NORTHWESTERN UNIVERSITY

Language-Being-Spoken and Other Indexical Dimensions in Monolingual and Bilingual Speech Processing

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Linguistics

By

Charlotte Reiss Vaughn

EVANSTON, ILLINOIS

September 2014

ABSTRACT

Language-Being-Spoken and Other Indexical Dimensions in Monolingual and Bilingual Speech Processing

Charlotte Reiss Vaughn

This dissertation explores the relationship between the processing of various indexical dimensions of speech for both monolinguals and bilinguals. These relationships were examined in several contexts for English monolingual and Mandarin-English bilingual listeners through the use of two experimental paradigms, the speeded classification task (Experiments 1-3) and the multi-phase binned classification task (Experiment 4). The speeded classification task taps into an earlier point in processing than does the binned classification task. In Experiment 1, the indexical dimension *language-being-spoken* (e.g. English or Mandarin Chinese in these experiments) showed a mutual and symmetrical pattern of interference with the indexical dimension *gender* (male or female). In Experiment 2, *language-being-spoken* showed a mutual and asymmetrical pattern of interference with the indexical dimension *talker* (identified by a name, e.g. Wei or Li), where it was harder for listeners to ignore language-being-spoken when attending to talker than the reverse. And, in Experiment 3, *language-being-spoken* showed asymmetrical interference with the non-linguistic dimension *amplitude* (e.g. loud or soft), where listeners could not ignore amplitude when attending to language-being-spoken, but could ignore language-being-spoken when attending to amplitude. Taken together, these results demonstrate language-being-spoken's place within a dimensional processing hierarchy: in this paradigm, language-being-spoken was equally as salient as gender, more salient than talker, and less salient

than amplitude. Importantly, this hierarchy did not differ according to language background; monolinguals and bilinguals behaved similarly in this task, whether or not the bilinguals had knowledge of Mandarin. Thus, the integrality of language-being-spoken with these other dimensions appears to be independent of the language experience of listeners. In Experiment 4, the results of a multi-phase binned classification task revealed a markedly different hierarchy, where language-being-spoken was most salient. Notably, the salience of language-being-spoken, and general task performance, was again similar for English monolinguals and Mandarin-English bilinguals. Comparing results across experimental paradigms reveals that the relative salience of a dimension is dependent on processing stage and task demands. In this dissertation, language-being-spoken was found to be: integrated with gender, talker, and amplitude, but in different ways, equally important to listeners regardless of language background, and perhaps more relevant later in processing than earlier.

ACKNOWLEDGEMENTS

Thank you first to my committee, Ann Bradlow, Matt Goldrick, and Satoru Suzuki for their help and support. Thanks also to Susanne Brouwer, who was involved at the early stages of this project and who remained a much-appreciated source of encouragement throughout the dissertation process. Many thanks go to Vanessa Dopker for subject running, recruitment, and other technical and on-the-ground support. Huge thanks also go to Chun Liang Chan for programming the binned classification experiment in Flash, and for helping with very many other things. Emily Kahn and Lindsay Valentino helped out with subject running, for which I am thankful. I gratefully acknowledge Grant R01-DC005794 from NIH-NIDCD, under which this work was partially supported.

I would not have gotten to this point without my brilliant cohort, Jordana Heller, Lisa Dawdy-Hesterberg, Jenna Luque, Kyounghee Lee, and Elizabeth Mazzocco. To Jordana, especially, I cannot imagine going through graduate school without you. I'm glad I never have to know what that would be like. Thanks also to Brady Clark for being a source of stability and understanding, and to Janet Pierrehumbert for support on many occasions. Thanks ever so much to Lori Allen for continuing to help me get out of my head and find balance.

I feel very lucky to have such a supportive family, and thank my parents for a lifetime of unending encouragement. Thank you to my parents and my sisters for inspiring me, and also for giving me some really good words. Thanks to Artie for accompanying me on so many walks and for reminding me what is important. Most of all I thank Tyler Kendall, for everything. Here is to the rest of our lives together, where I will no longer space out thinking about my dissertation.

CHAPTER 1: TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Introduction

One of the central tasks of psycholinguistics is to understand how humans process language. Research has increasingly shown that indexical information is implicated in many levels of language processing. However, little is known about the internal structure of the system of indexical dimensions itself. Deeper investigations into these indexical dimensions and the relationships between them, then, furthers the general understanding of human language processing. Thus, this dissertation explores the relationships between various indexical dimensions of speech in processing, paying special attention to a little-explored dimension relevant to bilinguals, namely, which language is being spoken.

1.2 The goals and structure of this dissertation

This dissertation reports on a series of four experiments exploring how the processing of the language that is being spoken is related to the processing of other speech dimensions, for both monolinguals and bilinguals. This relationship is examined at several points in processing through the use of two experimental paradigms: the speeded classification task, used in Experiments 1-3, taps into an earlier point in processing than does the multi-phase binned classification task, used in Experiment 4. All of these experiments test participants across different language backgrounds; each experiment compares the results of English monolinguals and Mandarin-English bilinguals, and one experiment (Experiment 2) also includes a group of bilinguals who speak English as well as a language that is not Mandarin Chinese. Taken together, the results of these experiments will provide a sense of language-being-spoken's place within the dimensional processing hierarchies of listeners from different language backgrounds.

A series of experiments using the speeded classification task pairs the language-being-spoken dimension with three other dimensions of the speech signal. Experiment 1 tests the dependency relation between the indexical dimension *language-being-spoken* (always English or Mandarin Chinese in these experiments) and the indexical dimension *gender*. Experiment 2 assesses *language-being-spoken*'s relationship with the indexical dimension *talker*. And, in Experiment 3, the relationship between *language-being-spoken* and the non-linguistic dimension *amplitude* is examined. A different paradigm is employed in Experiment 4, a modified version of the free classification task (here called the multi-phase binned classification task) as a way to determine which dimensions are salient to participants under different task demands, namely when they are asked to form open-ended groups of stimuli without time pressure. The use of multiple paradigms allows for the assessment of the relative salience of dimensions at different points in processing.

The dissertation is organized as follows. The rest of this chapter provides the background necessary to situate the questions addressed by this work within the broader literature. Then, Chapter 2 gives an overview of the speeded classification paradigm, describing its theoretical and methodological underpinnings and reviewing previous studies using this paradigm to test speech dimensions. Chapter 3 describes the hypotheses, methodology, and results of the three speeded classification experiments (Experiments 1-3), and Chapter 4 discusses these results in light of the hypotheses raised and speculates about ways in which methodological and theoretical considerations may account for the results observed. Next, Chapter 5 presents the hypotheses, methodology, and results of the multi-phase binned classification task (Experiment 4), and integrates these results with those of the first three experiments. The dissertation closes with Chapter 6, which presents conclusions and possibilities for future work.

1.3 Chapter outline

This introductory chapter first provides a terminological overview of the types of information conveyed by speech dimensions, and positions the dimension language-being-spoken within this framework. Then, I review studies documenting how gender, talker, and language-being-spoken are classified in isolation, before moving on to describe how certain types of dimensions—namely linguistic dimensions and indexical dimensions—are related in processing. Based on this information, I describe how the language-being-spoken dimension appears to relate in processing to dimensions conveying indexical information. The next section of this chapter situates the language-being-spoken dimension, and its relationships with other dimensions, in the context of bilingual mental representation. Then, properties of bilingual cognition are discussed before the chapter concludes with a summary.

1.4 Information conveyed by speech dimensions

I begin by positioning indexical dimensions in terms of a broader collection of dimensions present in the speech signal. First, I use the term *dimensions* here rather than a similar term like *categories* or *attributes* following the line of studies using the speeded classification paradigm (Garner, 1974; also called the Garner paradigm or task after its pioneer, W.R. Garner), since the majority of experiments in this dissertation make use of this paradigm.

Following Garner (1978, p. 98), the term dimension is used to mean “any variable attribute of a stimulus which exists at two or more levels.” Any snapshot of the speech signal, then, is made up of a myriad of dimensions, ranging from dimensions like amplitude (with a continuous range of levels) to dimensions like talker gender (with a limited set of levels, namely “male” and “female”) and the manner of articulation of a particular consonant (with a range of levels depending on the language). Different dimensions in the speech signal, then, provide different types of information to the listener. As mentioned above, the specific dimensions that will be tested in this dissertation are: the language that is being spoken, talker gender, talker identity, and amplitude. I now present the terminology that this dissertation uses to describe speech dimensions in terms of the types of information they provide to listeners, and note where each of the dimensions tested in this dissertation fit within this framework.

Prior work has typically divided the types of information carried in the speech signal into two major categories, linguistic information and indexical information (Abercrombie, 1967). Thus, the main terminological division made between dimensions of the speech signal here is between those which convey linguistic information and those which convey indexical information. It is important to note, however, that the type of information conveyed by a dimension is context-dependent. For example, some dimensions, such as whether a consonant is produced in breathy voice, convey linguistic information in some languages (e.g. Hindi) but convey indexical information in others (e.g. English). Even within a language, certain acoustic cues may be used to convey multiple types of information. For example, in Mandarin Chinese, fundamental frequency provides indexical information (e.g. talker gender), but also provides linguistic information (e.g. lexical tone).

From the perspective of the listener, linguistic dimensions involve those aspects of speech that can be used to decode the talker's message, from the phonological level to the morphological to the syntactic to the semantic to the pragmatic. Although some linguistic dimensions rely on few acoustic correlates (e.g. lexical tone is carried by the fundamental frequency of the signal), all linguistic dimensions are abstract and symbolic; in order to interpret them, listeners must rely on learned concepts and categories, a different set for each language with which a listener is familiar. Linguistic dimensions carry information which makes up the propositional content of utterances.

From the perspective of the listener, indexical dimensions are those properties of the speech that are "pointers" to the context of the utterance (following Peirce, 1940). Some indexical dimensions convey information about a talker that is stable over the course of the interaction, such as his or her gender, socioeconomic status, ethnicity, age, level of education, sexual orientation, and other such information. Other indexical dimensions convey information that may not stay constant over the course of the interaction, such as the talker's emotional state or attitude towards the topic at hand. Like linguistic dimensions, the interpretation of indexical dimensions also relies on abstract and learned categories; listeners must rely on associations between particular acoustic correlates and their social meanings in order to glean indexical information from these dimensions. Several dimensions tested in this dissertation, including language-being-spoken, talker gender, and talker identity, are considered indexical.

Indexical dimensions can be and are used by listeners to help decode linguistic dimensions. For example, knowing the identity of a talker may make a difference in interpreting meaning on a pragmatic level. Listeners would likely interpret the utterance "You have a green light" differently depending on whether their interlocutor is a passenger in the car they are driving, or is a Hollywood executive in a meeting at which they are making a pitch. This talker identity dimension is also important on a phonetic level: given a particular acoustic signal, a listener may not know whether a word is /bæt/ or /pæt/ given the potentially overlapping voice onset time (VOT) distributions of /p/ and /b/. Accessing the talker dimension aids speech perception by allowing the listener to calibrate to a distribution of tokens appropriate to the specific talker. Note that this process may be done without overtly identifying the talker, but merely by attending to talker-related acoustic properties of the signal. In other words, listeners may not need to make explicit judgments regarding indexical categories in the course of language comprehension, but certainly access indexical dimensions on some level. The relationship between linguistic and indexical dimensions in processing will be explored in detail in Section 1.6, below.

This dissertation makes one other division between dimensions, delineating those dimensions that provide information that is neither linguistic nor indexical in a category called *non-linguistic* (throughout this dissertation I use this term as a shorthand for *non-linguistic and non-indexical*; other authors may refer to dimensions that communicate this type of information as *acoustic*). Some non-linguistic dimensions, like noise, are not necessarily carried in the speech stream itself, but are part of the sound signal more generally. These dimensions communicate information that is not usable for understanding the talker's meaning, nor is it straightforwardly attributable to the talker or communicative context. Non-linguistic dimensions may be purely psychophysical properties of the speech signal, like amplitude. Or, in the case of noise, they may represent a different sound signal that is co-present in the communicative environment, temporally overlapping with the speech signal. Certain dimensions that can portray linguistic information, like the pitch of a syllable, can also be non-linguistic in certain cases, like the pitch of a non-speech pure tone. Likewise, certain non-linguistic dimensions can also provide indexical information, such as when amplitude alerts listeners to how close the talker is to them in physical space, or when a listener learns that a talker habitually speaks quietly or loudly. In the present experiments, amplitude does not readily convey indexical information, which leads to its categorization as non-linguistic for the present purposes. Non-linguistic dimensions tend to be lower-level, perceptual properties which do not require learned information for interpretation, in opposition to both linguistic and indexical information, which require the prior existence of concepts or categories for interpretation.

1.4.1 The language-being-spoken dimension

This dissertation will pay special attention to one particular dimension, the language that a particular talker is speaking, which will be referred to as the *language-being-spoken* dimension (often abbreviated here as L-B-S). As previously mentioned, language-being-spoken will be considered an indexical dimension here, by analogy with other dimensions that convey indexical information, like the language variety spoken by a talker, whether regional (i.e. Southern American English), ethnic (i.e. African American English), or L1-related (i.e. Spanish-accented English). This category is highly relevant to bilinguals, who must determine which language is being spoken as soon as they encounter a new speech stream. Language-being-spoken, while certainly related to language, is not a linguistic category, since it by itself does not convey linguistic meaning. As is true of talker-related indexical dimensions, language-being-spoken likely also interacts with linguistic dimensions; in order to decode the linguistic information in incoming speech tokens, listeners must know which language's distribution to pull from. Note, again, that this process may occur with or without explicit classification of the language being spoken.

1.5 Independent classification of dimensions

As was previously mentioned, each of the indexical dimensions tested in this dissertation (language-being-spoken, gender, and talker) may be implicated in the course of linguistic processing. However, classification of stimulus values along each of these dimensions can be accomplished as an end in itself. In other words, listeners are able to process indexical dimensions to extract indexical information just as they are able to process linguistic dimensions to extract linguistic information. These kinds of indexical classification tasks are performed by listeners in the course of everyday life. For example, listeners often have to identify the identity of a person by their voice (on the phone or in the dark, for example). The next three sections review what is known about how listeners perform gender classification, talker classification, and language identification, the indexical dimensions that will be examined in this dissertation. The terms *identification*, *classification*, and *recognition* will be used interchangeably, and the use of any other task (e.g. discriminability, same/different, etc.) will be noted accordingly.

1.5.1 Gender classification

Listeners can successfully classify unfamiliar talkers by gender (Lass, Hughes, Bowyer, Waters, & Bourne, 1976). Gender classification performance is impaired when low-pass filtered or whispered speech is presented to listeners as compared with recordings of the full speech signal, but classification in these degraded conditions is still above chance (Lass et al., 1976). Listeners appear to use few acoustic cues to perform gender classification; fundamental frequency (F0) is the major cue, though formant characteristics are also important (Childers & Wu, 1991; Coleman, 1971). Listeners are able to make a gender classification "within a few tens of milliseconds following the burst release of the consonant (Swartz, 1992) or, if the decision is based on voice pitch, within a few fundamental frequency cycles (a few tens of milliseconds more) into the vowel following the first consonant (Robinson & Patterson, 1995)" (as reported in Kaganovich, Francis, & Melara, 2006, p. 167). Despite the relative robustness and rapidity of gender processing, however, there is evidence that gender is not stored abstractly as two monolithic "male" versus "female" categories, but rather that it is likely represented as auditory-based perceptual representations in memory (Mullennix, Johnson, Topcu-Durgun, & Farnsworth, 1995). These representations "are probably an auditory composite of the various acoustic factors relevant to voice gender like F0, formant frequencies, breathiness, etc." (Mullennix et al., 1995, p. 3091). Nonetheless, gender is a highly salient indexical dimension, and its ability to be processed accurately and quickly is well-established.

1.5.2 Talker classification

Listeners are capable of identifying small sets of familiar talkers using sentence-length samples with very high rates of accuracy (Pollack, Pickett, & Sumby, 1954; Van Lancker, Kreiman, & Wickens, 1985). Acoustic cues used by listeners to perform talker identification largely appear to be global, or supra-segmental, features. For example, fundamental frequency, voice quality, and speaking rate correlate with listeners' ratings of talker similarity (Klatt & Klatt, 1990; Walden, Montgomery, Gibeily, Prosek, & Schwartz, 1978). In terms of time course, Andics, McQueen, & van Turennout (2007) suggest that talker identification can be performed in less than 500 ms.

It is worth noting that not all acoustic cues used by listeners to recognize talkers are directly related to talker anatomy (e.g., vocal tract length, oral cavity size, and fundamental frequency range); it is not the case that listeners only make use of physiological characteristics when processing talker information, but that listeners also make use of speech patterns of talkers that are less inherent, such as VOT. This indicates that listeners represent certain characteristic aspects of talkers' voices in memory. Several lines of evidence have found that the voice of a specific talker has been shown to cue alterations in phoneme perception specific to that talker (Allen & Miller, 2004; Bőhm & Shattuck-Hufnagel, 2009; Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2006, 2007). For instance, listeners pay attention to the characteristic voice onset time of a talker such that they can identify whether a given VOT token is consistent with their experience of that talker's speech (Allen & Miller, 2004). Further, listeners are surprised when a familiar talker's characteristic utterance-final phonation type is replaced with a different one (Bőhm & Shattuck-Hufnagel, 2009). Also, listeners shift phoneme category boundaries in a talker-specific way after perceptual learning (e.g., Kraljic & Samuel, 2005, 2006, 2007). These studies also demonstrate that talker information appears to be used when processing linguistic information, a relationship which will be explored in detail in Section 1.6.1, below.

Finally, it has been shown that talker classification can be performed given a degraded signal, such as in reversed speech (Van Lancker et al., 1985) and sinewave speech (Remez, Fellowes, & Rubin, 1997) suggesting that the presence of linguistic information is not necessarily required in talker classification, a relationship which will be explored in detail in Section 1.6.2, below.

1.5.3 Language-being-spoken classification

Language classification can be performed successfully from short samples of speech (e.g. Bond & Fokes, 1991; Lorch & Meara, 1995; Stockmal, Moates, & Bond, 2000). Lorch & Meara (1995) showed that English monolingual listeners could distinguish between two unknown languages (Farsi and Greek) at a rate slightly above chance when asked to make a same-language vs. different-language judgment on the basis of 2-second speech samples spoken by two different speakers of each language. Stockmal et al. (2000) showed that listeners could perform the same task accurately even when the samples were produced by the same bilingual talkers producing each of their two languages (eight bilinguals, each representing a different pair of languages). In this case, same-different judgments by listeners unfamiliar with all languages tested were better than chance on sentence-length stimuli, though discriminability was different for different language pairs.

Language classification appears to depend on a variety of acoustic cues, both local and global, such as rhythm, pitch patterns, and distinctive segments (Muthusamy, Jain, & Cole, 1994; Stockmal, Muljani, & Bond, 1996). There are also indications that listeners may use talker voice information when making these judgments (Muthusamy et al., 1994), which suggests that talker information may be implicated in language-being-spoken processing, a relationship specifically explored in this dissertation (its relevant background is given in Section 1.7.2, below). While no studies have directly measured the time course of language classification, evidence from an ERP study testing listeners' detection of unexpected language changes suggests that language recognition may be performed quickly: Highly proficient Welsh-English bilinguals detected an unexpected language change between Welsh and English as early as 200 ms following word onset (as indexed by an increased N1 response), and English monolinguals did so at least 100 ms later (as indexed by N400 modulation; Kuipers & Thierry, 2010). However, so far, no studies have measured the behavioral time course of language classification.

1.6 Processing relationships between dimensions: Examining the case of linguistic and indexical dimensions

After having considered the processing of specific dimensions in isolation, I now move on to review what is known about how different dimensions are related in processing. The speech signal is complex and multidimensional, and individual dimensions are rarely, if ever, encountered on their own. Thus, examining the ways in which multiple dimensions interact in processing—the primary goal of this dissertation—is an important component of understanding speech processing more generally. In particular, many studies, which will be reviewed in the following sections, have investigated whether one dimension is implicated when processing a separate dimension. Taken together, such information can be used to map out a processing hierarchy of relationships between dimensions. In the sections that follow, I review the well-studied processing hierarchy between linguistic dimensions and indexical dimensions, which partially motivates the current research. It will be seen that while indexical information appears to be implicated during linguistic processing, indexical processing may be accomplished more independently of linguistic processing. In many of these studies, the talker dimension is used to be representative of indexical dimensions more generally.

1.6.1 Linguistic processing is at least partially contingent on indexical processing

Though it has been proposed that that listeners abstract indexical dimensions away from linguistic dimensions when processing speech (see Klatt, 1989 for an overview), for decades it has been widely accepted that talker-related characteristics are not stored or processed entirely independently of the linguistic content of the signal. That is, when processing speech, listeners are sensitive to both the phonemes being uttered and the voice uttering them. In one of the earliest demonstrations of the interaction between linguistic and indexical information, a so-called talker normalization effect, Ladefoged & Broadbent (1957) found that participants categorized an ambiguous vowel in a /bVt/ word differently (as either /bɪt/ or /bɛt/) depending on the formant frequencies present in a preceding carrier sentence, demonstrating that phonetic processing is contingent on the acoustic properties of a talker. Two main lines of work, described in the following sections, provide evidence that linguistic processing is contingent on indexical processing. Then, other studies are described which document situations where listeners process linguistic information without the ability or the propensity to recognize voices, indicating that there are cases when indexical information is not implicated in linguistic processing.

1.6.1.1 Talker interference effect

Part of the evidence for the relatedness of talker and linguistic dimensions in processing comes from studies investigating what has been called the *talker interference effect:* listening to speech by multiple talkers as compared to one talker results in slower reaction times and disrupted accuracy on many tasks. If no dependency existed between talker and linguistic properties in the signal, then the number of talkers present should not impede linguistic processing, yet many studies have found just this type of disruption. For example, listeners are slower to respond in a word monitoring task when there are multiple talkers than when there is only one talker (e.g. Assmann, Nearey, & Hogan, 1981 for vowels; Nusbaum & Morin, 1992 for words), and this slowing is affected by working memory load (Nusbaum & Morin, 1992). Likewise, when given a set of utterances, listeners are slower and less accurate at naming a word spoken in noise if the utterances are spoken by a mix of talkers instead of one talker (e.g. Creelman, 1957; McLennan, 2006; Mullennix, Pisoni, & Martin, 1989; Peters, 1955; Sommers, Nygaard, & Pisoni, 1994). Also, listeners recall fewer words from a list spoken by multiple talkers as compared to a list spoken by one talker (Martin, Mullennix, Pisoni, & Summers, 1989, but see Goldinger, Pisoni, & Logan, 1991 and Nygaard, Sommers, & Pisoni, 1994 for evidence that inter-stimulus-interval modulates this effect). Further, there is evidence that talker identity across genders also has an influence on linguistic processing. Expectations about the gender of a talker can alter the perceived phoneme (e.g. /s/ or /ʃ/) of an acoustically identical stimulus (Strand, 1999), indicating that linguistic processing is also affected by talker gender. Such studies support the idea that linguistic processing is talker-contingent.

1.6.1.2 Talker specificity effect

Other evidence that talker and linguistic processing are interdependent comes from a line of studies demonstrating what has been called the *talker specificity effect*: repeated exposure to a single talker aids speech processing. If talker and linguistic information were processed separately, one would not expect to find a benefit for familiar talkers, but studies have repeatedly found that the speech of familiar talkers is easier to understand. For example, in word recognition paradigms, recognition is enhanced when the listener has some experience with a specific talker. In noise, words spoken by familiar talkers are understood better than words spoken by unfamiliar talkers (Nygaard et al., 1994; Nygaard & Pisoni, 1998). In terms of memory, voice familiarity appears to increase memory for words or sentences, though the results for different types of tasks are mixed (see Goh, 2005 and Luce & Lyons, 1998, for discussions). Recognition memory for words is more accurate when the voice is the same at exposure and at test, and this same-voice priming can last for up to a week (Goldinger, 1996). Talker-specific adaptation has also been shown to increase word recall; listeners are better at remembering whether a given word was previously heard if the word is presented in the same voice as before (e.g., Bradlow, Nygaard, & Pisoni, 1999; Martin et al., 1989; Palmeri, Goldinger, & Pisoni, 1993; cf. Goldinger et al., 1991 for serial recall). These studies also provide evidence that linguistic processing is talker-contingent.

1.6.1.3 Insensitivity to talker information

Other work, however, points out cases where linguistic processing can be done seemingly independently of indexical processing. First, patients with phonagnosia demonstrate a relative independence of talker identification from the linguistic signal. Phonagnosics are individuals who have unimpaired linguistic processing, but who show deficits in the recognition of familiar voices, most often resulting from a lesion in the right hemisphere (Van Lancker, Cummings, Kreiman, & Dobkin, 1988). The existence of individuals who exhibit such behavior supports the idea that linguistic processing is not completely contingent on indexical processing.

Even among listeners without impairments in voice recognition, there is evidence from the change deafness paradigm that listeners are not always sensitive to the voice of a talker. When performing a shadowing task, more than 40% of participants did not notice when the voice they were shadowing changed (Vitevich, 2003), suggesting that the relevance of encoding talker characteristics may be task-dependent.

1.6.2 Talker processing is less contingent on linguistic processing

While the results described above overwhelmingly show that talker information is implicated in linguistic processing in most cases, there is evidence that this relationship is asymmetrical. That is, studies have found that talker classification may make use of linguistic information when it is available, but that talker classification can also be done independently of linguistic processing. Below, studies are reviewed that demonstrate listeners' ability to perform talker identification on a speech signal that has been modified to remove or distort the linguistic content, results which provide evidence for this dissociation. Then, evidence is presented that if linguistic information is present, it can be used in the service of talker processing.

1.6.2.1 Talker identification is independent of linguistic information

A number of studies have found that familiar talkers can be accurately identified even in when the linguistic content of the signal has been stripped away, or when the signal is otherwise degraded. Van Lancker et al. (1985) found that listeners can successfully identify famous voices even when their speech is played in reverse, removing listeners' ability to access linguistic information. Remez, Fellowes, & Rubin (1997) found that their colleagues could recognize one another by voice even when their speech had been converted to sinewave speech (which removes much of the voice quality information from the signal). In a follow-up, Sheffert, Pisoni, Fellowes, & Remez (2002) showed that training on discriminating speakers based on sinewave speech generalized to new sinewave sentences as well as new sentences produced naturally. These results indicate that talker processing is not completely contingent on linguistic processing.

1.6.2.2 Talker processing may make use of linguistic information

When linguistic information is available in the signal, however, it does appear that listeners can make use of it when classifying talkers. Andics et al. (2007) showed that listeners' voice discrimination ability was related to the specific segments present in the sample of speech they heard. Dutch listeners heard a series of Dutch /CVC/ words spoken by 13 male native Dutch talkers and were asked to perform a same/different talker task. The phonetic content of the speech affected talker discrimination performance (e.g. talker discrimination was higher for words containing onset /m/ versus onset /l/), and this was true across segments in different positions in the syllable. The previously described result that listeners can recognize familiar talkers based on sinewave speech may also be used to illustrate that linguistic processing can be implicated in talker processing, as sinewave speech does retain phonetic information (Remez, Fellowes, & Rubin, 1997).

1.6.3 Linguistic–indexical interference: applying the speeded classification paradigm

The relationship between indexical and linguistic dimensions has been directly tested by several studies through the use of the speeded classification paradigm (or Garner paradigm). Since this task is used in Experiments 1, 2, and 3 of this dissertation, a much more thorough treatment of its rationale, history, and use will be given in Chapter 2, so for now it is only described briefly. Under an assumption that limited resources impose processing constraints, the speeded classification task tests how difficult it is to ignore irrelevant variability from one dimension when attending to the other dimension. The amount of variability from the irrelevant dimension is manipulated across different blocks of stimuli. If irrelevant variability makes it hard for participants to selectively attend to one dimension, this indicates that the two dimensions are related in processing. In this way, this paradigm provides a direct measure of whether two dimensions interact in processing, in what ways, and how much.

Using the speeded classification paradigm, Mullennix & Pisoni (1990) tested the processing relationship between a linguistic dimension, word-intial phoneme (/b/-initial or /p/-initial), and an indexical dimension, the gender of the speaker (male or female). Participants could not ignore variability from either dimension when attempting to selectively attend to the other, meaning that the two dimensions do indeed interact in processing. More interference was found as variability in the stimuli increased (using more talkers and more words). Further, there was an asymmetry between the dimensions such that it was harder to ignore gender variability when attending to phoneme than it was to ignore phoneme variability when attending to gender, a result replicated by Jerger et al. (1993) across age groups ranging from 3 years to 79 years of age.

Also using the Garner task, Green, Tomiak, & Kuhl (1997) found that the linguistic dimension phoneme interacted in processing with two indexical dimensions, gender and speaking rate, though the specific pattern of interference differed from Mullennix & Pisoni's (1990) results. For the phoneme–gender comparison, when classifying syllable-initial phoneme (/b/-initial vs. /p/-initial), interference from gender was found, but when classifying gender, there was no interference. For the phoneme–speaking rate comparison, interference from speaking rate (fast vs. slow) was found when processing phoneme, and an equal amount of interference from phoneme was found when processing speaking rate.

Cutler, Andics, & Fang (2011) paired the linguistic dimensions vowel (Dutch /ɛ/ vs. /ɒ/) and consonant (Dutch /t/ vs. /s/), with the indexical dimension talker identity ("Peter" vs. "Thomas") in a speeded classification paradigm and found that that it was harder for listeners to ignore the indexical dimension than the linguistic dimensions. Similarly, Kaganovich et al. (2006) tested the indexical dimension talker identity (male talker 1 vs. male talker 2) and the linguistic dimension vowel (/ɛ/ vs. /æ/). In this case, however, participants experienced equal amounts of interference from vowel when classifying talker as they experienced from talker when classifying vowel. This study also measured participants' ERP response during the speeded classification task, and results showed accompanying sustained negativity for blocks containing irrelevant variability as compared with blocks without irrelevant variability, starting at about 100ms after stimulus onset for both dimensions. The authors used the ERP results to suggest that the dimensional interference occurred at an early point in processing (as opposed to at the response selection stage).

These speeded classification studies provide a direct test of the processing relationships between linguistic dimensions and indexical dimensions. Though previous work showed that talker processing was implicated in linguistic processing (Section 1.6.1), and that linguistic processing can be implicated in talker processing, though perhaps less so (Section 1.6.2), the Garner paradigm experiments described above directly assessed the relationship between the dimensions. The results of these experiments are somewhat mixed, but in general, interference was found between indexical and linguistic dimensions, and in some cases, it was harder for listeners to ignore the indexical dimension (mirroring the results described in Sections 1.6.1 and 1.6.2 above).

1.7 Processing relationships between language-being-spoken and other dimensions

The fact that variation in indexical dimensions has been shown to affect linguistic processing makes it all the more important in the field of psycholinguistic speech perception to better understand relationships between indexical dimensions. To that end, this dissertation investigates processing dependencies between several indexical dimensions, concentrating on the language-being-spoken dimension. The central comparison in this dissertation will be between language-being-spoken and talker identity, but tests of language-being-spoken's relationship with gender and with amplitude will provide important complementary comparisons.

As was just noted, in the case of the relationship between linguistic and indexical dimensions, a number of studies investigated whether indexical information is implicated in linguistic processing (Section 1.6.1), and other studies investigated whether linguistic information is implicated in indexical processing (Section 1.6.2). In order to test the relationship between the dimensions directly, several studies using the Garner interference paradigm measuring interference from one dimension in the processing of the other (Section 1.6.3). Likewise, in the case of language-being-spoken and other indexical dimensions, the relationship of interest in this dissertation, several studies (which will be reviewed in Section 1.7.1, below) have investigated whether language-being-spoken information is implicated in talker processing. Some, but fewer studies (which will be reviewed in Section 1.7.2, below) have investigated whether talker information is implicated in language-being-spoken processing. However, as of yet, the relationship between language-being-spoken and other indexical dimensions has not been directly examined via the speeded classification task. This dissertation fills in this gap by conducting such tests in a series of Garner experiments described in Chapters 3 and 4. The language-being-spoken dimension's involvement in talker processing, and the talker dimension's involvement in language processing, will be taken into consideration when making predictions about the results of the current experiments (Sections 3.4.2).

1.7.1 Involvement of language-being-spoken information in talker/indexical processing

Work on cross-linguistic talker processing has shown that language information is at least partially involved in the process of talker identification. A number of studies have demonstrated that it is harder to identify talkers when they are speaking an unfamiliar language than in a familiar language, a phenomenon known as the *language familiarity effect* (Bregman & Creel, 2014; Goggin, Thompson, Strube, & Simental, 1991; Köster, Schiller, & Künzel, 1995; Perrachione & Wong, 2007; Thompson, 1987; Winters, Levi, & Pisoni, 2008). In an early demonstration of this effect, Thompson (1987) found that monolingual English listeners were better at identifying talkers speaking English than they were at identifying talkers speaking

Spanish, or talkers speaking English with a Spanish accent. Similarly, Goggin et al. (1991) found that monolingual English listeners identified bilingual English-German talkers more accurately when they were speaking English than when they were speaking German. They also found that English-Spanish bilingual listeners were equally accurate at identifying talkers in Spanish and in English. Similarly, in ABX talker similarity judgments, two talkers speaking the same language were judged as more similar to each other than two talkers speaking different languages, a finding that held true even when the two talkers were of different genders (Stockmal, Moates, & Bond, 2011). Johnson, Westrek, Nazzi, & Cutler (2011) demonstrated a version of this effect in infants, finding that 7-month-old infants are able to detect a talker switch in their native language (Dutch) but not in other languages. In a variation on this effect, Perrachione, Chiao, & Wong (2010) found a dialect familiarity benefit for talker identification between listeners familiar with the "General American" dialect versus listeners familiar with the African American English dialect.

Further, both language-general and language-specific properties appear to be used to discriminate between talkers. Winters et al. (2008) found that monolingual English listeners trained to identify German-English bilingual talkers speaking in one language could generalize that information and perform well on talker identification when those talkers were speaking their other language, indicating that there is sufficient language-general information available in the speech signal for talker identification. However, monolingual listeners trained on the talkers in their own native language (English) were not able to generalize this knowledge when tested on a language they did not know (whereas monolingual listeners trained on the talkers in a language unfamiliar to them—German—were able to generalize this knowledge to a test in their native language), indicating that these listeners used language-specific cues to identify the talkers. Further, English-trained listeners got to higher talker identification levels overall, which suggests using language-specific cues may allow for better representation of talker information. Along similar lines, Wester (2012) found that listeners were better at discriminating between bilingual talkers when the two samples were in the same language than when they were in different languages. Listeners performed well when doing unfamiliar language–unfamiliar language talker discrimination (as they could focus on the language-general indexical cues without being distracted by the linguistic content), and also performed well on native language–native language talker discrimination (as they had both language-general and language-specific cues to talker identity at their disposal). However, poorer performance was found for native language-unfamiliar language talker discrimination, which the author speculates was caused by increased task difficulty owing to the cognitive load involved in a familiar–unfamiliar language mismatch. Here it is important to note that while some language-general properties of talkers are certainly tied to the talker's anatomy (e.g. vocal tract length, oral cavity size, fundamental frequency range), certain non-physiological speech patterns also appear to transcend the language that is being spoken. For example, within a group of bilingual talkers, speaking rate in L1 was correlated with speaking rate in L2 (Kim et al., 2013).

Talker identification appears to improve if a listener even has minimal knowledge of the language (Köster et al., 1995; Sullivan & Schlichting, 2000), and recent evidence suggests that talker identification ability in a second language correlates gradiently with listeners' age of acquisition of L2 (Bregman & Creel, 2014). Köster et al. (1995) found that listeners with some knowledge of German were better at identifying a trained voice from a set of 108 utterances than were listeners with no knowledge of German, but there was no difference in performance between native speakers of German and native-English learners of German. Similarly, Sullivan & Schlichting (2000) found that British students in their first year of studying Swedish at university were better able to identify an imitation voice in a voice lineup than students with no knowledge of Swedish, but this ability did not improve over their course of Swedish study at university, nor was it as good as the performance of native Swedish speakers. Recently, Bregman & Creel (2014) demonstrated that Korean-English bilinguals who acquired English earlier learned to identify English talkers at a faster rate than those who acquired English later (as measured by the number of experimental blocks taken by participants to achieve a particular accuracy criterion). Notably, the bilinguals who acquired English earlier learned to identify English talkers as quickly as the monolingual listeners did, and also learned to identify Korean talkers as quickly as the bilingual Korean-dominant participants did. For these bilinguals, age of acquisition also correlated with a number of other measures of language background, including a metric of language dominance and a metric of lexical inventory, though the authors argue that age of acquisition is the factor that drives this effect. The authors summarize: “not only is it easier to learn voices in a language you know, it is also easiest to learn voices in a language learned early” (p. 93).

Typological closeness between the language background of the talker and the listener may also play a role in cross-language talker identification, but this relationship is not as clear when there is no knowledge of the target language. Köster & Schiller (1997) found that, out of listeners from four native language backgrounds who all had knowledge of German, English and German listeners performed better at identifying a target German voice than did Chinese and

Spanish listeners. However, out of listeners from these four native language backgrounds who did *not* have knowledge of German, English listeners performed better at identifying a target German voice than Spanish listeners, but Chinese listeners performed better than both of these groups (with German listeners performing the best).

The leading proposal about the basis of the language familiarity benefit in talker identification posits that it stems from stronger integration with linguistic processing systems in the listener's native language than in an unknown language (Perrachione & Wong, 2007). Listeners identifying voices in an unknown language do not have auditory representations of talkers speaking in that language to which new talkers in that language can be compared; they cannot integrate talker information with linguistic information and therefore are disadvantaged in that way as compared with a listener who knows the language of the talkers. In order to test this suggestion, Perrachione, Pierrehumbert, & Wong (2009) examined listeners' brain activation while identifying talkers speaking in a language familiar to them, and in a foreign language. Previous results indicate that linguistic information is lateralized in the left hemisphere of the brain within a familiar language (Galuske, Schlote, Bratzke, & Singer, 2000), while voice perception is thought to be lateralized in the right hemisphere (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; though cf. Kaganovich et al., 2006). Given these findings, Perrachione et al. (2009) tested whether talker identification in known and unknown languages would have differential representations across hemispheres. A dichotic listening paradigm for five-talker identification was used, where participants were first binaurally trained on the voices with feedback, then tested on performance one ear at a time such that listeners first had to identify target voices presented in the left ear while ignoring distractor voices from the right ear, and then had to repeat that test for the other ear. As predicted, right-ear (left-hemisphere) performance predicted talker identification better in the listener's native language than in a foreign language, which is consistent with the prediction that linguistic processing is recruited in native language talker identification. Additional evidence for this proposal is found in the results of Schiller, Köster, & Duckworth (1997), who found that the language familiarity effect did not appear to hold when participants were tested on speech where the linguistic content had been removed and replaced with the syllable /ma/.

In a further test of the hypothesis that talker identification is contingent upon linguistic processing, Perrachione, Del Tufo, & Gabrieli (2011) examined the cross-language voice recognition abilities of different populations: native English listeners with normal reading abilities, compared to those diagnosed as dyslexic, who have impaired linguistic processing. Under the aforementioned hypothesis, the authors speculated that participants with dyslexia would perform as poorly in identifying talkers in their native language as in an unfamiliar one due to their impaired phonological processing. This is indeed what they found; native English dyslexic listeners were worse than normal listeners at voice identification in English, and they performed equally well on voices speaking English and Chinese.

A recent study, however, presents a case that may be inconsistent with this line of thinking. Neuhoff, Schott, Kropf, & Neuhoff (2014) found that monolingual listeners of English and Spanish showed greater change deafness when listening to their familiar language than when listening to an unfamiliar language. When participants listened to a speech passage and were instructed to count the talker's breaths, fewer Spanish monolingual listeners detected a talker change when the talkers were speaking Spanish than when they were speaking English (meaning that they showed more change deafness in Spanish). The same was true of English monolinguals; fewer detected the talker change when listening to English than when listening to Spanish (meaning that they showed more change deafness in English). The idea that talker identification is performed better with access to linguistic information (the language familiarity effect) would suggest that listeners should be more sensitive to a talker change (and thus exhibit less change deafness) in a familiar language than an unfamiliar language. However, the opposite was found. The authors interpret this result based on previous processing asymmetries found in linguistic and indexical processing (such as those outlined in Section 1.6.3, above), and suggest that "attention to lexical/semantic information can override the language familiarity effect when listeners are not cued to listen for changes, resulting in greater change deafness in familiar languages" (p. 221). This result suggests that certain task demands may dampen the role of language-being-spoken in processing indexical information.

1.7.2 Involvement of talker/indexical information in language-being-spoken processing

Whether indexical information is implicated in the processing of language-being-spoken has not been as well studied as the reverse case, described above, but several studies examining the cues listeners use when identifying languages suggest that indexical dimensions may play a role. Muthusamy et al. (1994) trained participants to distinguish between ten languages, based on multiple speakers of each language. When asked about the cues they used to perform the task, some participants reported that they compared the voice quality of a talker producing a given language sample to the voice quality of a person whose native language they knew. In a task requiring participants to make same-different judgments on unknown languages, presenting the same sentence in the same language by two different talkers caused a decrease in accuracy (Stockmal, 1995). In an ABX task, listeners matched languages somewhat less accurately if they were spoken by talkers of different genders rather than by talkers of the same gender (Stockmal et al., 2011). Finally, in a multidimensional scaling analysis of perceptual cues used for language identification, one dimension out of two was construed as representing talker-specific information, being made up chiefly of voice quality and speaking rate (Stockmal et al., 1996).

1.8 The language-being-spoken dimension in bilingual representation and perception

Having established that language-being-spoken and other indexical dimensions may be processed in an interdependent manner, I now turn to discuss how these dimensions might be related for bilingual listeners in addition to monolingual listeners. So far, most of the work reviewed here on the processing of indexical dimensions, including language processing, has been based on the behavior of monolingual participants. With the addition of the language-being-spoken dimension to the landscape of indexical dimensions, the lack of work on indexical processing in bilinguals becomes a conspicuous omission.

This dissertation is concerned with bilinguals' and monolinguals' use of the language-being-spoken dimension and other dimensions in perceptual processing (as opposed to in, for example, mental representation). Bilinguals and monolinguals are able to engage with the language-being-spoken dimension in different ways: Monolinguals can certainly interact with language-being-spoken in perceptual classification, but the fact that they only know one language means that their use of the dimension in mental representation is limited. Bilinguals, however, may form a level of representation based on language-being-spoken, and may use language-being-spoken as an organizational tool in their representation of other dimensions. Bilinguals' use of the language-being-spoken dimension in representation may influence how they use the dimension in perceptual processing, in turn affecting how it fits into their hierarchy of dimension in perceptual processing. Thus, the way in which language-being-spoken is represented, and its relationship with other dimensions in representation, provides relevant background for the main goal of this dissertation. As such, in the following sections, I first review studies which consider how language-being-spoken is represented by bilinguals, particularly during language acquisition, and how indexical information may be implicated in this process. Then, I review studies which document how bilinguals make use of the language-being-spoken dimension in organizing other types of information. Findings regarding bilinguals' representation of the language-being-spoken dimension, and language-being-spoken's role in representing other dimensions, will be taken into consideration when making predictions about the results of the current experiments (Sections 3.4.3.1 and 3.4.3.2).

1.8.1 Representation of language-being-spoken information by bilinguals

Figuring out how to represent multiple languages is an issue that begins early in bilingual language acquisition. In acquiring sound categories, bilingual infants must not only cope with the continuous, variable speech stream within one language, but also must learn to differentiate between languages. Across languages, sound inventories vary, and certain phoneme contrasts may be realized differently phonetically (e.g. VOT values in English and French overlap across categories), leading researchers to ask how multilingual infants deal with overlapping distributions of segments from their multiple languages. Put in different terms, the question is how simultaneous bilinguals learn to distinguish which language is being spoken.

Two non-mutually exclusive explanations have been put forward, a brute force approach and a language tagging approach (Sundara & Scutellaro, 2011). The brute force approach suggests that infants calculate statistics across both languages together, and eventually build up enough input to discriminate between the shallow peaks that form in the otherwise monolithic distribution. Evidence for this approach comes from the fact that bilingual infants' discrimination of highly frequent phonemes (e.g. /d/ in English and French; Sundara, Polka, & Molnar, 2008) does not follow the typical U-shaped developmental curve that has been found in bilingual infants' discrimination of other less-frequent phonemes (e.g. /e/-/ɛ/ in Catalan/Spanish bilinguals; Bosch & Sebastián-Gallés, 2003). Increased exposure to the frequent phoneme allows bilingual children to reliably discriminate at a younger age (by 8 months instead of 12 months), suggesting the simultaneous calculation of statistics over both languages by "brute force". The language tagging approach proposes that infants sort or "tag" tokens by language and then compute statistics separately for each. The language tagging approach, then, is consistent with the idea that bilinguals maintain a representational level for language-being-spoken. Such a proposal, of course, necessitates that infants can discriminate between languages in the first place. A variety of possible cues that infants might use for language differentiation have been proposed, including cues providing indexical information, which will be described in the following section. In the linguistic domain, it has been shown that newborn infants can discriminate between languages if the languages differ in rhythm class (Ramus, Hauser, Miller, Morris, & Mehler, 2000; Nazzi, Bertoncini, & Mehler, 1998), and that by 4-5 months of experience with one language, infants can discriminate between languages from within the same rhythm class (Bosch & Sebastián-Gallés, 1997; Nazzi, Jusczyk, & Johnson, 2000). Other prosodic cues such as pitch may also be used by infants in classifying languages (Vicenik & Sundara, 2013).

Given that bilinguals can use prosodic cues such as rhythm in order to distinguish between languages, Sundara & Scutellaro (2011) asked whether infants can actually latch onto the rhythmic distinction as a way to solve other language learning problems, namely the learning of two overlapping vowel distributions. They found that bilingual infants learning rhythmically dissimilar languages (Spanish and English) were better able to deal with overlapping /e/-/ɛ/ distributions than were bilingual infants learning rhythmically similar languages (earlier results of Bosch & Sebastián-Gallés, 2003 on Spanish/Catalan /e/-/ɛ/). In other words, the rhythmic difference between Spanish and English gave bilingual infants learning those languages a scaffold for tagging /e/-/ɛ/ tokens of each language as belonging to one language versus the other. (This explanation may also have played a role in Sundara et al.'s (2008) findings with /d/ discrimination in French-English bilinguals, as French and English also differ in rhythm class.) Thus, not only is the language tagging approach consistent with a bilingual representational level for language-being-spoken, but the Sundara & Scutellaro (2011) results leave open the possibility that indexical information could serve as a scaffold for language organization. This possibility is explored in the next section.

1.8.2 Indexical information may be implicated in the representation of language-being-spoken

Evidence from a series of studies using artificial languages can be taken to indicate that indexical information may be used in the representation of language-being-spoken. Several experiments have shown that adults can make use of certain indexical dimensions to aid in discriminating between artificial languages in a simulated bilingual environment. In Weiss, Gerfen, & Mitchel (2009), participants were tasked with learning artificial languages consisting of four CV nonsense words. Results showed that participants could only segment a pair of languages with "incongruent" statistics (where the transitional probabilities were noisier across languages than within languages) when they heard a consistent pairing between a voice and a language (the same male voice always produced one language and the same female voice always produced the other). Without differentiating indexical information present, participants were not able to accurately distinguish between the languages. Mitchel & Weiss (2010) extended these findings to include other indexical dimensions. When participants heard each language accompanied by a video of a different female face lip-synching along to the recording (the audio of both languages was presented in the same female voice, but a different face was presented "speaking" each language, where a particular face was always matched with a particular language), language segmentation was better than when participants only saw one face across both languages to accompany the audio (where the face did not provide an indexical cue). However, listeners did not latch onto just any cue that may aid language segmentation. When a different background color correlated with each language (the background on the screen was always teal when language 1 was presented and purple when language 2 was presented), participants segmented the languages no better than when there was no indexical cue present. Further, presentation of two static faces paired with languages (just as with the videos described above, but with a static picture), did not help listeners segment the two languages. Thus, listeners in these studies made use of a variety of indexical dimensions, but not just any differentiating information, to form multiple representations of these artificial language inputs. These results suggest a role for certain indexical dimensions in language-being-spoken representation, and that different types of dimensions may interact with language-being-spoken in various ways.

1.8.3 Semantic knowledge is represented language-specifically

Since bilinguals have access to more than one language it is also possible that they index information language-specifically, which would indicate that the language-being-spoken dimension is implicated in the representation of other dimensions. This appears to be true in a several domains. First, there is evidence that bilinguals represent semantic knowledge language-specifically. Marian & Fausey (2006) found that participants remembered academic knowledge better when tested on it in the language in which it was learned. Marian & Kaushanskaya (2007) found that participants were more likely to provide answers to general knowledge questions congruent with the language used when asking the question; when asked to "name a statue of someone standing with a raised arm while looking into the distance," Mandarin-English bilinguals were more likely to name the Statue of Liberty when asked the question in English, and to name the Statue of Mao when asked in Mandarin. These results indicate that semantic knowledge may be represented in a language-contingent manner in bilingual minds.

1.8.4 Phonetic information is represented language-specifically

Evidence from bilingual speech production and perception indicates that phonetic information is also represented language-specifically. For example, Bradlow (1996) showed that vowels of English are fronted relative to similar vowels in Spanish, suggesting a language-specific base of articulation for the two vowel systems. Similarly, English and French, as an example, both have voiced and voiceless stops, but the actual distributions of VOT values that distinguish voiced from voiceless overlap across the two languages; a voiced stop in English may have a VOT value which corresponds to the VOT of voiceless stops in French (e.g. Lisker & Abramson, 1964; Caramazza, Yeni-Komshian, Zurif, & Carbone, 1973). Further, Flege & Eefting (1988) demonstrated that bilinguals are sensitive to such distinctions by showing that early Spanish-English bilinguals perceptually separate short-lag (Spanish) and long-lag (English) versions of /t/, indicating that they have separate representations for English /t/ and Spanish /t/. Finally, as mentioned above, it has been well documented that languages vary along a continuum of rhythm classes (Ramus, Nespor, & Mehler, 1999), and that even newborns can distinguish between languages that vary along the rhythm timing scale (Ramus et al., 2000; Nazzi et al., 1998). Taken together, these results suggest that bilinguals organize phonetic information language-specifically.

1.9 Properties of bilingual processing

In addition to the ability of bilinguals to form representations of the language-being-spoken dimension, it has also been proposed that bilinguals differ from monolinguals in specific cognitive functions, which may affect performance on the experiments conducted in this dissertation. Thus, these reported differences will be summarized below.

It has been suggested, most notably by Bialystok and colleagues, that bilinguals perform better on tasks involving executive control than monolingual counterparts (e.g., Bialystok, Craik, Klein, & Viswanathan, 2004; Bialystok & Martin, 2004; see Bialystok, Craik, & Luk, 2012 for a review; but, see Hilchey & Klein, 2011; Paap & Greenberg, 2013 for failures to replicate these results). In these types of tasks a salient conflict is presented to participants, who have to ignore or inhibit certain information or responses in order to perform the task well. Experimental paradigms thought to invoke the executive control system and thus used in these studies include the Simon task, the flanker task, the Stroop task, and other tests of selective attention or inhibitory control. The executive control system of the brain, centered in the frontal lobes, is normally responsible for making selections and resolving conflicts (e.g. Miyake et al., 2000). Due to the language competition situation in the bilingual mind, it is thought that executive control systems are more engaged in language processing by bilinguals than in language processing by monolinguals. Evidence for the engagement of executive control areas of the brain in bilinguals is given in a meta-analysis of studies involving bilingual language switching, which found that four of the six brain areas recruited by bilinguals when doing language-switching tasks are areas considered to be part of the executive control system (Luk, Green, Abutalebi, & Grady, 2011).

Current explanations for the superiority of bilinguals at these types of tasks are based on the idea that both of a bilingual's languages are continually active, at least to some extent, even in situations strongly favoring one language or the other. Therefore, one account posits that bilinguals must suppress one language whenever they produce or comprehend language, and thus develop superior skills at inhibiting irrelevant information (e.g. Green, 1998). An alternative account suggests that bilinguals' experience with competing languages makes them better at active selection, or conflict resolution, rather than at interference suppression (e.g. Costa, Miozzo, & Caramazza, 1999; Hernández, Costa, & Humphreys, 2012).

Much is known about the advantages (and disadvantages) of being bilingual in linguistic processing, but these studies are beyond the scope of this dissertation, which examines the processing of indexical (and non-linguistic) dimensions. Since there is a dearth of research into indexical processing by bilinguals, which this dissertation attempts to fill, I now review a few studies demonstrating bilingual differences in the processing of non-linguistic information. Bialystok (1992) found that bilingual children are better than monolingual children at finding a simple visual pattern embedded in a complex figure, signaling that they may have superior abilities to ignore distracting information and to selectively attend to relevant information. Other evidence comes from a paradigm known as the dimensional change card sort task (which is similar in concept to the Garner paradigm), where participants are given cards displaying one of two shapes (e.g. circle or square) in one of two colors (e.g. red or blue) and are asked either to sort the cards by color or by shape, and then must switch to sorting by the other dimension. Bialystok (1999) and Bialystok & Martin (2004) found that 4-5 year old bilingual children are better at switching to the new sorting rule than are monolingual age-matched children. A similar task, the Simon task, has found analogous results in adult bilinguals. The Simon task manipulates the spatial congruency between the position of a stimulus (e.g. a colored shape) on the screen, and the position of the button participants must use to respond to the stimulus (e.g. based on color). Bialystok et al. (2004) found that while bilinguals and monolinguals across multiple age groups showed equal reaction times to stimuli presented in a baseline condition, monolinguals had a harder time than bilinguals in cases where the stimulus was presented on the opposite side of the screen than its corresponding response button. It should be noted, though, that no such difference was found for young adults (aged 20-30). Bilinguals have also been shown to react faster in color word-ink color conflict trials in a Stroop task (Bialystok, Craik, & Luk, 2008). Finally, adolescent bilinguals appear to more strongly encode the fundamental frequency of a target speech sound (the syllable /da/) than monolinguals as measured by auditory brainstem response, and that this enhancement was greater when multi-talker babble was present in the background of the signal, further underscoring bilingual advantages in tasks involving attention (Krizman, Marian, Shook, Skoe, & Kraus, 2012).

These results indicating domain-general differences in executive function between monolinguals and bilinguals will be taken into consideration when making predictions about the results of the current experiments (Section 3.4.3.3).

1.10 Summary

This chapter provided the background which contextualizes the present study. The literature reviewed here indicates that language-being-spoken may be an important dimension to bilinguals, and suggests that it may be related to other indexical dimensions in processing, but so far no work has tested this explicitly. To this end, the experiments in this dissertation will seek to determine language-being-spoken's relation to other dimensions, for monolinguals and bilinguals, at various phases in processing.

CHAPTER 2: THE SPEEDED CLASSIFICATION PARADIGM

2.1 Chapter outline

This chapter presents an overview of the speeded classification paradigm, also known as the Garner task, and reviews methodological and theoretical issues debated in the existing task literature that are relevant to this dissertation. First, I give a more detailed account of the paradigm, which will be used by the bulk of the experiments in this dissertation (Experiments 1-3). I discuss the logic of the task in detail and what it can say about the relationships between the dimensions it tests. Following that, I review several discussions in the history of the speeded classification task that will aid in interpreting the results of Experiments 1-3 of this dissertation. Specifically, I discuss how Garner interference is measured, ways in which dimensions can interact, and what it means for dimensions to interact in certain ways. The next major section of the chapter is devoted to a literature review of studies in the acoustic speech realm that have used the speeded classification paradigm. Following that, those findings from previous literature are summarized in a table in order to evaluate dimensions' dependency relations as related to their relative levels of processing. The potential role of discriminability in dependency relations is also raised. The chapter closes with a brief summary.

2.2 The speeded classification paradigm

The Garner speeded classification procedure is a forced-choice task requiring the listener to pay attention to one dimension of a stimulus while attempting to ignore variation in other dimensions of that stimulus; listeners must classify a multidimensional stimulus on the basis of one dimension at a time. The paradigm is designed to test whether two dimensions of a stimulus interact in processing or whether they are independent. Such relationships between stimuli are not properties of the stimuli themselves, but, rather, are properties of listeners' cognitive processing. In the task, stimuli are arranged into three stimulus sets, or blocks (named control, correlated, and orthogonal[1]), each with differing amounts of variation in the dimension not being attended to in the task at hand (control: no variation in un-attended to dimension, correlated: redundant variation, orthogonal: random variation). The task was originally developed to examine processing dependencies between dimensions in visual perception (Garner, 1974; Garner & Felfoldy, 1970). As such, the example used below describes a design where participants have to classify circles either on the basis of brightness or of size. In the control block of the brightness task (where participants are presented with a choice between brightness values, dark or light) there is no variation on the dimension not being attended to (size); the participant would only see stimuli such as those in the following figure. Only stimuli that are uniform in the un-attended dimension (in this case, size) are presented in the control block, making it impossible for the classification of brightness to be interfered with by size. (The control block of the size task would present only dark or only light circles for classification as "big" or "small".)

Figure 2.1. Control block stimulus options for brightness task, Version 1.

Control: Brightness task (Big only)		*Brightness*	
		Dark	*Light*
Size	*Big*	O	O
	Small		

There is, of course, another option that also does not introduce variation in size, where

[1] Authors vary with regard to the names of these dimensions; in other work, control blocks may be called baseline or single-dimension blocks, and orthogonal blocks may be called filtering blocks.

participants see only small circles. Participants are typically divided equally between the two control conditions:

Figure 2.2. Control block stimulus options for brightness task, Version 2.

Control: Brightness task (Small only)		Brightness	
		Dark	Light
Size	Big		
	Small	o	o

In the orthogonal block, there *is* variation in the un-attended dimension, as well as in the dimension being classified; the stimulus set includes random variation in both dimensions. Participants in an orthogonal block of both the brightness task and the size task would see all of the options given below:

Figure 2.3. Orthogonal block stimulus options, brightness and size tasks.

Orthogonal		Brightness	
		Dark	Light
Size	Big	O	O
	Small	o	o

In the correlated block, one particular value of one dimension is always paired with one particular value of the other dimension, such that there is redundant information, as in the two versions below which could each be used for both the brightness task and the size task (as with the multiple control conditions, there are two ways to create the correlated block, which are counterbalanced across subjects).

Figure 2.4. Correlated block stimulus options, brightness and size tasks, Version 1 (left).
Figure 2.5. Correlated block stimulus options, brightness and size tasks, Version 2 (right).

Correlated (1)		Brightness	
		Dark	Light
Size	Big	O	
	Small		o

Correlated (2)		Brightness	
		Dark	Light
Size	Big		O
	Small	o	

As the only difference between blocks is the amount of extraneous variation in the irrelevant dimension, differences between the blocks—measured in terms of accuracy and reaction time—are used to assess the extent to which that variation interferes with the processing of the other dimension. Thus, differences in performance across the different blocks determine what relationship holds between the two dimensions.

2.2.1 Testing indexical dimensions with the speeded classification paradigm

Since the Garner paradigm's initial development for use in the visual modality, it has also been applied in the domain of speech, beginning with Wood & Day (1972) which tested the relationship between pitch (F0) and place of articulation. As will be seen in an extensive review of previous studies in Section 2.7, below, many of the studies applying the Garner paradigm to dimensions within the speech domain have tended to investigate perceptual or non-linguistic dimensions, such as pitch and amplitude, or linguistic dimensions such as place and manner of articulation, vowel, and tone. There is a precedent for testing indexical dimensions in this paradigm, as well, and such studies will also be reviewed below. Studies investigating indexical dimensions, along with work investigating other abstract, higher-level types of dimensions (e.g. Melara & Marks, 1990b), demonstrate that dimensions requiring learned, conceptual categories can indeed be successfully tested using the Garner paradigm. For example, Eimas, Tartter, Miller, & Keuthen (1978) had participants classify stimuli based on place of articulation (e.g. bilabial vs. alveolar) versus manner of articulation (e.g. stop vs. nasal) of consonants; participants needed to be taught the meaning of each of these dimensions before being able to classify them. Further, Melara & Marks (1990b), had participants classify non-speech tones varying in frequency and amplitude as “apple” or “banana” (based on a non-psychologically real 22.5 degree rotation of a frequency-amplitude matrix), an arbitrary dimension which clearly had to be taught.

2.3 Measuring Garner interference

The measure of exactly how much one dimension interferes with the other, called *Garner interference*, is taken by subtracting a subject’s performance (as measured by accuracy and/or reaction time) in the control block from their performance in the orthogonal block. If one dimension of the stimulus cannot be ignored, the subject will have longer reaction times in the block where the other dimension is variable (orthogonal) than when it is constant (control). The existence of Garner interference, then, is a failure of selective attention that implies that the two dimensions are perceptually integrated. If there is no Garner interference present, the two dimensions are said to be separable. In visual perception, for example, the size of a circle and the angle of its diameter have been shown to be separable dimensions (Garner & Felfoldy, 1970).

Accuracy and reaction time (RT), the two measures of Garner interference, have been shown to have an inverse monotonic relationship (Ashby & Maddox, 1994), both representing perceptual distance between dimensions. If two dimensions are close together in perceptual space, values in one dimension in the presence of variability from the other dimension should be harder to classify (lower accuracy), and correct classifications of that value should take longer (higher RT), than if those dimensions are farther apart. In practice, many dimensions used in Garner tasks are discriminable enough that participants’ error rates are quite low, making error rates rather uninformative to compare across blocks. Therefore, most studies focus on the reaction time metric of Garner interference. As in any task of this type, it is important to rule out the presence of speed-accuracy tradeoffs before focusing on reaction time results.

2.4 Types of dependency relations

The type of dependency relations exhibited between two dimensions is “thought to be due to a hierarchical relation between those dimensions… The dimension's rank in the hierarchy determines how easy or how hard the dimension is to ignore” (Jerger, Martin, Pearson, & Dinh, 1995, p. 932). If two dimensions of a stimulus are found to be integral (i.e. exhibit Garner interference), there are several ways in which they can be integrated. Eimas, Tartter, & Miller (1981) outline this possibility space, giving three dependency relations, as they term them. I will adopt these terms with slight modification. If two dimensions have a mutual dependency relation, what I will refer to as having *mutual and symmetrical interference*, then “analysis of the information for one feature affects the analysis of the information for a second feature and conversely. Furthermore, the magnitude of the processing interaction or dependency is approximately equal in both directions” (p. 287). In other words, the amount of Garner interference is the same for the two tasks in mutual and symmetrical dimensions. Dimensions that show *mutual and asymmetrical interference* are characterized by “processing interactions in both directions, but the extent of the interaction is greater in one direction than the other” (p. 287). Put differently, for mutual and asymmetrical dimensions there is Garner interference overall, but the amount of interference is greater for one dimension than the other. Finally, dimensions that exhibit *asymmetrical interference* are termed as such when “analysis of one feature occurs independently of a second feature, but the analysis of the second feature requires information from the prior analysis of the first feature" (p. 287). In the case of asymmetrical pairs, there is only interference for one dimension. Both types of asymmetrical interference, mutual or otherwise, suggest that in some sense one dimension dominates the other in processing.

Another measure, often termed *redundancy gain*, consists of the difference between reaction times in the control block and the correlated block. If the presence of redundant information in one dimension actually helps with processing in the other dimension, then the reaction times in the correlated block will be faster than in the control block. As such, the presence of a redundancy gain is often interpreted as converging evidence that two dimensions are integrated. Further, redundancy gains in the absence of Garner interference may indicate that the dimensions in question have been effectively integrated into a new dimension (Flowers & Garner, 1971). However, it has been found that redundancy gains can exist for perceptually separable dimensions (Biederman & Checkosky, 1970)—dimensions where no Garner interference is found—so "the presence or absence of a redundancy gain should not therefore be considered a strong indication of the way in which the two stimulus dimensions are processed" (Green et al., 1997, p. 280). As such, the presence or absence of redundancy gain will be reported for all experiments in this dissertation but will not be discussed in detail.

2.5 What does it mean for two dimensions to exhibit Garner interference?

Determining what it means for any two dimensions to demonstrate Garner interference is not entirely straightforward. The Garner paradigm goes as far as demonstrating the interference (or other dependency relation) between dimensions, but what that means for the relationship between those dimensions in theories of processing has been up for debate. Indeed, the Garner paradigm has been criticized for its purely operational definitions of integrality and separability, lacking theoretical grounding (e.g., Kadlec & Townsend, 1992). Some researchers, however, have attempted to position the definitions in theoretical frameworks, attempts which will be described in the rest of this section.

One of the earliest proposals to explain interference claimed that if two dimensions show interference, it is almost as if they are not distinct dimensions at all, but instead a singular "blob" which must first be processed holistically (Garner 1974; Lockhead 1972, 1979): "The distinction phenomenologically being between dimensions which can be pulled apart seen as unrelated, or unanalyzable *[separable dimensions]*, and those which cannot be analyzed but are somehow perceived as single dimensions *[integral dimensions]*" (Garner & Felfoldy, 1970, p. 325). While the "blob" account may successfully describe the perceptual relationship between some dimensions, this account has not fared well in recent years; it particularly breaks down when applied to more abstract, learned dimensions such as indexical dimensions, or crossmodal dimensions found to be integral. A basic "blob" account would claim that participants who showed Garner interference in an auditory–visual crossmodal task, for example, were unable to distinguish what they saw from what they heard, which is not a particularly realistic proposal. The higher-level, indexical dimensions under investigation in the present dissertation are likewise hard to square with the "blob" account, as it is difficult to imagine a model of processing where language-being-spoken could not be analyzed separately from the gender of a talker, for example.

Another proposal regarding the processing implications of different dependency relations posits a continuum between integral and separable dimensions, with perceivers having varying amounts of access to the individual constituent dimensions, from less (integrated end) to more (separable end) (e.g. Smith & Kemler, 1978).

A different interpretation of integrality was proposed by Melara & Marks (1990b), which they believe better accounts for interference effects found in dimensions such as those pairs they termed corresponding, which are those that "may have a semantic-linguistic origin" (p. 413). These authors suggest that all multidimensional stimuli are processed by *attribute-level processing*, where individual dimensions (attributes) are coded according to a set of psychologically meaningful axes. While separable pairs of dimensions only undergo attribute-level processing, integral pairs of dimensions also undergo what they call *stimulus-level processing*, wherein the context of one dimension affects how the values of the other dimension are experienced. They give the example of classifying high and low pitch in the context of amplitude (which they refer to as loudness): "variation on loudness creates two perceptually different high pitches, a loud high pitch and a soft high pitch, and two perceptually low pitches, a loud low pitch and a soft low pitch" (p. 399). It is thus the existence of intraclass variation within a dimension that causes the slowing in reaction times for integral dimensions (if the dimensions were separable, there would be no intraclass variation and therefore faster performance), rather than the previously held view that integral dimensions are first processed holistically (i.e. like "blobs") and then analyzed separately (which leads to slower reaction times). A benefit of this type of account for the purposes of this dissertation is that it is not limited to low-level perceptual dimensions, but also can be successfully applied to learned, abstract stimuli, such as those belonging to linguistic and indexical dimensions.

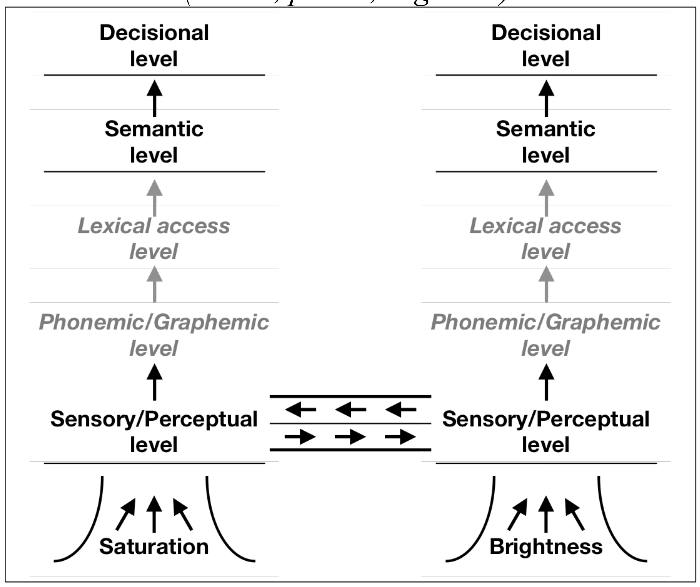
Another way to conceptualize the processing origins of effects in the Garner paradigm is to turn to what the selective attention literature refers to as *salience*. As described by Tong, Francis, & Gandour (2008, p. 702): "In selective attention, the degree of excitation of the target and the inhibition of distractor are determined by the salience of the dimensions (Melara & Algom, 2003). Salience refers to how much attention a stimulus or a dimension can capture in a certain task. Two dimensions compete for attention and greater salience of one dimension is achieved at the cost of weaker salience of the other. As a consequence, the greater the salience along the distractor dimension, the more difficult it is for observers to focus selectively on the target dimension." The salience of a dimension is a useful heuristic for discussing the relationship between two dimensions. Dimensional salience has been regarded as being independent of dimensional discriminability, an issue which will be discussed later in Section 2.9.

No matter what definition of integrality is used, however, one major criticism of the Garner paradigm remains: It cannot determine whether the dependency relation observed operates on a perceptual or on a decisional level (Kadlec & Townsend, 1992; Maddox, 1992). In other words, there is not a firm understanding of where in the time course of processing the dimensional integrality or separability occurs. Any commentary on when in processing a particular dependency relation may apply is based on speculation rather than on observable dependency relations from the Garner paradigm. Relatedly, it is not known whether the sequence of processing of two dimensions in relation to each other (i.e. are the two dimensions processed in serial, parallel, holistically, or some combination thereof?) can be inferred from those dimensions exhibiting a certain type of dependency relation, beyond what can already be surmised from the speed of processing of the two dimensions at baseline. Melara & Marks (1990c) explain: "…Garner interference is a limited diagnostic of dimensional interaction: Finding Garner interference may indicate that channels interact, but it does not indicate at what level(s) they interact." (p. 494).

2.6 Crosstalk between dimensional levels

Before undertaking the literature review, I now briefly detail an approach that schematizes the way in which dimensions with different relationships may interact, which will help in specifying the types of dependency relations observed in the review and in the experiments of this dissertation. The notion of *crosstalk* between dimensions, put forward by Melara & Marks (1990a), suggests that "when dimensions interact, channels *crosstalk* (cf. Pomerantz, Pristach, & Carson, 1989) at some level of information processing, causing a failure of selective attention to Dimension A" (p. 540). Their visual representation of this process is shown in Figure 2.6 below.

Figure 2.6. Illustration of crosstalk between dimensions. Adapted from Melara & Marks (1990a, p. 540, Figure B).



In this illustration, each dimension has a channel that flows from bottom to top (representing the temporal sequence of processing). If there is integration between dimensions, it is indicated with arrows connecting the two dimensions at the level of processing at which the crosstalk is thought to occur. If the integration is mutual and symmetrical (as is the case for saturation and brightness in the illustration in Figure 2.6), this is indicated by arrows flowing in both directions. For asymmetrical dimensions, the arrows would be unidirectional. Melara & Marks (1990a) argue that crosstalk can occur for dimensions whose levels are matched (what they term *within level*), as is indicated in the illustration in Figure 2.6, but that it may also occur for dimensions at different levels of processing (what they term *upstream* or *downstream* depending which dimension constrains the other and their relative levels). As was suggested by the physical relations approach to characterizing relationships between dimensions, dimensions processed later should be more affected by dimensions processed earlier (in this terminology, upstream crosstalk) than vice versa (downstream crosstalk). Despite this assumption, there do appear to be instances of both upstream and downstream crosstalk in the literature, as will be seen in the following review.

2.7 Review of speeded classification experiments in speech research

I now review previous studies that have used the Garner paradigm to investigate relationships between dimensions within the realm of acoustic speech stimuli. This review is organized by the type of information conveyed by the dimensions being tested. As described in Section 1.4, dimensions can be characterized according to what they convey to the listener, and in this dissertation I distinguish between linguistic, indexical, and non-linguistic dimensions. Because of the wide range of studies investigating linguistic dimensions, I subdivide linguistic dimensions here into *segmental* and *non-segmental* dimensions. Segmental dimensions convey information about phones (e.g. /b/ vs. /p/), whereas nonsegmental dimensions convey information above or below the phonemic level (e.g. tone 1 vs. tone 4 in Mandarin Chinese). In the review to follow, in certain cases I have standardized the authors' own label for the informational function of a dimension into the terminology used in this dissertation; for instance, if amplitude had been called *auditory* in a particular study, it is instead called *non-linguistic* here. Also, recall from Section 1.4 that a given dimension may convey different information in different settings. In this review, then, dimensions are labeled according to the type of information they are likely to convey in the particular situation of a laboratory study. For example, there are certainly instances where pitch and amplitude can convey indexical or linguistic information, but for the purposes of these studies, these dimensions are classified as non-linguistic. To continue with the example of pitch, it is not likely that pitch conveys linguistic of indexical information to participants in these experimental settings where they are asked to classify stimuli (often synthetic vowels or CV syllables) according to pitch using the response choices “high” or “low”. Thus, pitch is described here as conveying non-linguistic information (while pitch used as lexical tone, of course, is described as a linguistic, non-segmental dimension).

The review begins by covering studies pairing two dimensions that convey segmental information (segmental–segmental, Section 2.7.1), then moves to studies testing a segmental dimension and a non-linguistic dimension (segmental–non-linguistic, Section 2.7.2), then details studies pairing two non-linguistic dimensions or two non-segmental dimensions (non-linguistic–non-linguistic/non-segmental–non-segmental, Section 2.7.3), then a segmental dimension and a non-segmental dimension (segmental–non-segmental, Section 2.7.4), then a segmental dimension and an indexical dimension (segmental–indexical, Section 2.7.5), and closes with a study pairing a non-linguistic dimension with an indexical dimension (non-linguistic–indexical, Section 2.7.6). Following this review of the literature, I explore a claim made in many studies that the level of processing match or mismatch between dimensions can be used to predict the dependency relations observed for those dimensions (explained in detail in Section 2.8). To this end, the findings of the experiments in this literature review will be summarized in a comprehensive table (Table 2.1).

2.7.1 Segmental–Segmental

Many studies have employed the Garner paradigm to investigate the dependency relation between two segmental dimensions. Of these, some studies have found mutual and symmetrical integration between dimensions, and others have reported asymmetries.

Wood & Day (1975) paired consonant place of articulation (bilabial /b/ vs. alveolar /d/) and vowel (/a/ vs. /æ/) in the synthetic syllables /ba/, /da/, /bæ/, and /dæ/, and found mutual and symmetrical interference. However, Eimas et al. (1978; see also Eimas et al., 1981 for variations and extensions of these experiments) found mutual and asymmetrical interference across several permutations of the place of articulation and manner of articulation dimensions. Instead of comparing dimensions tied to two different segments of the syllable (in Wood & Day, one dimension varied on the consonant and the other on the vowel), these experiments tied both dimensions, place vs. manner of articulation, to the same segment, the consonant. Values within each dimension were varied across experiments (e.g. Experiment 1 used synthesized /Ca/ syllables, varying the place (bilabial /b/ and /m/ vs. alveolar /d/ and /n/) and manner (stop /b/ and /d/ vs. nasal /m/ and /n/) of the consonant, and Experiment 3 used resynthesized /Ca/ syllables, varying the place (front of oral cavity /b/ and /w/ vs. back of oral cavity /g/ and /j/) and manner (stop /b/ and /g/ vs. glide /w/ and /j/)). In all cases, interference was mutual and asymmetrical, and always such that manner of articulation interfered more when processing place of articulation than vice versa, differing from the results of Wood & Day (1975).[2]

[2] One difference between these two studies which to my knowledge has not been pointed out is the added task demands on the participant in Eimas et al. (1978)'s design, where subjects must respond based on dimensions unfamiliar to most people besides trained linguists, place and manner of articulation. Wood & Day's participants were able to respond to the segments themselves (/b/ or /d/?), while Eimas et al.'s participants had to undergo pretraining to learn which segments were bilabial and which were alveolar, for example. While it is unclear why

Tomiak, Mullennix, & Sawusch (1987) further examined the segmental–segmental pairing by varying the relationship that listeners had to the stimuli. Listeners classified noise-tone analogs of fricative-vowel syllables (containing no coarticulatory information), along the fricative (/f/ vs. /ʃ/) and vowel (/æ/ vs. /u/) dimensions, but were divided into two groups: one group was told that the stimuli were computer-produced sounds, while the other group was told that the stimuli were computer-modeled speech. When listeners were told the stimuli were speech, mutual and asymmetrical integration was found such that it was harder to ignore the vowel when processing the fricative than vice versa.[3] However, when listeners were told the sounds were non-speech, the dimensions were separable. Thus, listeners' understanding of the function of the dimensions being classified can have an influence on the dependency relation found between those dimensions.

Soli (1980) compared dependency relations between two segmental dimensions differing in the amount that the two were acoustically related. The dimensions that did not interact acoustically, place of articulation (bilabial /ba/ vs. alveolar /da/)[4] and voicing (voiced /ba/ and /da/ vs. voiceless /pha/ and /tha/), were found to be separable. The other pair, place of articulation (bilabial /ba/ vs. alveolar /da/) and vowel height (low /a/ vs. high /u/), which are acoustically related, were integral.[5,6]

---

such a difference may have caused responses to be asymmetrical rather than symmetrical, it is a notable difference between the two otherwise quite similar studies.

[3] A second experiment replicated the stimuli with naturally produced fricative-vowel syllables and mutual and symmetrical interference was found, as in Wood & Day, 1975.

[4] A slightly different methodology was employed in this series of experiments, using cues followed by target or catch trials, making the stimulus design a bit different.

[5] It is hard to tell from the analyses conducted, but it appears that the result was mutual and symmetrical.

[6] It should be noted that these results were obtained using a modified Garner paradigm, and very few participants (three) in each experiment, making the results not seamlessly comparable to the other work reported here.

2.7.2 Segmental–Non-linguistic

In pairing segmental with non-linguistic dimensions, experimenters have observed a diverse range of dependency relations. In perhaps the earliest demonstrations of the Garner paradigm applied to speech stimuli, Day & Wood (1972) tested the relationship between the segmental dimension consonant place of articulation (bilabial /b/ vs. alveolar /d/) and the non-segmental dimension pitch (high, 140 Hz vs. low, 104 Hz), and found a mutual and asymmetrical relationship between the two dimensions such that pitch interfered when processing place of articulation more than vice versa (upstream crosstalk). Other variations of this experiment by Wood found similar asymmetry. Wood (1974) replicated this mutual and asymmetrical result with a different consonant place of articulation pair (bilabial /b/ vs. velar /g/) compared with the same values of pitch in the context of a different vowel, /æ/. Wood (1975) in Experiment 1 found asymmetrical interference between the same dimensions tested in his 1974 study. Wood used these results to argue for the existence of a separate segmental (what he called phonetic) level in speech processing (despite his finding that the two types of dimensions were actually equated at baseline—see Table 2.1 below—which many would count as an indication that they were matched in terms of levels of processing).

Several subsequent studies used non-speech stimuli modeled after the speech stimuli used in the segmental vs. non-linguistic studies by Wood in order to argue that the existence of a separate segmental or phonetic level of processing is not required to find the same pattern of asymmetry. Blechner, Day, & Cutting (1976) tested synthetically created noise stimuli varying rise time (“pluck” vs. “bow” sound) and amplitude (loud vs. soft; difference of 7 dB), and found an asymmetry such that amplitude interfered more when processing rise time than vice versa.

Pastore et al. (1976) tested stimuli consisting of a tone pip (high, 3400 Hz vs. low, 1550 Hz) followed by a buzz (low, 600-1200 Hz vs. high, 2000-2600 Hz), and found asymmetrical interference such that buzz interfered when processing tone pip, but tone pip did not interfere when processing buzz. These studies were designed and their results interpreted in terms of the debate surrounding speech's "special" status in auditory processing, the details of which are not directly relevant for the purposes here. Of importance here is that these two non-speech studies revealed that it is problematic to assume that two non-speech dimensions will necessarily show symmetry. In other words, symmetry or asymmetry in processing may be related to properties of the stimuli themselves that are relevant auditory system-wide.

Melara & Marks (1990a, Experiment 3) also found asymmetry between the segmental dimension vowel (either /ai/ vs. /o/ or /i/ vs. /e/) and the non-linguistic dimension amplitude (loud vs. soft; intensities are reported in Amiga volume levels), but the asymmetry was in the opposite direction of Wood (1974, 1975), instead indicating downstream crosstalk. In this experiment, participants could not ignore the segmental dimension (vowel) when processing the non-linguistic dimension (amplitude), but did not find interference from the non-linguistic when processing the segmental. The authors were puzzled by this discrepancy.

In contrast to these asymmetrical findings, Miller (1978) found mutual and symmetrical interference between a segmental dimension, vowel, as compared with two different non-linguistic dimensions pitch and amplitude. Miller used the same non-linguistic dimension as Wood, pitch (low, 104 Hz vs. high, 140 Hz), and compared it to a different segmental dimension, vowel (/a/ vs. /æ/ following /b/) in synthesized syllables, with the idea that vowel may be more tied to non-segmental information in processing than consonant. A second experiment compared the non-segmental dimension amplitude (loud vs. soft; 20 dB difference) with vowel. In both experiments, mutual and symmetrical integration was found. In both experiments, mutual and symmetrical integration was found. Carrell, Smith, & Pisoni (1981) provided a more nuanced look at the existing picture by testing a range of values within dimensions previously tested (vowel versus pitch) differing in discriminability. This experiment will be reviewed in detail in the section on discriminability, below (Section 2.9).

2.7.3 Non-linguistic–Non-linguistic/Non-segmental–Non-segmental

As part of a series of experiments first introduced in the previous section, Wood (1975) also paired the non-linguistic dimension pitch (104 Hz vs. 140 Hz) with two other non-linguistic dimensions. First, pitch was paired with amplitude (loud vs. soft; difference of 20 dB between a reference level was calibrated per subject) in the synthetically produced syllable /bæ/. Another experiment in this series paired pitch with isolated second formant transitions distinguishing between places of articulation (/b/ rising vs. /g/ falling), which are perceived as nonspeech "chirps." Both of these studies found mutual and symmetrical interference.

Another study, Brunelle (2012), investigated three pairs of dimensions most accurately described as conveying non-segmental information in the Chad language. This study tested participants from three dialects of the Austronesian language Cham on the integrality of three non-segmental dimensions by using three pairwise comparisons. Using synthetic stimuli made to resemble a mid-front rounded vowel (a sound not present in Cham), the three pairings Pitch–F1[7],

[7] In this study, F1 is not used as a measure of a segmental distinction, but rather as part of a cluster of properties used to describe a difference in registers between dialects of Cham. Thus, it is classified as a non-segmental rather

Pitch–Voice quality, and F1–Voice quality all showed mutual and symmetric integrality (with some variations across dialects having to do with the inclusion of congruence measures, which are not relevant here). Further, when classifying the least discriminable dimension, voice quality, some subjects underwent a dimension shift, wherein they inadvertently began to use the more discriminable dimension for classification, another sign of the potential influence of discriminability on this task. (Discriminability will be discussed further in Section 2.9.)

2.7.4 Segmental–Non-segmental

A number of experiments that examine pairings of segmental and non-segmental dimensions are centered around the role of contrastiveness in dependency relations. These studies tested listeners from different language backgrounds on segmental–non-segmental pairs of dimensions that are either contrastive or not for the different listener groups (in these studies, English or Mandarin). These studies propose that the linguistic function of dimensions (instantiated here as whether a dimension is contrastive for a listener of a given language) may modulate dependency relations between dimensions at segmental–non-segmental levels of processing. These studies will be reviewed in detail here due to their relevance to the examination of language background in the experiments of this dissertation.

Repp & Lin (1990) paired segmental dimensions considered contrastive by both groups of listeners (they tested both consonant, /b/ vs. /d/, and vowel, /a/ vs. /u/), with tone (variations in F0), a dimension which is contrastive in Mandarin but is not considered to be contrastive to

---

than a segmental dimension here. Likewise, pitch and voice quality convey linguistic information in Cham, which is why they are classified as non-segmental rather than non-linguistic information here.

English listeners (while pitch is certainly manipulated by English speakers, it does not operate contrastively on the level of the syllable, as it does in Mandarin). The tones used were either real tones present in Mandarin (high level/tone 1 and high falling/tone 4; all stimuli were thus valid lexical items of Mandarin Chinese), or else variations in pitch uncharacteristic of the citation forms of tones in Mandarin (low level and low rising-falling). Mutual interference was found for all pairs of dimensions for both listener groups. For English listeners, where the dimensions were mismatched for contrastiveness (segmental/contrastive segment–non-segmental/not contrastive tone), the interference was mutual and symmetrical for the consonant/tone task (differing with Wood's (1974, 1975) asymmetrical result for consonant vs. pitch). However, the interference was mutual and asymmetrical for the vowel/tone task such that it was harder to ignore vowel when processing tone, indicating downstream crosstalk (differing with Miller's (1978) symmetrical result for vowel vs. pitch). The authors attribute these discrepancies to the fact that their participants responded slower to tone than to segment decisions overall, meaning that tone was less discriminable than segment dimensions, making it more susceptible to interference (cf. Section 2.9, below for more on discriminability). For Chinese listeners, where the dimensions were matched in contrastiveness (the segmental dimension segment and the non-segmental dimension tone are both considered contrastive), the interference was mutual and symmetrical for all tasks. Singh, Lee, & Goh (2011) also tested Mandarin Chinese listeners on consonant (/b/ vs. /p/) and tone (rising/tone 2 vs. falling/tone 4), and found very similar results: mutual and symmetrical interference for Chinese listeners, but mutual and asymmetrical interference for English listeners such that it was harder for English listeners to ignore segment when processing tone than vice versa. Thus, these experiments presents at least partial evidence that the linguistic function of dimensions may moderate the influence of levels of processing on dependency relations.

Lee & Nusbaum (1993) also paired consonant (/ba/ vs. /da/) with tone (either Mandarin Chinese tones: low rising/tone 3 vs. falling/tone 4, or non-Mandarin flat pitches: 104 Hz vs. 140 Hz) in synthetically produced syllables, in effect a combination of the stimuli of Repp & Lin (1990) and Wood (1974, 1975). For English listeners, where the dimensions were mismatched in terms of contrastiveness (segmental/contrastive consonant–non-segmental/not contrastive tone), there was mutual and symmetrical interference between consonants and Mandarin tones, but asymmetrical interference between consonants and non-Mandarin tones (flat pitches) such that variation in non-Mandarin tone interfered with consonant classification but variation in consonant did not interfere with non-Mandarin tone classification, indicating unidirectional upstream crosstalk). While potentially demonstrating a role for contrastiveness, the divergence of these results across types of tone are puzzling: Why should English listeners show greater interference from Mandarin-like tones than flat pitches when English does not use any type of tone contrastively at the syllable level? The authors attribute the difference to the fact that English listeners never encounter speech with a flat pitch, but are very sensitive to the linguistic role that varied pitch can play in English (e.g., intonation). For Mandarin listeners, where the dimensions were matched in contrastiveness (the segmental dimension segment and the non-segmental dimension tone could both be considered contrastive), there was mutual and symmetrical interference for both stimulus sets. Again, this may partially confirm that dimensions' linguistic function for listeners may affect the dependency relations between them.

Tong et al. (2008) extended the work of Repp & Lin (1990) and Lee & Nusbaum (1991), using a modified Garner paradigm which simultaneously compared three stimulus dimensions, consonant (/b/ vs. /d/), vowel (/a/ vs. /u/) and tone (rising/tone 2 vs. falling/tone 4) using synthesized syllables. This study only used native Mandarin-speaking subjects, for whom both segments and tone are both contrastive. The results showed mutual and asymmetrical integration for all pairs of dimensions, such that it was harder to ignore segmental than non-segmental dimensions: for the tone–consonant and tone–vowel pairings, consonant and vowel interfered more with tone than vice versa, and for the consonant–vowel pairing, vowel interfered more with consonant than vice versa.

One final study in this group paired the segmental dimension consonant with a different type of non-segmental dimension, phrase boundary, without a language background manipulation. Nakai & Turk (2011) had participants listen to two-syllable nonce phrases (e.g., /gʌdlɪdʒ/, /gʌglɪdʒ/, /gʌd#lɪdʒ/, /gʌg#lɪdʒ) and classify the coda consonant of the first syllable as /d/ or /g/, and whether there was a phrase boundary (e.g., /gʌd#lɪdʒ/) or no phrase boundary (e.g., /gʌdlɪdʒ/). There were two sets of stimuli such that one set had only one cue to the presence or absence of phrase boundary (preceding vowel duration), while the other set had two cues (both duration and F0 of preceding vowel). For both stimulus sets, mutual and symmetrical interference was shown. Interestingly, there was more interference overall when there was only a single cue to phrase boundary, the less discriminable version of that dimension. This result could not be accounted for by the discriminability account which will be described in Section 2.9, below. The authors use these results to support their hypothesis that the presence of multiple cues to signal prosodic information facilitates the processing of temporally overlapping segmental and non-segmental material.

2.7.5 Segmental–Indexical

In the first study to bring in indexical dimensions to the Garner task, Mullennix & Pisoni (1990) paired the segmental dimension initial consonant (/b/ vs. /p/) with the indexical dimension talker gender (male vs. female). Participants showed mutual and asymmetrical interference such that gender interfered more with the processing of consonant than vice versa; it was harder to ignore the indexical dimension than the segmental dimension, indicating downstream crosstalk. Jerger et al. (1993) replicated this result across age groups ranging from 3 years to 79 years of age.

Similarly, an experiment of Green et al. (1997) compared the segmental dimension consonant (/b/ vs. /p/ preceding the vowel /i/) with the indexical dimension gender (male vs. female). They found asymmetrical interference, where gender interfered with consonant but consonant did not interfere with gender, indicating unidirectional downstream crosstalk. This finding squares with Mullennix & Pisoni (1990) in that the indexical dimension was harder to ignore, but Mullennix & Pisoni found mutual interference where Green et al. did not. This discrepancy may have resulted from methodological differences: Green et al. used a between-subjects design across tasks, whereas Mullennix & Pisoni used a within-subjects design, and Green et al. used meaningless syllables, while Mullennix & Pisoni used real words.

Another experiment of Green et al. (1997) paired two different segmental dimensions, consonant voicing (voiced /b/ vs. voiceless /p/ preceding the vowel /i/) and consonant place of articulation (bilabial /b/ vs. alveolar /d/ preceding the vowel /i/), with the indexical dimension speaking rate (naturally produced “fast” vs. “slow”; tokens had a 0.46 fast-to-slow ratio on average), and found mutual and symmetrical interference for both pairs.

Kaganovich et al. (2006) paired the segmental dimension vowel (/ɛ/ vs. /æ/, no matrix consonant) with the indexical dimension voice (male talker 1 vs. male talker 2), but found symmetrical interference between the two dimensions. Similarly, Cutler et al. (2011) paired two segmental dimensions, vowel (Dutch /ɛ/ vs. /ɒ/) and consonant (Dutch /t/ vs. /s/), with the indexical dimension talker identity ("Peter" vs. "Thomas"), and found asymmetrical interference, such that it was harder to ignore the indexical dimension (upstream crosstalk). The rationale for this reversal in direction from Mullennix & Pisoni (1990) is attributed to differences in discriminability between dimensions, and is discussed further in Section 2.9. No rationale is given by the authors for their study's discrepancies with Kaganovich et al., who found symmetrical interference between the same dimensions.

Finally, a second experiment of the Singh et al. (2011) study tested English listeners and Chinese listeners on the segmental dimension consonant (/b/ vs. /p/) and the indexical dimension emotion (happy vs. sad; specific acoustic correlates not specified). In this experiment, Chinese and English listeners were not expected to differ in their dependency relations, as the relationship between emotion and consonant did not manipulate contrastiveness across listener language groups as in their first experiment, described above. However, the listener groups did differ, the Chinese listeners showing mutual and symmetrical interference and the English listeners showing mutual and asymmetrical interference such that it was harder to ignore emotion when processing consonant than vice versa.

2.7.6 Non-linguistic–Indexical

Only one study has tested the relationship between a non-linguistic dimension and an indexical dimension, Jerger, Pearson, & Spence (1999). This study found mutual and symmetrical interference between the indexical dimension gender (male vs. female) and the spatial location of the speech signal (right vs. left azimuth; speakers placed 45 degrees to the left and right relative to center of participant's head). This pattern of results held for all age groups tested (children between 4 and 10 years old, and adults).

Here it should be noted that the paucity of studies investigating the relationship between non-linguistic and indexical dimensions provides additional motivation for Experiment 3 of the current study, which tests the non-linguistic dimension amplitude and the indexical dimension language-being-spoken. Further, the complete lack of prior work comparing two indexical dimensions calls out for further research in this area. The present study tests two such pairs: language-being-spoken and gender of talker (Experiment 1), and language-being-spoken and talker identity (Experiment 2).

2.8 Are dependency relations between dimensions related to their relative levels of processing?

Having reviewed previous studies using the speeded classification paradigm, I now examine whether the observed dependency relations between dimensions could have been predicted by characteristics of those dimensions, namely their relative levels of processing. The idea behind what will be called the *levels of processing hypothesis* in this dissertation is that if two dimensions are matched in levels of processing, then they will show mutual and symmetrical interference, while mismatched dimensions will show a type of asymmetry (mutual or not), where the dimension processed earlier will interfere more with the dimension processed later than vice versa. Along these lines, some researchers have used observed dependency relations as evidence that two dimensions may or may not operate on the same level of processing. For example, Wood (1975) portrayed asymmetries observed between linguistic (segmental) and non-linguistic dimensions as evidence for a separate linguistic level of processing in the "speech is special" debate. Though such a correlation between levels of processing and dependency relations has been found by some researchers, others have not found a correlation, as will be demonstrated in Table 2.1, below.

While many researchers employing the Garner paradigm have attempted to relate dependency relations to levels of processing, the precise metric used to measure levels of processing is often left undefined, or particular levels of processing are assumed to be true a priori. The most common metric for evaluating the relative levels of processing of a pair of dimensions, however, is those two dimensions' respective processing speeds at baseline (in control blocks). If two dimensions show equivalent RTs in control blocks, they are thought to be processed at the same level (and thus the two dimensions demonstrate a levels of processing *match*). If two dimensions differ in processing speed, the dimension that is processed faster in control blocks (where there is no irrelevant interference) is thought to be processed earlier than the dimension processed more slowly under the same conditions (and thus the two dimensions demonstrate a levels of processing *mismatch*). As will be discussed in Section 2.9 below, this definition can become confounded with the notion of discriminability. Nonetheless, this approach to measuring the relative levels of processing of a pair of dimensions is widely used in the literature and will therefore also be used in this dissertation.

It should be noted that the levels of processing of dimensions can be, but are not always, correlated with the type of information conveyed by dimensions. As will be seen in Table 2.1, it does seem to be the case segmental dimensions are often (though not always) processed faster than indexical dimensions, for example. At the same time, dimensions conveying the same type of information my not always be processed at the same speed. For example, vowel and consonant are both linguistic, segmental dimensions, but this does not necessarily mean that they are processed at the same level (i.e. that their RTs in control blocks would be equivalent), as was demonstrated by Wood & Day (1975).

I now examine the studies reviewed in the previous sections with an eye toward levels of processing. In order to do so, I pull together the findings of all studies reviewed and display them in Table 2.1, below. In addition to the dependency relations observed, in this table I also present whether the two dimensions in each study were matched or mismatched in their processing speeds in control blocks. Taking these two pieces of information from each study, the relative speed of processing at baseline for the two dimensions (labeled "Levels of processing (RT in control) match?" in Table 2.1) and the type of dependency relation that held between the two dimensions (labeled "Observed dependency relation"), I can then determine whether the levels of processing hypothesis was confirmed (labeled "Levels of processing hypothesis upheld?"). Again, the levels of processing hypothesis is upheld (and "Yes" appears in the relevant column) under the following conditions: the two dimensions show equivalent processing speed in control blocks and they exhibit mutual and symmetrical interference, or one dimension is faster than the other in processing at baseline and they exhibit asymmetrical interference such that the faster dimension causes more interference. In all other cases, levels of processing and dependency relations are not correlated, and "No" appears in the relevant column.

It should be noted here that not all studies report statistical tests of dimensions' reaction times in control blocks. In these cases, "N.R." (for "not reported") is entered into the appropriate column, and the levels of processing hypothesis cannot be tested for these studies, denoted with an "N/A" in the appropriate column. The only other instances where the levels of processing hypothesis cannot be tested are the two studies reviewed (Tomiak et al., 1987; Soli, 1980) which found dimensions to be separable, which has not been thought to directly relate to levels of processing. Also, a few studies only reported dimensional differences between reaction times across all blocks rather than control blocks only as a measure of levels of processing differences, and, in those cases, the combined RTs were used as their measure of levels of processing in the table. The studies listed in Table 2.1 are in roughly the same order as they appeared in the literature review. Studies appear multiple times if they contain multiple experiments testing different pairs of dimensions, or if they test different populations of listeners. Further, dimensions are called Dimension 1 and Dimension 2 in the table simply as a way to standardize the presentation of dimensions across experiments.

Table 2.1. Summary of speech studies using the Garner paradigm evaluating the relationship between levels of processing and dependency relations.

Paper	*Dimension 1*			*Dimension 2*			*Results*			*Levels of processing hypothesis confirmed?*
	Dimension 1	*Values of dimension 1*	*Information type of dimension 1*	*Dimension 2*	*Values of dimension 2*	*Information type of dimension 2*	*Levels of processing (RT in control) match?*	*Observed dependency relation*	*If asymmetry, which dimension harder to ignore?*	
Wood & Day (1975)	Consonant	/b/ vs. /d/	Segmental	Vowel	/a/ vs. /æ/	Segmental	Vowel < Consonant	Mutual and symmetrical		No
Eimas et al. (1978)	Consonant (place)	Bilabial vs. alveolar and Bilabial vs. velar	Segmental	Consonant (manner)	Stop vs. nasal and Stop vs. glide	Segmental	Manner < Place (though manipulated discriminability)	Mutual and asymmetrical	Consonant (manner)	No
Tomiak et al. (1987)	Fricative	/f/ vs. /ʃ/	Segmental	Vowel	/æ vs. /u/	Segmental	=	Mutual and symmetrical when told sounds were speech		Yes
Tomiak et al. (1987)	Fricative	Same as above	Segmental	Vowel	Same as above	Segmental	Fricative < Tone	Separable when told sounds were noise		N/A
Soli (1980)	Consonant (place)	/b/ vs. /d/	Segmental	Consonant (voicing)	/b/ vs. /p/	Segmental	N.R.	Separable		N/A
Soli (1980)	Consonant (place)	Same as above	Segmental	Vowel	/a/ vs. /u/	Segmental	N.R.	Mutual (and probably symmetrical)		N/A
Day & Wood (1972)	Consonant	/b/ vs. /d/	Segmental	Pitch	High vs. low	Non-linguistic	N.R.	Mutual and asymmetrical	Consonant	N/A
Wood (1974, 1975)	Consonant	/b/ vs. /g/	Segmental	Pitch	High vs. low	Non-linguistic	=	Mutual and asymmetrical	Consonant	No
Miller (1978)	Vowel	/a/ vs. /æ/	Segmental	Pitch	High vs. low	Non-linguistic	=	Mutual and symmetrical		Yes
Miller (1978)	Vowel	/a/ vs. /æ/	Segmental	Amplitude	Loud vs. soft	Non-linguistic	=	Mutual and symmetrical		Yes

Carrell et al. (1981)	Vowel	2 of a 5-step continuum from /i/ to /ɛ/ depending on condition	Segmental	Pitch	2 of 3 possible pitches (70, 130, or 145 Hz) depending on condition	Non-linguistic	(Experimentally manipulated)	When equally discriminable, mutual and symmetrical; otherwise, asymmetrical	Pitch (When not equally discriminable)	Yes
Carrell et al. (1981)	Vowel	Same as above but preceded by /b/	Segmental	Pitch	Same as above	Non-linguistic	(Experimentally manipulated)	When equally discriminable, mutual and symmetrical; otherwise, asymmetrical	Pitch (When not equally discriminable)	Yes
Melara & Marks (1990a)	Vowel	Either /ai/ vs. /o/ or /i/ vs. /e/	Segmental	Amplitude	Loud vs. soft	Non-linguistic	=	Asymmetrical	Vowel	No
Wood (1975)	Pitch	High vs. low	Non-linguistic	Amplitude	Loud vs. soft	Non-linguistic	=	Mutual and symmetrical		Yes
Wood (1975)	Pitch	Same as above	Non-linguistic	Isolated F2 transitions	/b/ vs. /g/	Non-linguistic	=	Mutual and symmetrical		Yes
Brunelle (2012)	Pitch	Synthesized to resemble a mid front rounded vowel (in isolation)	Non-segmental	F1	Synthesized to resemble a mid front rounded vowel (in isolation)	Non-segmental	N.R.	Mutual and symmetrical for 3 dialects of Cham		N/A
Brunelle (2012)	Pitch	Same as above	Non-segmental	Voice quality	Same as above	Non-segmental	N.R.	Mutual and symmetrical for 3 dialects of Cham		N/A
Brunelle (2012)	F1	Same as above	Non-segmental	Voice quality	Same as above	Non-segmental	N.R.	Mutual and symmetrical for 3 dialects of Cham		N/A
Repp & Lin (1990)	Segment (Consonant or Vowel)	/b/ vs. /d/ or /a/ vs. /u/	Segmental	Tone	Tone 1 vs. Tone 4 or Low level vs. low rising-falling	Non-segmental	Segment < Tone	English listeners: mutual and symmetrical for consonant–tone but mutual and asymmetrical for vowel–tone	Segment (For vowel–tone pairing)	Yes
Repp & Lin (1990)	Segment (Consonant or Vowel)	Same as above	Segmental	Tone	Same as above	Non-segmental	Segment < Tone	Chinese listeners: mutual and symmetrical for all tasks		No

Lee & Nusbaum (1993)	Consonant	/b/ vs. /d/	Segmental	Tone	Tone 3 vs. Tone 4 and High vs. low flat pitches	Non-segmental	Consonant < Mandarin tones Consonant = non-Mandarin tones	English listeners: mutual and symmetrical for consonant–Mandarin tones but asymmetrical for consonants–non-Mandarin tones	Tone (for consonant–non-Mandarin tone pairing)	No, No
Lee & Nusbaum (1993)	Consonant	Same as above	Segmental	Tone	Same as above	Non-segmental	=	Chinese listeners: mutual and symmetrical interference for both		Yes
Tong et al. (2008)	Segment (Consonant and Vowel)	/b/ vs. /d/ and /a/ vs. /u/	Segmental	Tone	Tone 2 vs. Tone 4	Non-segmental	=	Chinese listeners: mutual and asymmetrical	Segment	No
Singh et al. (2011)	Consonant	/b/ vs. /d/	Segmental	Tone	Tone 2 vs. Tone 4	Non-segmental	=	English listeners: mutual and asymmetrical	Segment	No
Singh et al. (2011)	Consonant	Same as above	Segmental	Tone	Same as above	Non-segmental	=	Chinese listeners: mutual and symmetrical		Yes
Nakai & Turk (2011)	Consonant	/d/ vs. /g/ in coda position of syllable 1 of 2-syllable nonce phrases	Segmental	Phrase boundary	Phrase boundary vs. No phrase boundary, distinguished by either 1 cue or 2 cues	Non-segmental	N.R.	Mutual and symmetrical for both conditions. Less interference for multiple cue than single cue		N/A
Mullennix & Pisoni (1990)	Consonant	/b/ vs. /p/	Segmental	Gender	Male vs. Female	Indexical	=	Mutual and asymmetrical	Gender	No
Jerger et al. (1993)	Consonant	/b/ vs. /p/	Segmental	Gender	Male vs. Female	Indexical	Gender < Consonant	Mutual and asymmetrical	Gender	Yes
Green et al. (1997)	Consonant (voicing) and Consonant (place)	/b/ vs. /p/ and /b/ vs. /d/	Segmental	Gender	Male vs. Female	Indexical	=	Asymmetrical	Gender	No

Green et al. (1997)	Consonant (voicing) and Consonant (place)	Same as above	Segmental	Speaking rate	Fast vs. slow	Indexical	Consonant < Speaking rate (but only for slow rates)	Mutual and symmetrical		No
Kaganovich et al. (2006)	Vowel	/ɛ/ vs. /æ/	Segmental	Talker	Male 1 vs. Male 2	Indexical	=	Mutual and symmetrical		Yes
Cutler et al. (2011)	Segment (Consonant or Vowel), in Dutch	/t/ vs. /s/ or /ɛ/ vs. /ɒ/	Segmental	Talker	Peter vs. Thomas	Indexical	Segment < Talker	Mutual and asymmetrical	Segment	Yes
Singh et al. (2011)	Consonant	/b/ vs. /p/	Segmental	Emotion	Happy vs. Sad	Indexical	=	English listeners: Mutual and asymmetrical	Emotion	No
Singh et al. (2011)	Consonant	Same as above	Segmental	Emotion	Same as above	Indexical	=	Chinese listeners: Mutual and symmetrical		Yes
Jerger et al. (1999)	Spatial location of speech	Right vs. left	Non-linguistic	Gender	Male vs. Female	Indexical	Gender < Location	Mutual and symmetrical		No

Key > : Slower than = : Equal to < : Faster than
N.R. : Not reported N/A : Hypothesis cannot be tested

From this table, it is evident that levels of processing and dependency relations are not always correlated. If this were the case, the far right column would be full of “Yes” entries, indicating that levels of processing matches resulted in symmetry and mismatches resulted in asymmetry. Instead, a much more varied picture is presented, where approximately the same number of “Yes” and “No” entries appear in the table. Further, it does not appear to be the case that the relationship between levels of processing and dependency relations is stronger for certain types of dimensions more than others; the two are not correlated more in segmental–non-segmental pairs than in segmental–indexical pairs, for example.

Thus, there is little agreement across studies regarding whether dimensional integrality is related to level of processing matches or mismatches between dimensions. What this inconsistency means is still up for debate. Perhaps it means that reaction time measures at baseline are not the best measure of levels of processing. It may also reflect the limitations of reaction times as an interpretive tool in the assessment of this type of human behavior. Reaction times, without converging evidence from a different measure, cannot be decomposed into their component sources. The comparison of reaction times to two tasks cannot, for example, determine whether the faster reaction time was a result of a faster serial processing mechanism, or the result of several mechanisms operating in parallel.

Or, the lack of consistency may mean that a continuum may be a better theoretical construct to use than discrete levels of processing (e.g. while vowels and consonants are both informationally segmental, vowels may be functionally processed more similarly to tone/pitch than to consonants, and thus vowels should be in-between consonants and tone). Or, it could be that other factors such as the linguistic functions of dimensions (recall the discussion of the contrastiveness of tone for Mandarin versus English listeners in Section 2.7.4), or factors particular to specific pairs of dimensions, mediate whatever role levels of processing play in predicting dependency relations. Or, the lack of consensus among these results may mean that levels of processing cannot be used to predict dependency relations. Discussing the relationship between segmental and non-segmental dimensions in particular, Eimas et al. (1981) express optimism that the relationship between levels of processing (or dimensional characteristics more generally) and dependency relations may still be uncovered: “Although we are able to state that certain patterns of dependency are consistently obtained with particular combinations of features, we are not as yet able to extend this analysis further and describe the specific spectral (or perhaps functional) characteristics that yield a particular dependency relation. However, it is our belief that these descriptions exist and that continued research will yield these lawful specifications” (p. 306). The work reviewed here shows that as of this writing, such a relationship has still not been found. Nonetheless, relative levels of processing is a useful way to characterize relationships between dimensions, and predictions of the levels of processing hypothesis will be assessed in the experiments of this dissertation.

2.9 The role of discriminability

Related to the discussion of dimensional levels of processing is the issue of dimensional discriminability. One long-standing suggestion in the literature holds that classification decisions for less discriminable dimensions (those which may show longer RTs in the control block) may be more susceptible to irrelevant variation than more discriminable dimensions (e.g. Carrell et al. 1981; Cutler et al. 2011; Lee & Nusbaum, 1993). In other words, the task dimension that is more difficult shows more Garner interference. As described by Tong et al. (2008), the idea is that "if the values along one dimension are more discriminable than the other, small changes in the irrelevant dimension will not be detected. As a result, classification along the less discriminable dimension will be greatly affected by variation along the more discriminable dimension" (2008, p. 692).

On first blush, this account may seem to be merely a different way of stating the same mechanism responsible for the levels of processing account of differences between dimensions; a dimension deemed to be operating at a deeper level of processing than another dimension may also be called more discriminable. In fact, discriminability is often operationalized in the same way as levels of processing, and is typically measured using processing speed in control blocks. The difference between the two constructs is that discriminability can operate within a dimension (and thus within a level of processing). For example, the difference in pitch between a steady-state vowel at 70 Hz and at 145 Hz is a more discriminable difference than the difference between pitch at 130 Hz and 145 Hz of that same vowel. Presumably, this difference does not reflect that the dimension pitch changes the level on which it is processed, but rather only reflects a difference in the discriminability of particular values within the dimension (and level of processing). The role of discriminability in dependency relations, and the distinct influence of discriminability as separate from levels of processing, is best illustrated in the series of experiments carried out by Carrell et al. (1981), which explicitly manipulated discriminability within two dimensions.

Carrell et al. (1981) assigned listeners to one of five groups whose stimuli varied in the extent to which they were discriminable, the stimuli in each group chosen from a selection of values for each dimension (two vowels chosen from a synthesized 5-step continuum from /i/ to /ɛ/, versus two pitch settings chosen from 70, 130, or 145 Hz). Listeners in the large vowel-small pitch group, for example, would have encountered stimuli chosen from opposite ends of the vowel continuum, but close together in pitch. While the results were complex and varied, relevant here is the fact that vowels in isolation as paired with pitch showed mutual and symmetrical interference if the dimensions were matched in terms of discriminability. When the dimensions were mismatched in discriminability, however, there was an asymmetry such that it was harder to ignore vowel when processing pitch than vice versa. But, differences in the magnitude of discriminability between dimensions completely accounted for the asymmetry created, because "manipulations of the discriminability of the irrelevant dimension clearly affects the identification of pitch and vowel in the same way and to the same degree" (p. 6). The authors asserted that the asymmetry between isolated vowels and pitch was caused by discriminability rather than a true difference in levels of processing between the pitch and vowel dimension.

The authors' second experiment found a different relationship between discriminability and dependency relations for the same dimensions, but in a different context. This experiment made the synthesized stimuli of the previous experiment more speech-like by presenting the vowels following /b/. In general, the same results held (mutual and symmetrical interference when dimensions equal in discriminability, mutual and asymmetrical otherwise such that it was harder to ignore vowel than processing pitch than vice versa), but this time, "changes in the magnitude of irrelevant pitch variation affect[ed] vowel identification much more than changes in the magnitude of irrelevant vowel variation affect[ed] pitch identification" (p. 7-8). Because of the nature of the interaction with discriminability, the authors ascribe this difference to be a "true" asymmetry, reflective of the different levels of processing at which vowels (in a speech context) and pitch are classified. Thus, the authors argue that levels of processing and discriminability may exert independent effects.

The role of discriminability has been used by some studies to account for discrepancies between findings across studies. Cutler et al. (2011), for example, argued that their results (consonant and talker identity were asymmetrical such that it was harder to ignore the segmental than the indexical dimension) differed from Mullennix & Pisoni's (1990) results (consonant and talker gender were mutual and asymmetrical such that it was harder to ignore the indexical than the segmental dimension), due to the fact that talker is a "harder" (less discriminable) dimension, as compared with segment, than gender. The authors claimed that Mullennix & Pisoni found gender to be a more discriminable dimension than segment at baseline, and gender thus should and does exhibit more Garner interference (however, Mullennix & Pisoni actually found that the two were equally discriminable at baseline, making Cutler et al.'s argument somewhat problematic), thus attributing the difference to discriminability. A similar explanation has been made to account for discrepancies between Repp & Lin's (1990) and Wood's (1974, 1975) findings (for segmental−segmental dimensions). It should be noted, though, that for many of these studies discriminability is still confounded with levels of processing, as often only one set of values was chosen from within each dimension; thus, the difference between segment−talker and segment−gender, for example, may be better explained by levels of processing than discriminability.

Not all researchers have found that Garner interference can be attributable to dimensional discriminability, however. Eimas et al. (1978) tackled this issue directly by running a follow-up experiment on two dimensions that they had already found to exhibit mutual and asymmetrical interference (consonant place vs. manner of articulation). When they manipulated the previous stimuli so that the two dimensions became less discriminable, the same pattern of mutual and asymmetrical interference was still found, showing that discriminability could not account for the dependency relation exhibited between those two dimensions.

Overall, it can be seen that differences in dimensional discriminability, as was the case with levels of processing, are not straightforwardly related to dependency relations. Further, the matters of dimensional discriminability and levels of processing are often confounded. Whether observed dependency relations should be attributed to the relative discriminability between dimensions, or to the relative levels of processing of dimensions (or to other factors particular to the specific dimensions tested), is still unclear. In a review chapter, Garner himself discusses the difficulty with distinguishing between discriminability and levels of processing: "There is no easy resolution to this problem. Just because an equivalent result can be found with other dimensions that are unequally discriminable does not guarantee that the result found with any particular pair of dimensions is only due to the unequal discriminabilities. And it may even be, as Logan (1980) argues, that the differences in reaction times are due to differences in levels of processing. Discriminability or processing levels? The problem seems necessarily to remain with us for some time yet" (1983, p. 13).

Despite the lack of agreement in the literature on these issues, it is important to take them into consideration in the present study. Especially with the introduction of a new dimension to the Garner literature, language-being-spoken, it will be valuable to pair this dimension with other dimensions whose relationships to it are likely to differ. These considerations are detailed in

Chapter 3.

2.10 Summary

This chapter provided an overview of the speeded classification task, the primary experimental paradigm used in this dissertation. The theoretical and methodological concepts introduced in this chapter will be taken into consideration in the design, analysis, and discussion of the results of Experiments 1-3, which will be described in Chapters 3 and 4. The literature reviewed here underscores the importance of investigating the dimensions that will be tested in this dissertation, as no previous work has investigated the processing relationships between multiple indexical dimensions. Many of the concepts considered in this chapter will be called on in generating hypotheses for Experiments 1-3, and again in interpreting the results of these experiments.

CHAPTER 3: EXPERIMENTS 1-3, SPEEDED CLASSIFICATION

3.1 Chapter outline

In this chapter, I begin with an overview of Experiments 1, 2, and 3 of this dissertation. Then I re-introduce the dimensions being tested in these experiments in terms of their processing characteristics. I do this in order to provide a rationale for their inclusion in these experiments, and to make predictions about how each dimension will behave relative to language-being-spoken. Then, I discuss predictions as to the results of these three experiments based on various hypotheses derived from the diverse bodies of work reviewed in Chapters 1 and 2. First, I make predictions regarding the speed of processing of language-being-spoken at baseline as compared to the other dimensions tested based on previous results on the processing of these dimensions independently (reviewed in Chapter 1). Then, I make predictions about the dependency relations that will be observed in each experiment based on the match or mismatch in levels of processing between dimensions (as in work reviewed in Chapter 2). Next, I use the literature about the relative language-specificity of talker processing and talker-generality of language processing (reviewed in Chapter 1) to make alternative predictions about the language-being-spoken and talker pairing (Experiment 2), which may also extend to the language-being-spoken and gender pairing (Experiment 1). Finally, I bring in the role of listener language background and present a set of three predictions based on hypotheses generated from an examination of the bilingualism literature (reviewed in Chapter 1), two implicating more detailed representations of dimensions as the operative mechanism for language background differences, and one implicating language executive function as the operative mechanism.

In the next major section of the chapter, I describe the general methods and procedures used in all Garner paradigm experiments. Then, each experiment is detailed in turn, beginning with a review of its rationale and the predictions made by various hypotheses, followed by an overview of its individual methods and procedure, and closing with its results. The chapter ends with a summary of major results found in Experiments 1, 2, and 3.

3.2 Brief rationale for Experiments 1-3

Experiments 1-3 apply the speeded classification paradigm in order to determine the dependency relations between the indexical dimension language-being-spoken and other dimensions of the speech signal, across participants varying in language background. Experiment 1a tests the relationship between language-being-spoken and talker gender, also an indexical dimension, for English monolingual and Mandarin-English bilingual listener groups. Experiment 1b replicates Experiment 1a with a different pair of talkers to ensure generalizability of results, using English monolinguals as participants. Experiment 2 pairs language-being-spoken with another indexical dimension, talker identity. Talker identity differs from talker gender in several ways (reviewed in the following section) which may conspire to make talker identity harder to classify than talker gender. This experiment uses three different listener groups as participants: English monolinguals, Mandarin-English bilinguals, and non-Mandarin-English bilinguals. Finally, Experiment 3 pairs language-being-spoken with amplitude, a non-linguistic dimension, testing English monolinguals and Mandarin-English bilinguals. Amplitude is a more peripheral dimension which is likely to differ from the other dimensions tested in a variety of ways (reviewed in the following section).

3.3 Surveying dimensions: Language-being-spoken, talker, gender, and amplitude

I begin re-introducing the dimensions tested in this dissertation with a discussion of the pairings between the indexical dimension language-being-spoken and the other two indexical dimensions, talker gender in Experiment 1 and talker identity in Experiment 2. The dimensions chosen for Experiments 1 and 2 allow for better understanding how the language-being-spoken dimension might interact with two indexical dimensions that themselves differ with respect to how quickly they are processed. A similar configuration found in previous literature was described in Section 1.6.3; the dimensions gender and talker have both also been paired with segmental dimensions (consonant and/or vowel): gender was paired with segment by Green et al. (1997), Jerger et al. (1993), and Mullennix & Pisoni (1990); and talker was paired with segment by Cutler et al. (2011) and Kaganovich et al. (2006). The present set of experiments mirrors this setup but instead of being paired with a segmental dimension, gender and talker are here paired with language-being-spoken. The pairing of language-being-spoken with both gender and talker allows for an examination of dependency relations in two pairs of dimensions which both convey the same type of information (indexical–indexical), but whose level of processing match or mismatch likely differs across pairs.

While gender and talker both convey indexical information, they have a number of dissimilarities that position them differently with respect to the language-being-spoken dimension in terms of levels of processing. First, as pointed out in Section 1.5.1, classification of gender from speech relies on relatively few acoustic cues (F0 being the major one; Coleman, 1971) and occurs quite quickly, within a few tens of milliseconds following a consonant's release burst (Swartz, 1992). The multidimensional cues involved in performing talker classification (e.g. pitch, pitch range, rate, vocal quality, and vowel quality; Van Lancker et al., 1985, described in Section 1.5.2), however, make it a more complex process. Though the time course of talker classification is quite fast–listeners can distinguish between voices in less than 500ms (Andics et al., 2007)–this is still longer than the process of gender classification. The cognitive basis of language classification is comparatively less tested in the literature (as discussed in Section 1.5.3), but it is known that classification of speech by language depends on a variety of acoustic cues (e.g., rhythm, pitch patterns, distinctive segments, as well as talker voice characteristics, Muthusamy et al., 1994; Stockmal et al., 1996), and that experienced listeners can classify languages quite accurately (e.g., Bond & Fokes, 1991; Stockmal et al., 2000). However, these studies documenting language classification accuracy have not measured the time course of this process. Thus, it is not known exactly how quickly language-being-spoken can be classified as compared with gender and talker; this dissertation will be able to fill in this gap in the literature by comparing the reaction times in the control block for language-being-spoken classification with those for gender, talker, and amplitude.

Additionally, the set of possible genders is a closed set, consisting of only two choices, male and female (though cf. Mullennix et al., 1995), whereas talker is an open set, bounded only by the number of talkers that exist (though in the present study, it is confined to a binary choice between two male talkers). The set of possible languages is also an open set (though in this study also confined to a binary choice, between English and Mandarin Chinese), aligning language-being-spoken with talker along this metric.

Finally, in the course of speech comprehension, listeners must perform talker classification (or at least some type of talker calibration) and language-being-spoken classification in order to extract meaningful content from the signal. While listeners certainly can and do perform gender and amplitude classification, their roles in linguistic processing are arguably more incidental. Thus, again the talker and language-being-spoken dimensions are more matched, this time in the extent to which they are necessary for speech processing. Overall, then, between the two indexical properties, language-being-spoken appears to have more in common with talker than with gender.

Experiment 3 brings in amplitude, a non-linguistic dimension, to be paired with language-being-spoken. Amplitude was chosen as a way to explore the patterns of interdependency across a pair of dimensions that each convey a different type of information, and that likely differ in levels of processing to a larger degree. Wherever language-being-spoken fits in relation to other dimensions in terms of processing speed, it is likely to be closer to the indexical dimensions gender and talker than it is to amplitude, making amplitude an important complement to the other indexical dimensions tested here.

Amplitude's differences from gender, talker, and language-being-spoken are many. First, listeners in real world contexts may not need to form mental representations for amplitude, whereas forming representations for language-being-spoken, talker, and gender may all be relevant for navigating the everyday linguistic world (Foulkes, 2010), differentiating amplitude from the other dimensions in terms of its function.

Further, and perhaps more importantly here, amplitude has different processing properties than gender and talker. Amplitude has been shown to be more peripheral to processing than other dimensions of speech like talker and speaking rate (Bradlow et al., 1999; Nygaard, Sommers & Pisoni, 1995; Sommers et al., 1994). For example, listeners did not show a decrease in performance on L1 word identification when there was variability in the amplitude of those words (Sommers et al., 1994), while variability in speaking rate did cause a decrease. Similarly, in a continuous recognition paradigm, listeners did not show a decrease in performance for memory of a previously heard word when the amplitude of the word changed between the two repetitions, whereas a change in talker and a change in speaking rate did negatively affect recall performance (Bradlow et al., 1999). Explanations of these results have suggested a "phonetic relevance hypothesis", such that only variability in those dimensions in the signal that aid in phonetic identification will have an effect on speech memory and learning (Sommers & Barcroft, 2006). Changes in talker and speaking rate, for example, alter acoustic features like formant frequencies and speech timing characteristics (such as VOT) that serve to distinguish phoneme categories. While changes to amplitude may certainly affect speech perception accuracy (as amplitude can serve as a secondary cue to the distinction between stressed and unstressed vowels, and increased amplitude makes speech easier to hear in the first place), it is likely that amplitude is not retained in memory alongside the linguistic representation of speech in the same way that talker or speaking rate might be.

Finally, the amplitude dimension is a global acoustic feature in and of itself as compared with dimensions comprised of complex, multidimensional cues such as talker and language-being-spoken (and gender, though to a lesser extent), making it likely to be faster to classify at baseline.

3.4 Predictions

In this section I give an overview of the hypotheses and predictions for Experiments 1-3 stemming from the diverse bodies of literature reviewed in Chapters 1 and 2. These will be reviewed again briefly in the relevant section for each experiment individually.

3.4.1 Levels of processing of dimensions

The literature examined in Chapter 2 regarding the relationship between levels of processing of dimensions to the dependency relations between those dimensions will be used here to make predictions about dependency relations in the current experiments. Due to the fact that language-being-spoken is as of yet an untested dimension, it is not known how its level of processing will compare to the dimensions with which it is paired. Thus, predictions will first be made concerning the position of language-being-spoken at baseline relative to the other dimensions tested, which will be based on the literature reviewed in the previous section. These predictions about relative reaction time in control blocks will then be used to make a hypothesis about the dependency relations between dimensions, which will be referred to as the levels of processing hypothesis.

3.4.1.1 Predictions about dimensional differences at baseline

In terms of the type of information it conveys, language-being-spoken is conceived of in this dissertation as an indexical dimension alongside talker and gender (described in Section 1.4.1). However, the time course of language identification, and thus its level of processing in relation to each of the other dimensions tested here, is not yet known. Therefore, I now call on the background just reviewed in Section 3.3 to make predictions about the time course of language-being-spoken processing in relation to the other dimensions. Predictions will be summarized in Table 3.1, below.

Language-being-spoken processing appears to have the most in common with talker processing in that they both: are open sets, require the use of multiple acoustic cues, and are processes that are used by listeners in the course of speech comprehension. Thus, it is hypothesized that language-being-spoken and talker will be most similar in terms of time course (meaning that they are hypothesized to be on the same level of processing). This makes the prediction that reaction times in the control blocks of the talker task and the language-being-spoken task will be equal (in Experiment 2).

Amplitude was selected for use in Experiment 3 because it has less in common with language-being-spoken processing than the other dimensions have, as amplitude is a single, peripheral acoustic cue that requires no learning. As such, it is predicted that amplitude will be processed more quickly than language-being-spoken in control blocks (in Experiment 3), meaning that amplitude is hypothesized to be on a shallower level of processing than language-being-spoken.

It is likely that gender falls in-between talker and amplitude in terms of processing, sharing more with talker than with amplitude. Like language-being-spoken processing, gender processing also does not rely solely on one acoustic cue (though F0 is dominant). Unlike language-being-spoken, gender is a closed set. Gender, then, is predicted to be processed faster than language-being-spoken in the control blocks (Experiment 1), meaning that gender is hypothesized to be on a shallower level of processing than language-being-spoken, as well.

Table 3.1. Predictions regarding performance in control blocks based on prior work on the processing of dimensions in isolation.

Experiment	*Predicted reaction time at baseline*
Experiment 1: Language-being-spoken–Gender	Language-being-spoken > Gender
Experiment 2: Language-being-spoken–Talker	Language-being-spoken = Talker
Experiment 3: Language-being-spoken–Amplitude	Language-being-spoken > Amplitude
Key > : Slower than = : Equal to < : Faster than	

3.4.1.2 The levels of processing hypothesis

Although the review of the relationship between levels of processing and dependency relations in Section 2.8 revealed a rather inconsistent picture, I utilize levels of processing here as a way to make predictions about how language-being-spoken may be related to other dimensions. As a reminder, the levels of processing hypothesis (e.g., Melara & Marks, 1990a; Wood, 1975) suggests that two dimensions matched in levels of processing will have a mutual and symmetrical dependency relation, while dimensions mismatched in levels of processing will have some type of asymmetrical dependency relation (either mutual asymmetry or asymmetry). If there is a mismatch, the dimension processed more quickly at baseline (i.e. the dimension at the shallower level of processing) is predicted to be the one to cause more interference in the processing of the other.

In Experiment 1, gender is predicted to be faster than language-being-spoken in control blocks, making a mismatch in terms of levels of processing, which means that the two dimensions should have some type of asymmetrical dependency relation. Further, it is likely that gender will interfere more with language-being-spoken than vice versa because gender is hypothesized to be the dimension that is faster to process at baseline. Of course, if gender is not found to be faster than language-being-spoken in control blocks, then the levels of processing hypothesis will need to be refined to reflect the observed control block reaction times. As it stands, the prediction of the levels of processing hypothesis will be confirmed if there is Garner interference for both tasks in Experiment 1 but more interference in the language-being-spoken task than in the gender task.

Likewise, for Experiment 3, amplitude is also predicted to be faster than language-being-spoken in control blocks, another mismatch in terms of levels of processing, making the prediction that these two dimensions will have some type of asymmetrical dependency relation. Further, since amplitude and language-being-spoken are likely quite far apart in terms of their levels of processing, as described above, it is possible that the interference will not be mutual; language-being-spoken may not interfere with amplitude at all. In fact, the previous literature proposing that amplitude is not stored in memory with other indexical information (cf. Bradlow et al., 1999) might suggest that amplitude would be completely separable from other dimensions. However, given that most speech dimensions appear to be integral (e.g., out of the many studies reviewed in Chapter 2, dimensional separability was only found in two cases: for fricative and vowel when listeners were told the stimuli were non-speech noise in Tomiak et al., 1987; and for consonant place of articulation and voicing in Soli, 1980). Further, many non-speech auditory dimensions even appear to be integral (e.g. Blechner et al., 1976; Pastore et al., 1976). Thus, it is more likely that the dependency relation here will be some form of asymmetrical integrality. Of course, if amplitude is not found to be faster than language-being-spoken in control blocks, then the levels of processing hypothesis will need to be refined to reflect the observed control block reaction times. As it stands, the prediction of the levels of processing hypothesis will be confirmed if there is more Garner interference in the language-being-spoken task than in the amplitude task in Experiment 3.

Finally, in Experiment 2, talker is predicted to be a level of processing match to language-being-spoken based on processing time in control blocks, making the prediction that talker and language-being-spoken will have a mutual and symmetrical dependency relation. Of course, if talker and language-being-spoken are not found to be equal in control blocks, then the levels of processing hypothesis will need to be refined to reflect the observed control block reaction times. As it stands, the prediction of the levels of processing hypothesis will be confirmed if there is an equal amount of Garner interference in the talker and language-being-spoken tasks in Experiment 2.

Table 3.2. Predictions regarding dependency relations based on (predicted) levels of processing.

Experiment	*Predicted dependency relation*
Experiment 1: Language-being-spoken–Gender	Language-being-spoken ← Gender
Experiment 2: Language-being-spoken–Talker	Language-being-spoken ←→ Talker
Experiment 3: Language-being-spoken–Amplitude	Language-being-spoken ← Amplitude
Key X→Y: mutual and asymmetrical; X interferes with Y more than vice versa X←→Y : mutual and symmetrical X←Y : mutual and asymmetrical; Y interferes with X more than vice versa	

3.4.2 The relative language-specificity/talker-generality hypothesis (LS/TG)

The relative language-specificity/talker-generality hypothesis (LS/TG) is based on previous empirical findings about language-being-spoken and other indexical dimensions (most frequently, talker) in processing. From this hypothesis, another set of predictions about the dependency relations between these dimensions will be made. The discussion here centers on the pairing of language-being-spoken with the indexical dimension talker, the comparison at the heart of this dissertation. However, while most of the previous work investigating the relationship between language-being-spoken and other indexical dimensions focuses on the talker dimension, it is possible that these predictions for the language-being-spoken–talker pairing also extend to the language-being-spoken–gender pairing. This is the case for several reasons, discussed below. As such, the results of the gender–language-being-spoken pairing will also be assessed in relation to the LS/TG hypothesis.

The kind of predictions that can be made about the relationship between the language-being-spoken and other indexical dimensions can again be effectively modeled after what is known about the relationship between linguistic (particularly segmental) and indexical dimensions. The literature on linguistic and talker processing (reviewed in detail in Section 1.6) revealed that linguistic and indexical dimensions are integrated in processing, and sometimes that relationship is found to be asymmetrical. As was seen, linguistic processing is at least somewhat reliant on talker processing (e.g. the talker interference effect, Mullennix et al., 1989; Nusbaum & Morin, 1992; and the talker specificity effect, Martin et al., 1989; Nygaard et al., 1994; though cf. Van Lancker et al., 1988; Vitevich, 2003 for cases where linguistic processing can be done independently of talker information). Talker processing is perhaps less reliant on linguistic processing (e.g. talkers can be reliably identified even in a signal where linguistic information is absent or degraded, Van Lancker et al., 1985; Remez, Fellowes, & Rubin, 1997), though there is evidence that linguistic information is implicated in talker processing (Andics et al., 2007). These findings suggest that, when tested against each other directly in a speeded classification task, linguistic dimensions and the talker dimension (and perhaps indexical dimensions more broadly) will show Garner interference. Further, they suggest that this relationship may be asymmetrical, because linguistic processing may be less obligatory for indexical processing than vice versa. As was discussed in Chapters 1 and 2, previous Garner studies that directly tested the integration of a linguistic dimension and an indexical dimension are in line with these predictions. Most of these studies have shown asymmetrical interference (mutual or not) between these types of dimensions (gender−segment: Green et al., 1997; Jerger et al., 1993; Mullennix & Pisoni, 1990; talker−segment: Cutler et al., 2011), and one has shown symmetrical interference (talker−segment: Kaganovich et al., 2006). Of those exhibiting asymmetry, the direction of the asymmetry for all but one study (Cutler et al., 2011) was that it was harder to ignore indexical information when processing linguistic information than the reverse, consistent with previous research indicating that linguistic processing may be more contingent upon indexical processing than the reverse case. The direction of the Cutler et al. asymmetry was interpreted by the authors to be the result of a difference between dimensional discriminability (though levels of processing differences may instead be responsible; levels of processing and discriminability are confounded in this study). In sum, patterns of dependency relations found between linguistic and indexical (gender and talker) dimensions are consistent with previous work describing linguistic dimensions' role in indexical processing, and indexical dimensions' role in linguistic processing.

In this same vein, the previous literature can be used to make predictions about the pairing between indexical dimensions, in particular talker identity and talker gender, and the language-being-spoken dimension. While the picture is not as clear for this relationship, making the results of this dissertation an important complement to previous work, a well-formed hypothesis can still be developed. This hypothesis draws on three bodies of work−the role of language-being-spoken in talker processing (and in the processing of other indexical dimensions), the role of talker information (and other types of indexical information) in language-being-spoken processing, and the existence of language-specific phonetics–to propose that talker processing is relatively more language-specific than language-being-spoken processing is talker-specific. This proposal, described in detail below, predicts that in the talker–language-being-spoken pairing (Experiment 2) and perhaps also in the gender–language-being-spoken pairing (Experiment 1) there will be mutual and asymmetrical interference such that there is more interference from language-being-spoken when processing the indexical dimension than vice versa (summarized in Table 3.3, below).

In talker processing, the role of language-being-spoken has been relatively well-studied (refer to Section 1.7.1 for a detailed review), and the identification of talkers has been found to be at least partially reliant on language processing. (As noted, the role of language-being-spoken in gender processing has not been as well-studied, so for the present purposes results from talker processing will be generalized to gender processing as well, though a few studies related to gender are summarized at the end of this section.) When performing a talker identification task, it is harder for listeners to identify talkers speaking in an unfamiliar language than in a familiar one, a result known as the language familiarity effect (e.g. Bregman & Creel, 2014; Goggin et al., 1991; Köster et al., 1995; Perrachione et al., 2009; Perrachione & Wong, 2007; Thompson, 1987; Winters et al., 2008). These studies find that, while there is a benefit from language familiarity in talker identification, listeners can still identify talkers in unfamiliar languages at a rate above chance (or discriminate between them, as in Wester, 2012). This means that listeners are likely using both language-specific cues (because talkers in familiar languages have better rates of identification) and language-general cues (because listeners can generalize across languages and perform the task at all). The fact that idiosyncratic talker-specific characteristics exist that persist across languages (e.g. speaking rate in L1 is correlated with speaking rate in L2; Kim et al., 2013) is in line with the use of such characteristics as language-general cues in talker identification. However, the fact that language-being-spoken plays a role at all in talker processing, which is true given the robust effect of language familiarity in talker identification, language-being-spoken may interfere with the processing of talker.

In language-being-spoken processing, the role of talker and other indexical information has been less systematically studied (refer to Section 1.7.2 for a review of the existing work), but there are suggestions that the identification of language-being-spoken may also be at least partially reliant on talker processing. Language identification can be performed accurately even when the same talker is presented speaking in multiple languages (Stockmal et al., 2000). Although there is evidence that listeners use talker-general acoustic information in performing language identification (e.g., pitch patterns, rhythm, and segment; Stockmal et al., 1996), they may also be making use of talker-specific information (e.g., voice quality, speaking rate; Muthusamy et al., 1994; Stockmal, 1995; Stockmal et al., 1996). Thus, as was the case with talker processing using both language-general and language-specific cues, it appears that listeners use both talker-general and talker-specific cues in language processing. However, since talker information can be pointed to as a contributing factor at all in language identification, it is likely that talker will interfere when processing language.

Given these two bodies of work, a sensible hypothesis to make is that talker and language-being-spoken identification are reliant on each other, which predicts mutual interference in processing. The addition of another line of research documenting the language-specific organization of phonetic information (discussed in Section 1.8.4) serves to further refine this relationship, shifting the balance toward mutual asymmetry between these dimensions in processing. Evidence from speech production demonstrates the language-specific organization of phonetic information. Specifically, bilingual talkers employ a language-specific base of articulation for each language they speak. For example, English vowels are realized as more fronted than similar vowels in Spanish (Bradlow, 1996). Also, the distributions of VOT values distinguishing between voiced and voiceless stops are shifted in English relative to French, such that the value of a voiced stop in English may be equivalent to a voiceless stop in French (Lisker & Abramson, 1964). In other words, a bilingual speaker does not have just one setting with which they speak both languages, but instead utilize language-specific modes of speaking. Next, and more relevant here, listeners appear to make use of such distinctions. For instance, early Spanish-English bilinguals perceptually separate short-lag (Spanish) and long-lag (English) versions of /t/ (Flege & Eefting, 1988). Even newborns can distinguish between languages varying along a rhythm timing scale (Nazzi et al., 1998). Thus, listeners can and do make separate representations for multiple languages, even if the tokens come from the same talker. This evidence indicates that listeners abstract over talkers to form a representation of a language, as distinct from other languages. This point is rather self-evident, in fact: There must be global features of a language that persist across talkers in order for there to be multiple mutually intelligible speakers of the same language. While it was acknowledged earlier that there are also features of talkers that persist across languages (e.g. speaking rate), this point is rather less significant in light of the fact that it is a vital part of language use for listeners of a language to be able to abstract over information from many speakers of that language.

Given, then, the importance of talker-general, language-specific information, it is likely that language processing may not be as reliant on talker processing as the reverse. I now refine the hypothesis to reflect that proposal: While talker and language-being-spoken are likely to be mutually reliant on each other, it is likely that this relationship is imbalanced such that language processing may not rely as much on talker processing as talker processing relies on the processing of language. The prediction, then, is that talker and language-being-spoken will have a mutual and asymmetrical dependency relation, such that language-being-spoken will interfere more with talker classification than talker will interfere with language-being-spoken classification.

Though the work reviewed above concentrates on the relationship between language-being-spoken and talker processing specifically, there are indications that language-being-spoken and gender processing may have a similar relationship. For example, in an ABX task, listeners could match unknown languages even if the within-language stimuli were spoken by talkers of different genders, however, they were somewhat less accurate in language matching with different-gender talkers than within-gender talkers (Stockmal et al., 2011). Also, there is evidence that gender is conveyed language-specifically: Gender differences between average vowel formant values are not the same across languages (Johnson, 2006). For example, the difference between Danish men's and women's average vowel formants is much smaller than the difference between genders in Russian. Taken together, these studies are suggestive of the fact that the LS/TG hypothesis, while created on the basis of studies investigating the relationship between talker and language-being-spoken processing, may also extend to the relationship between gender and language-being-spoken processing.

The LS/TG hypothesis is not thought to extend to the language-being-spoken–amplitude experiment. Though previous work has not explicitly examined this pairing, it is not expected that amplitude would interact with language-being-spoken in a manner for any theoretically-motivated reason beyond what is accounted for by the levels of processing hypothesis.

Table 3.3. Predictions regarding dependency relations based on the relative language-specificity/talker-generality hypothesis.

Experiment	*Predicted dependency relation according to LS/TG hypothesis*
Experiment 1: Language-being-spoken–Gender	Language-being-spoken → Gender
Experiment 2: Language-being-spoken–Talker	Language-being-spoken → Talker
Key X→Y: mutual and asymmetrical; X interferes with Y more than vice versa X←→Y : mutual and symmetrical X←Y : mutual and asymmetrical; Y interferes with X more than vice versa	

3.4.3 Listener language background hypotheses

Finally, I make predictions for the current experiments based on listeners' language backgrounds. Bilinguals are able to interact with the language-being-spoken dimension in different ways than monolinguals (as detailed in Section 1.8), and as such, language-being-spoken's place in the dimensional hierarchy for the two groups is expected to differ. Here, I present three contrasting hypotheses based on previous literature, which lead to different predictions with respect to the language background of listeners. The hypotheses, which will be detailed in the following sections, are: (1) that language familiarity results in more detailed representations of dimensions, which serves to increase interference (the LF-DR hypothesis), (2) that bilingualism results in more detailed representations of dimensions, which serves to increase interference (the B-DR hypothesis), and (3) that bilingualism results in better selective attention, which serves to decrease interference (the B-SA hypothesis). Special attention will be paid to the results of Experiment 2, which was conducted on two different types of bilingual listeners (Mandarin-English bilinguals and non-Mandarin-English bilinguals), in order to distinguish between these possibilities.

3.4.3.1 Language familiarity enhances dimensional representations hypothesis (LF-DR)

The language familiarity enhances dimensional representations hypothesis (LF-DR) states that listeners familiar with the languages being spoken will have more detailed representations of both language-being-spoken and other indexical dimensions (namely, talker and gender), making the prediction that those listeners will show more interference between those dimensions. This hypothesis is targeted at language-being-spoken's relationship with the two indexical dimensions, gender and talker, because, similar to the reasoning above, it not predicted that amplitude classification would be dependent on listeners' language experience; there is not an empirically-driven reason to believe that listeners from one language background should differ in the amount of interference they experience between amplitude and language-being-spoken. In particular, the results of Experiment 2 are important for differentiating between this hypothesis and the hypothesis given in the following section, which rely on the non-Mandarin-English bilingual listeners to differentiate their predictions.

Many current models of speech processing suggest that listeners accumulate detailed exemplars in memory as they encounter speech (e.g. Goldinger, 1998; Johnson, 2006; Pierrehumbert, 2002). Representations are created not just for linguistic dimensions but for all relevant episodic details of the speech signal as well (especially indexical dimensions), as evidenced, for example, by the talker interference and talker specificity effects described in Section 1.6.1. Further, it has been suggested that talker identification within a familiar language is carried out by comparing speech from an incoming voice against those stored auditory representations of speech in that language (Perrachione et al., 2009; Perrachione & Wong, 2007; as discussed in Section 1.7.1). In terms of language familiarity, then, listeners unfamiliar with a language will not have access to the rich base of linguistic and indexical exemplars for that language as do listeners familiar with that language.

Further, I suggest that dimensions with more detailed representations will be harder to ignore when irrelevant to the task at hand (and will also be responded to more quickly when relevant to the current task), such that dimensions with more detailed representations will cause more interference. In a sense, experience with a dimension can be thought to render it more discriminable. Since familiarity with a language results in more detailed representations of talkers within that language (as evidenced by the language familiarity benefit), and since familiarity with a language by definition results in more familiarity with that language, both dimensions for listeners familiar with both languages (Mandarin-English bilinguals) have more detailed representation for the Mandarin stimuli than those same dimensions for listeners unfamiliar with one of the languages being tested (the English monolinguals, and non-Mandarin-English bilinguals in Experiment 2). That is, the English monolinguals and non-Mandarin-English bilinguals' unfamiliarity with Mandarin means that they will have access to less detailed representations than Mandarin-English bilinguals for half of the stimuli (the ones presented in Mandarin) in the language-being-spoken, gender, and talker classification tasks. This predicts that the Mandarin-English bilinguals will show more interference overall as compared with the

English monolingual and Mandarin-English bilingual listeners, as depicted in Table 3.4, below.

A previous study using this task provides evidence suggesting that familiarity with stimuli may affect patterns of interference. In the domain of face processing, Ganel & Goshen-Gottstein (2004) found an interaction between Garner interference and familiarity with stimuli. In this experiment, participants classified faces based on their expression (smiling vs. angry) or their identity (Person A vs. Person B). Familiarity with faces was manipulated between-subjects (whether the participants were in a psychology class taught by the two professors whose faces were used as Person A and Person B, or not). While there was asymmetrical Garner interference overall, more interestingly, the amount of Garner interference was greater for familiar faces than for unfamiliar faces; more familiar faces showed more perceptual integrality between expression and identity. The authors attributed the outcome to the fact that "representations of familiar faces contain richer and more detailed structural descriptions than representations of unfamiliar faces, for which only coarse, sketchy structural representations exist" (p. 586). These more detailed representations, in turn, make the dimensions more reliant on each other in processing and thus harder to selectively ignore.

In addition to more detailed indexical representations, another thing that listeners unfamiliar with a language do not have is access to the linguistic content of the signal. Even though the dimensions being tested here are not linguistic, it is known that indexical levels are integrated with linguistic levels (cf. Section 1.6). Furthermore, as discussed above, it is likely that linguistic processing is recruited when doing talker identification in a known language, and a proposal for the basis of the language familiarity effect rests on the ability to "compute the differences between the incidental phonetics of a specific vocalization and the abstract phonological representations of the words that vocalization contains" (Perrachione et al., 2011, p. 595), an ability that is unavailable to listeners when they cannot access the linguistic level of a language.

Previous studies comparing dimensional integrality between listener language groups have explicitly targeted linguistic dimensions. These studies have used the language background of the listener (either English or Mandarin Chinese) to vary the extent to which different acoustic cues are linguistically meaningful to the groups of listeners (e.g., Lee & Nusbaum, 1993; Repp & Lin, 1990, discussed in detail in Section 2.7.4). While it was not the case in these studies that listeners familiar with the language tested (Mandarin) always showed more interference than those unfamiliar with the language, this would not be predicted for the dimensions tested, which were differentially important to each listener group. Instead, it was found that the function of a dimension for a particular listener (in this case, whether it was contrastive or not) mediated the dependency relations between dimensions. Another study, Pallier, Cutler, & Sebastián-Gallés (1997), found a similar result for listeners across languages where stress placement is either highly predictable or not predictable from segmental content. In this study, the classification of two nonsense words ("deki" or "nusa") was interfered with by variability in the stress pattern of such forms (primary stress on first syllable versus on the second syllable)[8]. Importantly, the interference was greater for native Spanish listeners, in whose native language stress placement is not entirely predictable from segmental content, than it was for native Dutch listeners, for

---

[8] The reverse case, testing whether nonsense word classification interferes with stress classification, was not performed.

whom stress placement is correlated with syllable weight in their native language. These results again suggest that the function of a dimension for a particular listener may moderate the interference demonstrated. Similarly, Tomiak et al. (1986; explained in detail in Section 2.7.1) found that the dependency relations between dimensions were different depending only on whether listeners thought the stimuli were speech-related or not. Taken together, these previous results suggest that the perceived function of a dimension may affect how it is related to other dimensions in processing. In the present case, listeners who are able to access the linguistic content of stimuli may be afforded a different relationship to dimensions as a result of more detailed representations, in particular the language-being-spoken dimension, that is unavailable to listeners who do not understand the language.

The LF-DR hypothesis also makes secondary predictions about the particular stimuli that underlie the predicted difference between listener language groups. As previously mentioned, if language familiarity does result in more interference for Mandarin-English bilinguals than for English monolinguals and non-Mandarin-English bilinguals, it is because they are familiar with both languages while the other two groups are only familiar with English. This suggests that English monolinguals and non-Mandarin-English bilinguals would show more interference for stimuli presented in a familiar language (English) than an unfamiliar language (Mandarin Chinese). A difference in interference between stimulus languages for the two language groups unfamiliar with Chinese would count as further support that language familiarity plays a role in dimensional integration between talker and language-being-spoken.

Table 3.4. Predictions regarding dependency relations based on listener language backgrounds: Language familiarity enhances dimensional representations hypothesis.

Experiment	*Predicted difference between listener languages according to LF-DR hypothesis*
Experiment 1: Language-being-spoken–Gender (ENG & MAN)	MAN > ENG
Experiment 2: Language-being-spoken–Talker (ENG, MAN, & NMB)	MAN > ENG NMB
Experiment 3: Language-being-spoken–Amplitude (ENG & MAN)	MAN = ENG
Key > : More interference than = : Equal interference < : Less interference than ENG = English monolinguals MAN = Mandarin-English bilinguals NMB = Non-Mandarin-English bilinguals	

3.4.3.2 Bilingualism enhances dimensional representations hypothesis (B-DR)

The second hypothesis for language background performance, the bilingualism enhances dimensional representations hypothesis (B-DR) extends the previous hypothesis with respect to having detailed representations of dimensions to suggest that the experience of being bilingual enhances attention to indexical properties in general, which would then apply even when listening to an unfamiliar language. This hypothesis predicts that not just Mandarin-English bilinguals, but all bilinguals (including the non-Mandarin-English bilinguals in Experiment 2), will show more integration between dimensions than monolingual listeners, as given in Table 3.5, below. This hypothesis is meant to apply to the results of Experiments 1 and 2, and can only be differentiated from the LF-DR hypothesis with the results of the non-Mandarin-English bilinguals in Experiment 2.

The motivation for the B-DR hypothesis comes from the fact that detailed indexical representations of familiar languages may not just be a result of mental representations amassed from experience with the language, but instead may be useful cues to bilinguals when learning their languages. As reported in Section 1.8.1, it may be the case that simultaneous bilingual language learners scaffold onto supplementary cues in order to separate tokens of speech into distinct representations for each language. For example, there is evidence that bilingual infants “tag” speech tokens according to rhythm as a way to separate languages (Sundara & Scutellaro, 2011). While rhythm is not an indexical dimension, these findings are suggestive of the fact that learners may use other dimensions–potentially indexical ones–in organizing their two languages. Additional evidence for this claim comes from findings that adults learning two artificial languages make use of indexical cues, namely talker gender and identity, in order to separate speech tokens into groups according to language (Mitchel & Weiss, 2010; Weiss et al., 2009). These results would suggest that simultaneous bilinguals, at least when in the process of language learning, may have more of a reason to be “tuned in” to indexical features than do monolinguals. This increased attention to indexical features may persist long after initial language segmentation as a useful strategy for determining which language is being spoken. In fact, latching onto indexical features to aid in language classification is likely used by bilinguals of all types, regardless of whether they needed to use the strategy in language learning (i.e. simultaneous bilinguals). As an overly simplistic example, take a child who learned one language first at home but another language later at school. This child would be well-served to pay attention to who is speaking as a shortcut to determining what language is being spoken; if Mom is speaking, enter Language A mode, if my friend from class is speaking, enter Language B mode. Therefore, even though the bilingual listeners in this dissertation are mostly late learners of English (and not simultaneous bilinguals), they likely have a good deal of experience pairing indexical cues with language-being-spoken.

Further, an important complement to the potentially increased importance of indexical dimensions to bilinguals is the increased importance of language-being-spoken as a dimension to this listener group. Determining which language is being spoken is a decision more frequently made by listeners in multilingual settings, and it seems plausible that bilingual listeners would have more experience "tuning in" to the language-being-spoken dimension than would monolinguals.

As such, this hypothesis draws on the fact that bilinguals (not just those familiar with the specific language being spoken, but bilinguals in general) have more experience associating indexical (talker and gender) information with language information, and on the fact that all of these dimensions are likely more entrenched for bilinguals due to this experience, to predict that bilinguals will show more interference between dimensions. This is because, according to the same mechanism that was responsible for the previous hypothesis, more detailed representations would be harder to ignore when irrelevant to the task at hand, and would be responded to more quickly when relevant to the current task, creating more interference. Monolinguals, then, would not suffer as much interference because they are hypothesized to have less practice tuning into indexical and language-being-spoken dimensions. This is not to say that monolinguals could not adopt these strategies during the course of a laboratory experiment, but just that bilinguals have more experience with these strategies. As described above, this prediction holds for Experiments 1 and 2 only, as amplitude is not thought to be affected by language familiarity.

Table 3.5. Predictions regarding dependency relations based on listener language backgrounds: Bilingualism enhances dimensional representations hypothesis.

Experiment	*Predicted difference between listener languages according to B-DR hypothesis*
Experiment 1: Language-being-spoken−Gender (ENG & MAN)	MAN > ENG
Experiment 2: Language-being-spoken−Talker (ENG, MAN, & NMB)	MAN > ENG NMB
Experiment 3: Language-being-spoken−Amplitude (ENG & MAN)	MAN = ENG
Key > : More interference than = : Equal interference < : Less interference than ENG = English monolinguals MAN = Mandarin-English bilinguals NMB = Non-Mandarin-English bilinguals	

3.4.3.3 Bilingualism enhances selective attention hypothesis (B-SA)

The final hypothesis related to listener language background, the bilingualism enhances selective attention hypothesis (B-SA) brings to bear previous findings showing enhanced executive control for bilinguals, which predicts less interference overall for bilinguals (both Mandarin-English bilinguals and non-Mandarin-English bilinguals) as compared with monolinguals in the present experiments, as shown in Table 3.6, below.

Section 1.9 detailed several studies showing that bilinguals are better at many tasks involving executive function, an advantage thought to stem from their need to control both languages in speech production and comprehension. These advantages have been found not only in linguistic domains, but also in non-linguistic domains such as selectively attending to a simple pattern within a complex visual display (Bialystok, 1992), and inhibiting incongruous spatial information when identifying colors (i.e. Simon interference; Bialystok et al., 2004).

The Garner task is a test of selective attention, which falls under the domain of executive function. According to this hypothesis, bilinguals should be better at ignoring irrelevant dimensions and thus show less interference than monolinguals.

The B-SA hypothesis, in contrast to the previous two, also predicts differences between bilinguals and monolinguals in Experiment 3, the pairing of language-being-spoken and amplitude. If the difference between language backgrounds is predicted to be located entirely in the processing abilities of listeners, then this mechanism should operate equally over all pairings of dimensions.

Table 3.6. Predictions regarding dependency relations based on listener language backgrounds: Bilingualism enhances selective attention hypothesis.

Experiment	*Predicted difference between listener languages according to B-SA hypothesis*
Experiment 1: Language-being-spoken−Gender (ENG & MAN)	MAN < ENG
Experiment 2: Language-being-spoken−Talker (ENG, MAN, & NMB)	MAN < ENG NMB
Experiment 3: Language-being-spoken−Amplitude (ENG & MAN)	MAN < ENG
Key > : More interference than = : Equal interference < : Less interference than ENG = English monolinguals MAN = Mandarin-English bilinguals NMB = Non-Mandarin-English bilinguals	

3.5 Methods

Methodological considerations relevant to all experiments using the Garner paradigm (Experiments 1a, 1b, 2, and 3) are described in this section. Methods specific to individual experiments are explained in the appropriate section for each experiment.

3.5.1 Participants

Participants in all experiments had no history of uncorrected hearing or language impairment and had normal or corrected-to-normal vision. No person participated in more than one experiment of this dissertation.

3.5.2 Defining listener language groups

Following previous literature (e.g., Perani et al., 1998), for all experiments I classify participants as bilinguals (as opposed to monolinguals) depending on a variety of factors having to do with age of acquisition and amount of language use.

The monolingual English listener groups in all experiments of this dissertation were recruited from the Linguistics Department subject pool at Northwestern University. While these groups are referred to as monolingual here, participants in these groups have certainly been exposed to other languages over the course of their lives. However, these groups are heavily English-dominant. In order to be included in the monolingual listener group, participants must have reported that they learned English as their native language. Further, they must have reported that either they did not learn a language other than English before age 7, or if they did, that they currently use that language 10% or less of the time. Additionally, all participants in this group must have reported that they currently use English more than 85% of the time[9], and that they do not know Mandarin Chinese.

The bilingual Mandarin-English participants in all experiments of this dissertation were recruited from the greater Northwestern University community. In order to be included in the Mandarin-English bilingual group, participants must have reported that they learned Mandarin Chinese as their native language and English as a second (or third or fourth, etc.) language. They must have reported that they moved to an English speaking country from China only relatively recently (these data will be reported separately for each experiment). Length of residence (LOR)

[9] This cutoff for amount of English spoken applies to all participants in the monolingual groups except one participant in Experiment 3, who reported only using English 65% of the time. However, the language the participant reported using 30% of the time was only learned one year ago, and was given an average of 3.25/10 across a range of a self-reporting proficiency measures on that language, revealing a closer affinity to the monolingual group than to the bilingual group.

information for all experiments was calculated by subtracting the age at which a participant reported that they moved to an English-speaking country from their current age. Therefore, participants living in an English-speaking country for less than a year receive an LOR of zero. In the few cases where participants declined to report when they moved to an English-speaking country, this information was calculated from the year they reported moving to the Northwestern University area (Evanston/Chicago, Illinois).

The bilingual non-Mandarin-English participants in Experiment 2 were recruited from the greater Northwestern University community. In order to be included in this listener group, participants must have reported that they learned a language other than Mandarin Chinese as their native language and English as a second (or third or fourth, etc.) language, and that they do not know Mandarin Chinese. They must also have reported that they moved to an English speaking country only relatively recently, using the same criteria described for the bilingual Mandarin-English group, above.

Tables 3.7-3.9 summarize self-reported language information for each of the language groups in Experiments 1-4. Note that Experiment 4 is uses a different experimental paradigm and is not presented until Chapter 5, but its participants are included in these tables for the sake of comparison.

Table 3.7. Self-reported language background information given by English monolingual participants in Experiments 1-4.

Monolingual English listeners				
Experiment	*Amount of English usage*	*Number of participants who did not learn a language other than English before age 7*	*Number of participants who did learn a language other than English before age 7*	*Language-other-than English usage (of those participants who did learn language other than English before age 7)*
1a *(N = 18)*	94% (range 85%-100%)	11	7	5% (range 0%-9%)
1b *(N = 18)*	97% (range 94%-100%)	14	4	2% (range 0%-5%)
2 *(N = 18)*	97% (range 92%-100%)	10	8	2% (range 0%-10%)
3 *(N = 18)*	93% (range 65%-100%)	13	5	3% (range 1%-6%)
4 (Free classification; See Chapter 5) *(N = 19)*	97% (range 80%-100%)	18	3	0% (range 0%-0%)

Table 3.8. Self-reported language background information given by Mandarin-English bilingual participants in Experiments 1-4.

Mandarin-English bilingual listeners			
Experiment	*Amount of English usage*	*Age of acquisition of English*	*Length of residence in English-speaking country*
1a *(N = 18)*	29% (range 5-90%)	10.2 years old (range 6-15)	0.6 years (range 0-3 years)
2 *(N = 18)*	16% (range 1-80%)	10.2 years old (range 3-14)	0 years (range 0-0 years)
3 *(N = 18)*	33% (range 10-70%)	10.3 years old (range 6-14)	1.4 years (range 0-4 years)
4 (Free classification; See Chapter 5) *(N = 21)*	42% (range 10-90%)	8.6 years old (range 3-15)	1.7 years (range 0-6 years)

Table 3.9. Self-reported language background information given by non-Mandarin-English bilingual participants in Experiment 2.

Non-Mandarin-English bilingual listeners				
Experiment	*Amount of English usage*	*Age of acquisition of English*	*Length of residence in English-speaking country*	*Native languages represented*
2 *(N = 18)*	62% (range 30-96%)	8.9 (range 0-17)	2.4 years (range 0-9 years)	Spanish (4), Greek (2), Thai (2), Portugese (2), Korean (1), Croatian (1), Czech (1), Russian (1), Serbian (1), Farsi (1), Norwegian (1), Kannada (1)

3.5.3 Stimuli

3.5.3.1 Stimulus materials

Stimuli for all four experiments in this dissertation were taken from the Archive of L1 and L2 Scripted and Spontaneous Transcripts and Recordings (ALLSSTAR) developed at Northwestern University (Bradlow et al., 2010). As a part of a larger collection, this archive includes recordings of both monolingual English speakers and bilinguals reading a series of simple sentences originally developed for the Hearing in Noise Test (HINT, Soli & Wong, 2008), although the present experiments do not contain a noise manipulation. The following sample HINT sentences illustrate the length and level of grammatical and lexical complexity of the stimuli: “A boy fell from the window”, “The wife helped her husband”, and “Somebody stole the money”. The Mandarin version of these sentences was developed by Wong, Liu, Han, Huang & Soli (2007), and have comparable properties. From this corpus, a subset of HINT sentences spoken by three Mandarin-English bilingual speakers (two male, one female) was selected for use in these experiments, 64 spoken in English and 64 spoken in Mandarin. Information regarding stimulus recording procedures can be found in Bradlow et al. (2011). Acoustic characteristics of stimuli (duration and amplitude) are given in Section 3.5.3.3, below.

The present experiments are the first using the Garner paradigm to utilize sentences, or any units longer than two-word phrases (cf. Nakai & Turk, 2011), as stimuli. As such, a secondary goal of this dissertation will be to judge the efficacy of using longer stimuli in this task. One concern with using longer stimuli might be that they would afford too much processing time to participants, overshadowing any potential differences in response times between blocks. As will be discussed in Section 3.5.4, an attempt was made to offset this concern by instructing participants to respond as soon as they are able, emphasizing that they do not have to wait until the end of the stimulus to respond. However, stimuli are equal lengths across all blocks, so response tendencies for control blocks should not be any different than those for orthogonal blocks. Thus, the presence of Garner interference in any task of these experiments can be taken as evidence supporting the use of sentence-length stimuli in this experimental paradigm.

3.5.3.2 Stimulus talkers

Each experiment used a pair of talkers selected from the ALLSSTAR corpus (see Table 3.10). All talkers were late bilinguals with Mandarin as their L1 and English as their L2. Talkers completed Pearson's Versant test of spoken English proficiency as part of their participation in the ALLSSTAR project. The Versant test evaluates English proficiency in a variety of categories by having participants read and repeat sentences, answer questions, rearrange phrases into sentences, and retell a story. The test gives four subscores, two of which are relevant here: fluency ("measured from the rhythm, phrasing and timing evident in constructing, reading and repeating sentences", Pearson Education, Inc., 2008, p. 11) and pronunciation ("reflects the ability to produce consonants, vowels, and stress in a native-like manner in sentence context. Performance depends on knowledge of the phonological structure of everyday words as they occur in phrasal context", p. 11), as well as an overall score ("represents the ability to understand spoken English and speak it intelligibly at a native-like conversational pace on everyday topics. Scores are based on a weighted combination of the four diagnostic subscores", p. 11). Table 3.10 presents the results of these two subscores and the overall score for the three talkers used in this study. The overall score can range from 20 and 80 points. Based their overall scores, all talkers who produced the stimuli for this study fall into the "Independent user" level of English based on the Common European Framework global scale as reported by Pearson (Council of Europe, 2001, p. 24): 26-46 = Basic user, 47-68 = Independent user, and 69-80 = Proficient user. The potential role of English proficiency in this experimental design is considered in Section 3.6.1.

Table 3.10. Age and English proficiency information (as assessed by Versant) for talkers used in Experiments 1a, 1b, 2, and 3, and 4.

Talker	*Age*	*Versant: overall score*	*Versant: fluency score*	*Versant: pronunciation score*	*Experiments using this talker*
Male 1	25	50	54	49	1a, 2, 3, 4a, 4b
Female 1	22	65	80	69	1a, 1b, 4a
Male 2	23	59	77	61	1b, 2, 4b

3.5.3.3 Acoustic characteristics of stimuli

3.5.3.3.1 Duration of stimuli

Tables 3.11-3.14 below summarize the duration of the stimuli in each experiment. As was pointed out in Table 3.10, talkers are used in multiple experiments. For example, the male talker used in Experiment 1a (Male 1), is also used in Experiments 2 and 3; durations for that talker do not change across experiments as items were kept the same.

It should be noted that these durations represent all possible stimuli that a participant could have received from a particular talker, but no listener heard all 64 sentences from a particular talker speaking in a particular language. In actuality, in each block, a listener heard a pseudo-random sample of sentences spoken by each talker in each language within the specifications of the condition (as will be explained further in Section 3.5.3.4).

Tables 3.11-3.14. Durations for stimuli in each experiment, split by talker and language. Averages by talker (across languages) are given on the right-most column. Averages by language (across talkers) are given on the bottom row.

Table 3.11. Stimulus durations for talkers in Experiment 1a.

Experiment 1a: Average stimulus duration (in ms)				
Experiment	*Talker*	*Chinese (N = 64)*	*English (N = 64)*	*Talker average (N = 128)*
1a	*Male 1 (N = 64)*	1945.89	1554.11	1750.00
	Female (N = 64)	2097.28	1639.70	1868.49
	Language Average (N = 128)	2021.59	1596.91	

Table 3.12. Stimulus durations for talkers in Experiment 1b.

Experiment 1b: Average stimulus duration (in ms)				
Experiment	*Talker*	*Chinese (N = 64)*	*English (N = 64)*	*Talker average (N = 128)*
1b	*Male 2 (N = 64)*	1677.79	1567.48	1622.64
	Female (N = 64)	2097.28	1639.70	1868.49
	Language Average (N = 128)	1887.54	1603.59	

Table 3.13. Stimulus durations for talkers in Experiment 2.

Experiment 2: Average stimulus duration (in ms)				
Experiment	*Talker*	*Chinese (N = 64)*	*English (N = 64)*	*Talker average (N = 128)*
2	*Male 1, "Wei" (N = 64)*	1945.89	1554.11	1750.00
	Male 2, "Li" (N = 64)	1677.79	1567.48	1622.64
	Language Average (N = 128)	1811.84	1560.80	

Table 3.14. Stimulus durations for talker in Experiment 3.

Experiment 3: Average stimulus duration (in ms)				
Experiment	*Talker*	*Chinese (N = 64)*	*English (N = 64)*	*Talker average (N = 128)*
3	*Male 1 (N = 64)*	1945.89	1554.11	1750.00
	Language Average (N = 128)	1945.89	1554.11	

3.5.3.3.2 Amplitude of stimuli

All stimuli in Experiments 1a, 1b, and 2 were normed to 70 dB SPL using Praat (Boersma & Weenink, 2011), a comfortable listening level for listeners with normal hearing. The amplitude norming for stimuli in Experiment 3 (the amplitude–language-being-spoken pairing) is described in Section 3.8.4.3.2.

3.5.3.4 Stimulus selection and arrangement

Each participant completed three blocks (control, correlated, and orthogonal) for both dimensions of the stimulus. The blocks were grouped by stimulus dimension, so that a participant received all three blocks for one stimulus dimension (e.g. gender), and then all three blocks for the other stimulus dimension (e.g. language-being-spoken). The order of presentation of stimulus dimensions was counterbalanced across participants. The order of presentation of blocks was pseudo-randomized across participants; there were 6 different possible block orders a participant could receive. The order of the stimulus sets in the first task (e.g. the gender dimension) was held constant in the second task (e.g. the language-being-spoken dimension). See Figure 3.1 below for an overview of the order of blocks for two example participants, using the tasks of Experiment 1 as an illustration.

Figure 3.1. Block order for two sample participants in Experiment 1. The same protocol was followed in Experiments 2 and 3.

Example participant 1:	*Example participant 2:*
Block 1: Gender task, correlated block	Block 1: Language-being-spoken task, control block
Block 2: Gender task, orthogonal block	Block 2: Language-being-spoken task, correlated block
Block 3: Gender task, control block	Block 3: Language-being-spoken task, orthogonal block
Block 4: Language-being-spoken task, correlated block	Block 4: Gender task, control block
Block 5: Language-being-spoken task, orthogonal block	Block 5: Gender task, correlated block
Block 6: Language-being-spoken task, control block	Block 6: Gender task, orthogonal block

In each block of each task, participants heard all 64 stimulus sentences (produced in different talker-language configurations depending on the block), such that each participant heard all 64 items 6 times over the course of the experiment. Participants never heard the same sentence more than once in each block.

The following figure (Figure 3.2) outlines what stimulus dimensions were present in each block of each task, using the tasks of Experiment 1 as an example. Each participant was pseudo-randomly assigned to one of two control blocks for each task (for Experiment 1a/1b: Chinese (C) or English (E) held constant while classifying gender, and Male (M) or Female (F) held constant while classifying language-being-spoken), and one of two correlated blocks for each task (for Experiment 1a/1b: CF-EM or CM-EF for gender and CF-EM or CM-EF for language-being-spoken) so that an equal number of participants completed each. There is only one version of the orthogonal block, and therefore all participants had the same version. The number of sentences a listener heard from each of the talker-language combinations (e.g. male talker speaking in Chinese, female talker speaking in Chinese, etc.) is listed in each cell of the tables in Figure 3.2 below; the number of stimuli in each block in each task totals 64. As was mentioned in Section 3.5.3.3.1, the particular items presented by each talker-language combination varied across participants. For example, a listener in Experiment 1a in an orthogonal block would have heard 32 sentences by the male talker, 16 in Chinese and 16 in English, and 32 sentences by the female talker, 16 in Chinese and 16 in English. Each participant in that same block (Experiment 1a, orthogonal block) would have heard a different 16 items produced by each talker-language pair, as items were pseudo-randomly chosen from the pool of 64 items for each participant.

Figure 3.2. Number of stimuli presented to participants in each value of each dimension for each task in Experiment 1a or 1b. The same protocol was followed in Experiments 2 and 3.

Gender task:
Who is talking, Male or Female?

Orthogonal block		
	Chinese	*English*
Male	16	16
Female	16	16

Control block (C)		
	Chinese	*English*
Male	32	
Female	32	

or

Control block (E)		
	Chinese	*English*
Male		32
Female		32

Correlated block (1)		
	Chinese	*English*
Male	32	
Female		32

or

Correlated block (2)		
	Chinese	*English*
Male		32
Female	32	

Language-being-spoken task:
What language is being spoken, English or Chinese?

Orthogonal block		
	Chinese	*English*
Male	16	16
Female	16	16

Control block (M)		
	Chinese	*English*
Male	32	32
Female		

or

Control block (F)		
	Chinese	*English*
Male		
Female	32	32

Correlated block (1)		
	Chinese	*English*
Male	32	
Female		32

or

Correlated block (2)		
	Chinese	*English*
Male		32
Female	32	

As demonstrated in Figure 3.2 above, and keeping Experiment 1 as an example, the breakdown of stimulus dimension per block is as follows. In the orthogonal block, listeners heard 16 sentences produced by each of the four gender–language possibilities (16 by the male speaker in Mandarin, 16 by the male speaker in English, 16 by the female speaker in Mandarin, 16 by the female speaker in English). In the correlated and control blocks, the number of stimuli coming from each value of a dimension must be doubled in order to equal the same number of stimuli as in the orthogonal block. Thus, in the correlated block, listeners heard 32 sentences produced by each of two gender–language combinations (they either heard 32 of the male speaking Mandarin and 32 of the female speaking English, or 32 of the female speaking Mandarin and 32 of the male speaking English). In the control block for the gender classification task, listeners were assigned to one of two control groups, and therefore either heard all 64 sentences read in Mandarin or all 64 read in English, each receiving 32 by the male talker and 32 by the female talker. In the control block for the language-being-spoken classification task, listeners were assigned to one of two control groups, and therefore either heard all 64 sentences read by the male talker or all 64 read by the female, each receiving 32 in Mandarin and 32 in English. In each block, participants thus heard 64 sentences in each of 3 blocks (control, correlated, and orthogonal), making 192 sentences per task (e.g. gender classification and language-being-spoken classification in Experiment 1), and 384 sentences after completing both tasks.

3.5.4 Procedure

Listeners were seated in a sound attenuated booth equipped with a Mac Mini running Superlab 4.5, connected to Sony MDR-V700 headphones and a Cedrus RB-730 button box. Before the start of the experiment, each participant completed a web-based demographic language background questionnaire (Northwestern University Subject Database, or NU-subDb). After the completion of the questionnaire, listeners received instructions for the first part of the experiment, a screenshot of which is shown in Figure 3.3 for Experiment 1.[10]

[10] In some studies early on in the use of this paradigm, explicit information was given to participants regarding the nature of the variation in the blocks they were to encounter, particularly the presence of redundancy in the correlated

Participants first heard 16 practice trials with feedback at the beginning of each task before moving on to the main part of the experiment. Following the practice trials, the following screen appeared: “Practice trials complete. There will be 3 main sections of this part of the experiment, with a short break in between each section. We will now begin the experiment. You will not receive feedback on your answers from now on.” Participants then pressed a button to advance to the first block.

Stimulus sentences were presented one at a time over headphones set to a comfortable listening level, with a 50 ms inter-stimulus interval. Participants were given verbal instructions to respond using the buttons on the button box as soon as they knew the answer; they did not need to wait until the sentence was over before responding. The experimental software was set to automatically advance to the next trial and label the trial “no response” if the participant did not respond within 4000 ms of the onset of the stimulus. The order of the buttons was counterbalanced across participants.

After each block, a screen with the following message appeared: “[The section] is complete. The instructions for the next section remain the same.” Participants then pressed a button to advance to the next block. Between each task, participants were given a slightly longer break, during which they moved to a different sound-attenuated booth with an identical set up, received instructions appropriate to the task for the new stimulus dimension, and completed the second task. All participants completed the language background questionnaire and experiment in less than 1 hour.

---

block (e.g. Wood, 1974; Blechner et al., 1976; Felfoldy & Garner, 1971). No such instructions were given in the experiments in this dissertation; the opposing dimension was mentioned, but no reference was made to differences between blocks. See Section 3.6.5 for complete instructions given to participants.

3.6 Experiment 1a and 1b: Gender–Language-being-spoken

3.6.1 Experiment 1: Rationale

Experiment 1a tested the relationship between language-being-spoken and talker gender, for English monolingual and Mandarin-English bilingual listener groups. Experiment 1b was designed as a replication of Experiment 1a using a different pair of talkers—the same female from Experiment 1a paired with a different male talker—in order to discount the possibility of talker-specific effects and to test for a role of relative matches in talker L2 proficiency. Only one listener group, English monolinguals, completed Experiment 1b. While variability in L2 proficiency across bilingual talkers is always present, the particular concern in this design was that listeners might latch onto a separate dimension of the talkers' voices and use it in addition to or instead of gender cues in order to perform gender classification. As was pointed out in Table 3.10, the male and female talker used in Experiment 1a are not equally proficient in English (the female talker is more proficient than the male talker, as measured by the Versant test). It is a possibility, then, that English proficiency would become an additional dimension by which listeners could distinguish the voices, making the voices more discriminable and thus the gender decision easier to make. It could also be the case that proficiency might interact with the language-being-spoken dimension; the language-being-spoken decision may be harder if English stimuli spoken by one talker are more English-like, and English stimuli spoken by the other talker are less English-like.

In order to make sure that second-language proficiency was not involved in either the gender or language-being-spoken classification decisions, a different male-female talker pair was tested in Experiment 1b. The male voice used in Experiment 1b is slightly more proficient than the male voice in Experiment 1a, making it a more even proficiency match to the female voice, though the female voice is still more proficient than the male voice in Experiment 1b. Comparing the results of English listeners across Experiments 1a and 1b will allow for the examination of the potential role of talker L2 proficiency in the way listeners performed the task.

3.6.2 Experiment 1: Hypotheses and predictions

As described above, there are several predictions for the results of Experiment 1. First, in terms of processing speed at baseline, gender is predicted to be faster than language-being-spoken. Next, in terms of dependency relations, several predictions hold. Following the levels of processing hypotheses, language-being-spoken and gender are predicted to show mutual and asymmetrical interference such that gender interferes more with language-being-spoken than vice versa (if predictions about processing speed at baseline are confirmed). If gender turns out not to be faster than language-being-spoken in control blocks, the levels of processing hypothesis will be refined to reflect the observed levels of processing. The relative language-specificity/talker-generality hypothesis (LS/TG) may also extend to the gender–language-being-spoken pairing, which predicts mutual and asymmetrical interference between language-being-spoken and gender such that language-being-spoken would interfere more with gender than vice versa. With regard to listener language backgrounds, Mandarin-English bilinguals are hypothesized to show more interference overall than English monolinguals following both of the hypotheses related representations of dimensions (stemming either from language familiarity, the LF-DR hypothesis, or bilingualism, the B-DR hypothesis; the two are not distinguishable in this experiment, as non-Mandarin-English bilinguals are not tested), but the opposite is predicted to occur (less interference for Mandarin-English bilinguals relative to English monolinguals) under the B-SA hypothesis.

3.6.3 Experiment 1: Participants

3.6.3.1 Experiment 1a: Participants

The monolingual group of listeners in Experiment 1a consisted of 18 English-speaking Northwestern undergraduates. These participants were recruited using the Northwestern Linguistics Department subject pool, and were given partial course credit for their time. The monolingual group was comprised of 7 males and 11 females, with the mean age 19.4 (range 17-22). Participants in this group reported using English 94% of the time on average (range 85%-100%). Eleven of these participants did not learn a language other than English before age 7. The seven participants who did report learning a language before age 7 currently use that language on average 5% of the time (range 0%-9%). No participant in this group reported knowing any Mandarin Chinese. Four additional participants were run but their data were excluded unanalyzed because they did not conform to the language background requirements of the monolingual group.[11]

[11] Early on in this project, the goal was to have 36 total monolingual English listeners participate in this experiment (as well as in Experiments 1b and 2), and as such additional participants were run in an attempt to add an additional 18 to the existing 18 participants described above. Eighteen participants were needed to complete the quasi-Latin square factorial design, whereby task order, button order, and block order, correlated option, and control option were counterbalanced across participants. In the course of attempting to add this second group of 18 participants, 24 additional participants were run, some of which had knowledge of Mandarin (4), did not conform to the language background requirements of the monolingual group (i.e. they were bilingual in languages other than Mandarin) (1), or experienced technical glitches (2). When the decision was made to only include 18 participants in all listener groups of all Garner paradigm experiments of this dissertation, the data of all 24 of these participants were excluded, unanalyzed, because any number of participants besides multiples of 18 would result in an imbalance on one or more counterbalanced properties of the sample (e.g., more participants would have received only English stimuli than only Mandarin stimuli in the control block of the gender task).

The 18 Mandarin-English bilingual participants who participated in Experiment 1a were recruited both from the International Summer Institute (ISI) at Northwestern in 2012 and from the greater Northwestern community. These participants received compensation in cash for their time (less than 1 hour). Of these 18 Mandarin-English bilingual participants, 8 were male and 10 were female, with a mean age of 24.2 (range 21-29), though one participant declined to report his or her age. On average, the participants in this group reported learning English at age 10.2 (range 6-15), and report using English an average of 29% of the time (range 5%-90%). These participants have lived in an English speaking country for an average of 0.6 years (range 0-3 years).[12] Two additional participants were run but their data were excluded unanalyzed because they had trouble staying awake during the experiment.

3.6.3.2 Experiment 1b: Participants

The monolingual group of listeners who participated in Experiment 1b consisted of a different group of 18 English-speaking Northwestern undergraduates. These participants were recruited using the Northwestern Linguistics Department subject pool, and were given partial course credit for their time. The monolingual group was comprised of 5 males and 13 females, with a mean age of 19.8 (range 18-22). Participants in this group reported using English 97% of the time on average (range 94%-100%). Fourteen of these participants did not learn a language other than English before age 7. On average, the four participants who did report learning a language before age 7 currently use that language on average 2% of the time (range 0%-5%).

[12] This calculation is lacking for one participant in this group who declined to report his/her age, making it impossible to compute length of residence information. This participant did report moving to an English-speaking country at age 22.

Eight additional participants were run but their data were excluded unanalyzed because either they did not conform to the language background requirements of the monolingual group (6), or because of a technical glitch (2).[13]

There was no Mandarin-English bilingual group in Experiment 1b.

3.6.4 Experiment 1: Stimuli

3.6.4.1 Experiment 1: Stimulus materials

Sixty-four sentences of English and 64 sentences of Mandarin Chinese were used in this experiment, as described in Section 3.5.3.1.

3.6.4.2 Experiment 1: Stimulus talker characteristics

The sentences in Experiment 1a were read by a male talker (named Male 1 in Table 3.10) and a female talker (Female in Table 3.10). The male talker was 25 years old at the time of recording and had an overall Versant score of 50 (and a Versant fluency score of 54 and a Versant pronunciation score of 49). The female talker was 22 years old at the time of recording had an overall Versant score of 65 (and a Versant fluency score of 80 and a Versant pronunciation score of 69). These talkers' voices sounded typical of their genders.

The sentences in Experiment 1b were read by the same female talker (Female in Table

[13] As was described in Footnote 11 (to Section 3.6.3.1), an initial goal of this project was to have 36 total monolingual English listeners participate in this experiment, and as such additional participants were run in an attempt to add an additional 18 to the existing 18 participants described above. In the course of attempting to add this second group of 18 participants, 22 additional participants were run, some of which had knowledge of Mandarin (7), or experienced technical glitches (2). When the decision was made to only include 18 participants in all listener groups of all Garner paradigm experiments of this dissertation, the data of all 22 of these participants were excluded, unanalyzed.

3.10), paired with a different male talker (Male 2), producing a different male-female pair. The male talker in Experiment 1b was selected to more closely match the female talker in terms of English proficiency. He was 23 years old at the time of recording and had an overall Versant score of 59 (and a Versant fluency score of 77 and a Versant pronunciation score of 61), higher scores than the male talker in Experiment 1a. Thus, there was more of a match between the male and female talkers for L2 fluency in Experiment 1b than in Experiment 1a. Refer back to Table 3.10 to compare characteristics of talkers used across experiments.

3.6.4.3 Experiment 1: Acoustic characteristics of stimuli

3.6.4.3.1 Duration of stimuli

As given in Table 3.11, in Experiment 1a the mean duration of all 128 stimuli produced by the male talker (including both Chinese and English sentences) was 1750.00 ms, and for the female talker was 1868.49 ms. Averaged across talkers, the mean duration for all 128 Chinese stimuli was 2021.59 ms and for the English stimuli was 1596.91 ms.

As given in Table 3.12, in Experiment 1b the mean duration of all 128 stimuli produced by the male talker (including both Chinese and English sentences) was 1622.64 ms, and for the female talker was 1868.49 ms. Averaged across talkers, the mean duration for all 128 Chinese stimuli was 1887.54 ms and for the English stimuli was 1603.59 ms.

3.6.4.3.2 Amplitude of stimuli

As described in Section 3.5.3.3.2, for both Experiments 1a and 1b, all stimuli were amplitude-normalized to 70 dB SPL and were presented at a comfortable listening level.

3.6.4.4 Experiment 1: Stimulus selection and arrangement

In Experiments 1a and 1b, stimuli were selected and arranged in the manner reported in Section 3.5.3.4.

3.6.5 Experiment 1: Procedure

The procedure outlined in Section 3.5.4 was followed in Experiments 1a and 1b. For Experiments 1a and 1b, the buttons labels on the button box for the gender task were “MALE” and “FEMALE”, and for the language-being-spoken task were “CHINESE” AND “ENGLISH”. Screen shots of the instructions given to participants in Experiment 1 are shown below in Figure 3.3. Experiments 2 and 3 used these same instructions with modifications to reflect the dimension being compared against language-being-spoken.

Figure 3.3. Screenshot of instructions for participants for the gender task (left) and the language-being-spoken task (right).

In this task, you must decide whether each voice you hear is Male or Female. Sometimes the voices may be speaking in different languages, but your job is to respond to the Gender of the voice.

If the voice you hear is Male, press the button corresponding to "Male" on the button box.
If the voice you hear is Female, press the button corresponding to "Female" on the button box.

Please respond as quickly and accurately as possible. You may respond as soon as you have identified the speaker as Male or Female; you do not have to wait until they finish speaking in order to respond.

You will now be given a few practice trials. Press a button to begin the practice.

In this task, you must decide whether each voice you hear is speaking Chinese or English. Sometimes the speakers may be Male and sometimes they may be Female, but your job is to respond to the Language being spoken.

If the voice you hear is speaking Chinese, press the button corresponding to "Chinese" on the button box.
If the voice you hear is speaking English, press the button corresponding to "English" on the button box.

Please respond as quickly and accurately as possible. You may respond as soon as you have identified the language as English or Chinese; you do not have to wait until they finish speaking in order to respond.

You will now be given a few practice trials. Press a button to begin the practice.

3.6.6 Experiment 1: Results

Any trial in which a participant did not respond (or any response that occurred later than 4000 ms after stimulus onset) was recorded as "no response" and was eliminated. In Experiment 1a, these exclusions represented 0.38% of all trials for English listeners and 0.69% of all trials for Mandarin-English listeners. In Experiment 1b, these exclusions represented 0.38% of all trials for those English listeners.

3.6.6.1 Experiment 1: Accuracy analysis

Responses from practice trials were discarded. Means and standard deviations of accuracy (in percent correct) for each language group as a factor of block and task are presented in the first part of Tables 3.15 and 3.16, below. Accuracy analysis was conducted by first converting percent correct to empirical logit-transformed proportions (Cox, 1970; Jaeger, 2008). Then, ANOVAs were performed over empirical logits. While new techniques have been developed for the analysis of categorical data (e.g. Jaeger, 2008 for logit mixed models), ANOVA was used here for maximal comparability with existing Garner paradigm studies. Results from participants in Experiment 1a and Experiment 1b are presented separately. As can be seen in the accuracy results in Tables 3.15 and 3.16, performance was near ceiling for all listener language groups in all blocks in all tasks (above 97%).

To determine whether there were any differences in accuracy between blocks for each listener group within Experiment 1a, empirical logit-transformed proportions were calculated for each participant for each block in each task. These were submitted to a 2 x 2 x 3 repeated measures ANOVA containing the between-subject factor of listener language group (English monolinguals and Mandarin-English bilinguals), and within-subjects factors of task (gender and language-being-spoken, hereafter in the data analysis referred to simply as "language"), and block (correlated, control, and orthogonal). No main effects or interactions were significant. There was no main effect of listener language group, $F(1,33) = 0.188$, $p = 0.667$; no main effect of task, $F(1,33) = 0.670$, $p = 0.419$; and no main effect of block, $F(2,66) = 2.227$, $p = 0.116$. There was no interaction between language and task, $F(1,33) = 0.346$, $p = 0.560$; no interaction between listener language group and block, $F(2,66) = 0.152$, $p = 0.859$; no interaction between task and block, $F(2,66) = 2.923$, $p = 0.061$; and no interaction between listener language group, task, and block, $F(2,66) = 0.478$, $p = 0.622$). In all, listeners in both language background groups were equally accurate across tasks and blocks. Lacking a significant main effect of listener language group or any significant interactions involving listener language group, listener language groups will not be analyzed separately for accuracy.

To determine whether English listeners behaved similarly across Experiments 1a and 1b in terms of accuracy, empirical logit-transformed proportions were submitted to a 2 x 2 x 3 repeated measures ANOVA containing the between-subjects factor of experiment (Experiment 1a and Experiment 1b), and within-subjects factors of task (gender and language) and block (correlated, control, and orthogonal). There was no main effect of experiment, $F(1,33) = 0.064$, $p = 0.801$. There was a main effect of task, $F(1,33) = 6.512$, $p = 0.016$, such that gender (M = 98.93% correct) was more accurate than language (M = 98.32% correct), a mean difference of 0.61%. There was also a main effect of block, $F(2,66) = 6.239$, $p = 0.003$. Planned comparisons revealed that correlated blocks (M = 99.06% correct) were performed more accurately than control blocks (M = 98.44% correct), $t(69) = -2.306$, $p = 0.024$, correlated blocks were performed more accurately than orthogonal blocks (*M* = 98.37% correct), $t(69) = 3.482$, $p = 0.0009$, but there was no significant difference between accuracy in control versus orthogonal blocks, $t(71) = 0.822$, $p = 0.414$. There was no interaction of experiment and task, $F(1,33) = 0.865$, $p = 0.359$, no interaction of experiment and block $F(2,66) = 0.422$, $p = 0.657$, no interaction of task and block $F(2,66) = 2.246$, $p = 0.114$, and no interaction of experiment, task, and block $F(2,66) = 0.032$, $p = 0.968$. While there were main effects of task and block, there was no main effect of experiment or any significant interactions involving experiment, indicating that English listeners across experiments were equally accurate for both pairs of talkers. This indicates that proficiency differences between pairs of talkers did not impact English listeners' performance, at least in terms of accuracy.

Despite near ceiling performance in both experiments, for English listeners a very small but reliable difference emerged in task (where gender was slightly more accurate than language) and in block (where correlated was slightly more accurate than both control and orthogonal blocks). However, accuracy was very high across all conditions (greater than 97%), negating the need to investigate a possible speed-accuracy tradeoff. As such, the remainder of the analyses in this experiment will focus on reaction times.

Table 3.15. Experiment 1a. Mean by-participant error rates and reaction times (RTs) both with standard error.

Experiment 1a		Accuracy (Percent correct)				RT (ms)			
		English monolinguals		Mandarin-English bilinguals		English monolinguals		Mandarin-English bilinguals	
Task	Block	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Gender	Correlated	99.05	0.44	98.61	0.40	1379.48	112.22	1353.86	100.83
	Control	98.87	0.53	99.05	0.31	1443.66	132.63	1364.85	118.61
	Orthogonal	98.87	0.49	98.18	0.61	1503.73	124.44	1391.75	111.67
Language	Correlated	98.99[14]	0.52[14]	98.09	1.47	1564.17[14]	118.47[14]	1374.57	98.78
	Control	98.00	0.58	97.57	1.00	1623.91	115.21	1445.58	97.62
	Orthogonal	98.09	0.64	97.83	0.74	1683.90	111.02	1470.90	95.59

Table 3.16. Experiment 1b. Mean by-participant error rates and reaction times (RTs) both with standard error.

Experiment 1b		Accuracy (Percent correct)		RT (ms)	
		English monolinguals		English monolinguals	
Task	Block	Mean	SE	Mean	SE
Gender	Correlated	99.48	0.18	1246.97	53.32
	Control	98.52	0.95	1283.02	61.51
	Orthogonal	98.96	0.25	1313.84	71.22
Language	Correlated	98.78	0.61	1377.68	72.74
	Control	98.09	0.53	1519.54	79.93
	Orthogonal	97.74	0.49	1559.46	77.77

3.6.6.2 Experiment 1: Reaction time analysis

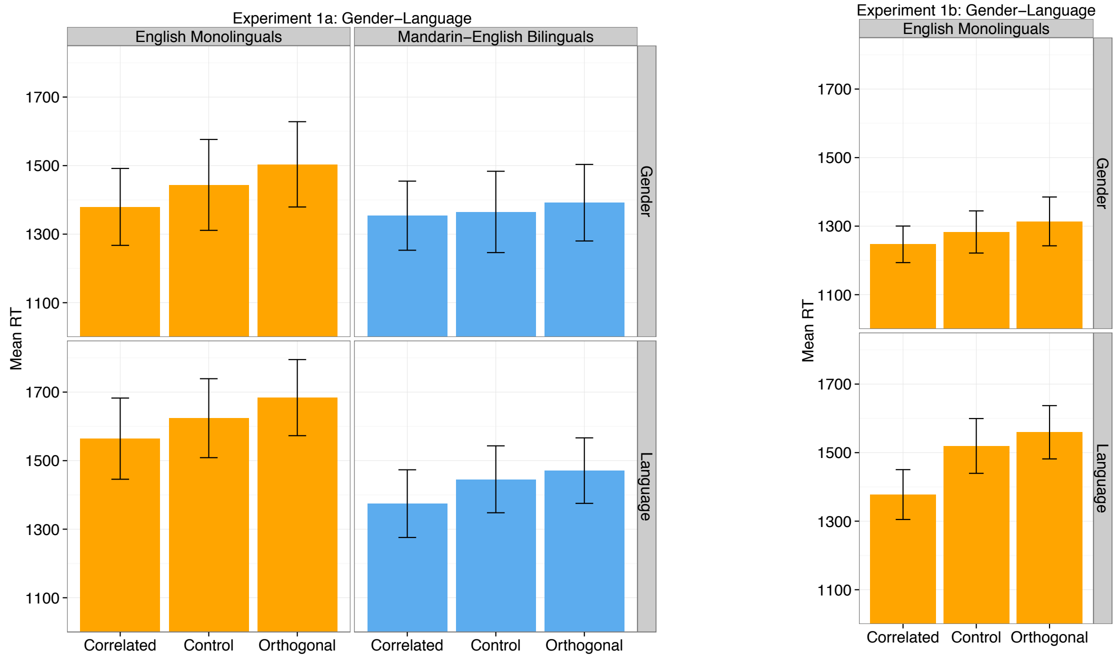
In order to analyze reaction time data, incorrect responses were discarded. Within the correct responses, data points greater than 2.5 standard deviations from each participant's mean for a particular block in a particular task were removed. For English monolingual listeners, this resulted in the removal of 2.26% of total responses, and for Mandarin-English bilinguals, the removal of 2.65%. In Experiment 1b, this resulted in the removal of 2.11% of responses for those English monolingual listeners. Means and standard errors of reaction times for each language

[14] Note that these means do not include the correlated responses for one participant, whose data for the correlated block in the language task were inexplicably corrupted.

group as a factor of block and task are presented in the second part of Tables 3.15 and 3.16. Figure 3.4 plots raw reaction times by task and block separately for each listener group. Reaction times were log-transformed before submission to ANOVA in order to reduce skewing.

Reaction time data was analyzed using repeated measures ANOVAs on by-participant means. While newer methods exist for the analysis of continuous dependent variables (e.g. linear mixed effects regressions, Baayen, 2008), repeated measures ANOVAs take into account the fact that a participant in this study performed the same action (responding to stimuli in an orthogonal block, for example) across two tasks (gender and language, for Experiment 1). Item-specific effects are absent from these analyses (by necessity for by-participant means), a decision driven by the fact that in these experiments, item-specific contributions to these effects are likely to be random; any difference in response times between control and orthogonal blocks should be the same for any two specific items. Moreover, repeated measures ANOVAs give maximal comparability with previous studies using the Garner paradigm, which favor this analysis. Thus, repeated measures ANOVAs on by-participant means will be used to analyze reaction time data for Experiments 1-3 of this dissertation.

Figure 3.4. By-participant means of reaction times by condition for Experiments 1a and 1b. Error bars represent +/- 1 SE of the mean.



3.6.6.2.1 Experiment 1: Baseline performance (control block only)

In order to determine how participants performed in the baseline (control) block, mean log-transformed reaction times for the control block were calculated for each participant. These were then submitted to a 2 x 2 repeated measures ANOVA containing the between-subjects factor listener language group (English monolinguals and Mandarin-English bilinguals) and the within-subjects factor task (gender and language). There was no main effect of listener language group, $F(1,34) = 0.746$, $p = 0.394$, but the main effect of task was significant, $F(1,34) = 11.996$, $p = 0.002$, such that the gender task (*M* = 1404.25 ms) was performed faster than the language task (*M* = 1534.75 ms) in the control block. Across listener groups, there was no difference in performance on the control block across tasks as evidenced by the lack of a significant interaction between task and listener language group, $F(1,34) = 1.138$, $p = 0.294$. In other words, the gender task was performed faster than the language task in the control block for both listener groups. Lacking a significant main effect of listener language group or a significant interaction involving listener language group, listener language groups will not be analyzed separately on control block performance.

To analyze whether the English listeners behaved similarly across experiments on control blocks, empirical logit-transformed proportions were submitted to a 2 x 2 repeated measures ANOVA containing the between-subjects factor of experiment (Experiment 1a and Experiment 1b), and the within-subjects factor of task (gender and language). There was no main effect of experiment, $F(1,34) = 0.655$, $p = 0.424$. There was a main effect of task, $F(1,34) = 27.058$, $p < 0.0001$, such that the gender task ($M = 1363.34$ ms) was performed faster than the language task ($M = 1571.73$ ms) in the control block. Across experiments, there was no difference in performance on the control block across tasks as evidenced by the lack of a significant interaction between task and experiment $F(1,34) = 0.166$, $p = 0.687$. This, and the lack of a main effect of experiment, indicates that the proficiency differences between talker pairs did not impact how quickly English listeners classified gender and language dimensions at baseline. In fact, reaction times in the control blocks to the pair of talkers more matched in English proficiency (Experiment 1b; 1283.02 ms for English monolinguals in gender task, 1519.54 ms for English monolinguals in language task) were numerically faster than those of the less matched pair (Experiment 1a; 1443.66 ms for English monolinguals in gender task, 1623.91 ms for English monolinguals in language task) for both tasks, indicating that the pair of talkers with more of a proficiency match actually tended to be slightly more discriminable. Though, as noted, these differences were not statistically significant. Nonetheless, this is secondary evidence supporting the fact that English proficiency did not seem to be used by listeners when performing the gender or language tasks. In sum, English listeners in both experiments perform faster on the gender task than in the language task in the control block, despite differences in talker proficiency pairs.

Overall, both listener groups perform the gender task faster than the language task in the control block, as predicted. Thus, the prediction that gender is processed more quickly than language was confirmed, and the levels of processing hypothesis does not need to be adjusted for these dimensions.

3.6.6.2.2 Experiment 1: Garner interference?

To determine whether participants exhibited Garner interference, mean log-transformed reaction times were calculated for each participant for each block (control and orthogonal only) in each task. These were submitted to a 2 x 2 x 2 repeated measures ANOVA containing the between-subjects factor of listener language group (English monolinguals and Mandarin-English bilinguals) and the within-subjects factors of task (gender and language) and block (control and orthogonal). In this test, the presence of Garner interference would be signaled by a significant main effect of block, and an asymmetry in interference would be signaled by a significant interaction between task and block. There was no main effect of listener language group, $F(1,34) = 0.979$, $p = 0.33$. There was a main effect of task, $F(1,34) = 18.710$, $p = 0.0001$, such that gender (M = 1426.00 ms) was performed faster than language (M = 1556.07 ms). There was a main effect of block, $F(1,34) = 5.647$, $p = 0.023$, such that control blocks (M = 1469.50 ms) were performed faster than orthogonal blocks (M = 1512.57 ms), indicating the presence of Garner interference across tasks and language background. There was no interaction between listener language group and task, $F(1,34) = 1.811$, $p = 0.187$; no interaction between listener language group and block, $F(1,34) = 0.326$, $p = 0.572$; no interaction between task and block, $F(1,34) = 0.022$, $p = 0.882$; and no interaction between listener language group, task, and block, $F(1,34) < 1$, $p = 0.986$. In sum, across listener language groups, the main effect of block indicates that there was Garner interference. Moreover, the lack of interaction between task and block reveals that the interference was stable across task, meaning that the interference was mutual and symmetrical. The lack of interaction between listener language group and block reveals that the interference was also stable across listener language groups. Lacking a significant main effect of listener language group or any significant interactions involving listener language group, listener groups will not be analyzed separately with respect to Garner interference.

To determine whether the English listeners had similar amounts of Garner interference across the two experiments, by-participant mean log-transformed reaction times were submitted to a 2 x 2 x 2 repeated measures ANOVA containing the between-subjects factor of experiment (Experiment 1a and Experiment 1b), and the within-subjects factors of task (gender and language) and block (control and orthogonal). There was no main effect of experiment, $F(1,34) = 0.898$, $p = 0.35$. There was a main effect of task, $F(1,34) = 43.265$, $p < 0.0001$, such that gender ($M = 1386.06$ ms) was performed faster than language ($M = 1596.70$ ms). There was a main effect of block, $F(1,34) = 5.895$, $p = 0.021$, such that control ($M = 1467.53$ ms) was performed faster than orthogonal ($M = 1515.23$ ms), indicating that English listeners showed Garner interference across tasks and experiments. There was no interaction of experiment and task, $F(1,34) = 0.413$, $p = 0.525$; no interaction of experiment and block, $F(1,34) = 0.519$, $p = 0.476$;

no interaction of task and block, $F(1,34) = 0.009$, $p = 0.926$; and no interaction of experiment, task, and block $F(1,34) = 0.066$, $p = 0.799$. The lack of a main effect of experiment, and of any significant interactions with experiment, indicate that the proficiency differences between talker pairs did not impact the amount of Garner interference experienced by English listeners.

In sum, listeners from both language backgrounds experienced mutual and symmetrical interference between gender and language-being-spoken. The degree of Garner interference is shown for each task by listener group in Figure 3.5, below. Group-level Garner interference (e.g. English monolinguals' mean orthogonal block reaction time minus their mean RT in the control block) is represented by the orange (English monolingual listeners) and blue (Mandarin-English bilingual listeners) bars. Each individual subject's Garner interference (e.g. one individual English monolingual participant's mean orthogonal block RT minus that participant's mean control RT) is plotted with a grey dot, showing the inter-listener variability in interference within each listener language background. Groups and participants above the $y = 0$ line experienced Garner interference, as this represents longer orthogonal RTs than control RTs. The mutual and symmetrical Garner interference is evident in the plot, and the lack of difference in the amount of interference between listener language groups is evident, as well. The variance between participants is also striking; many participants do not exhibit Garner interference, even showing longer RTs in the control block than in the orthogonal block. Most prior studies using the Garner paradigm do not show participant-level differences, so it is unknown if such variability is common, though Kimchi, Behrmann, Avidan, & Amishav (2012) report significant inter-participant variability in the interaction between two types of featural information in the visual processing of faces.

Figure 3.5. Experiments 1a and 1b, Garner interference. Bars represent group means of differences between orthogonal and control blocks, collapsed over participant. Dots represent individual participants' differences between the two blocks.

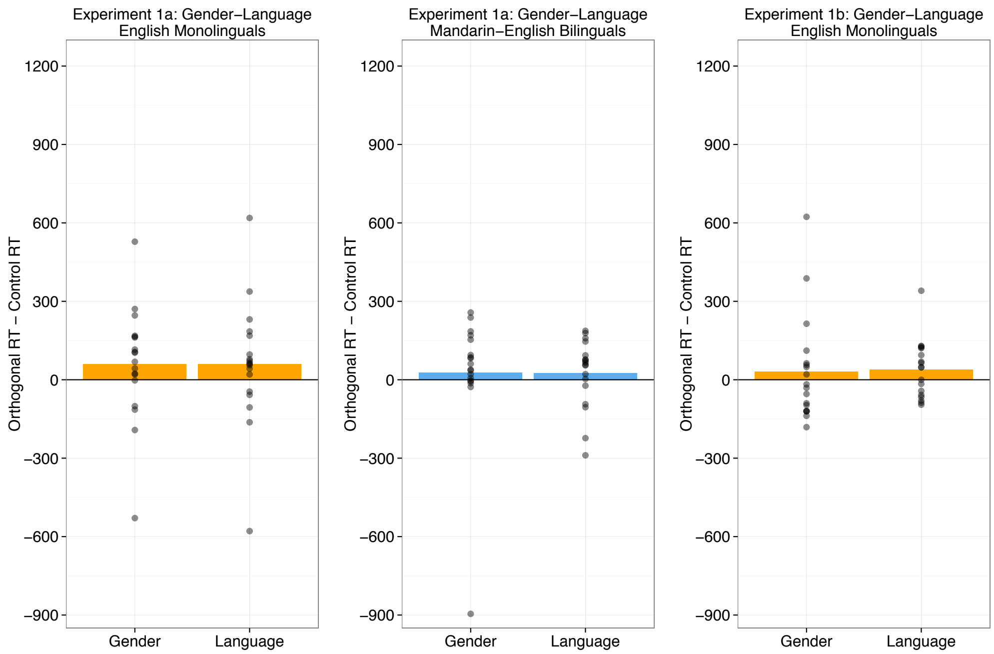


Figure 3.6 gives a different representation of individual participants' levels of Garner interference for the gender and language tasks. The *x* axes of these plots give the reaction time in the control block, while the *y* axes give the reaction time in the orthogonal block. Participants are represented by dots color coded by listener language group. If a listener exhibits Garner interference, their dot will be above the $y = x$ line, since orthogonal will be larger than control. This figure again illustrates the range of individual-level performance.

Figure 3.6. Experiment 1a and 1b, Garner interference for each task, by participant. Error bars represent +/- 1 SE of the mean.

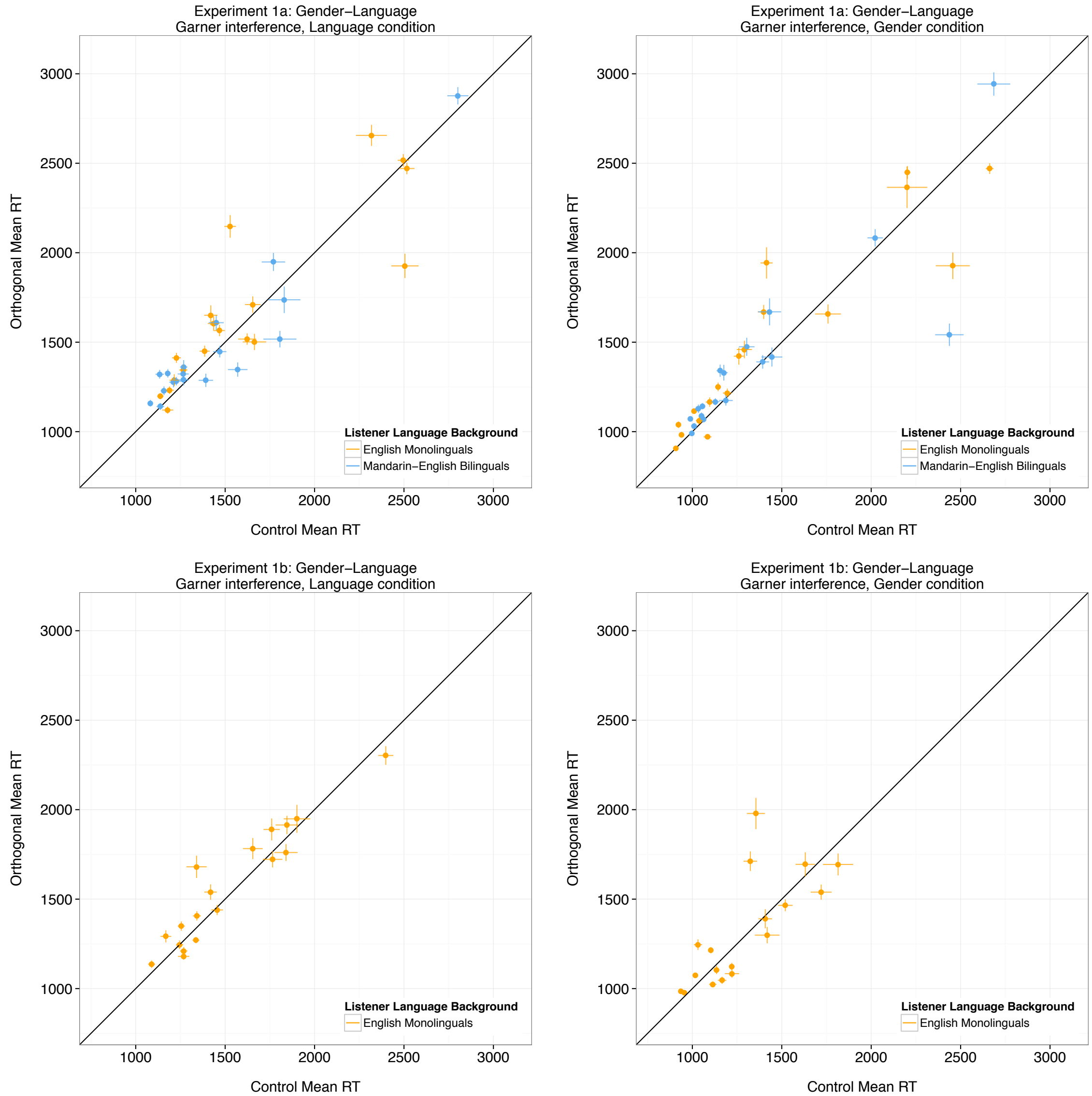
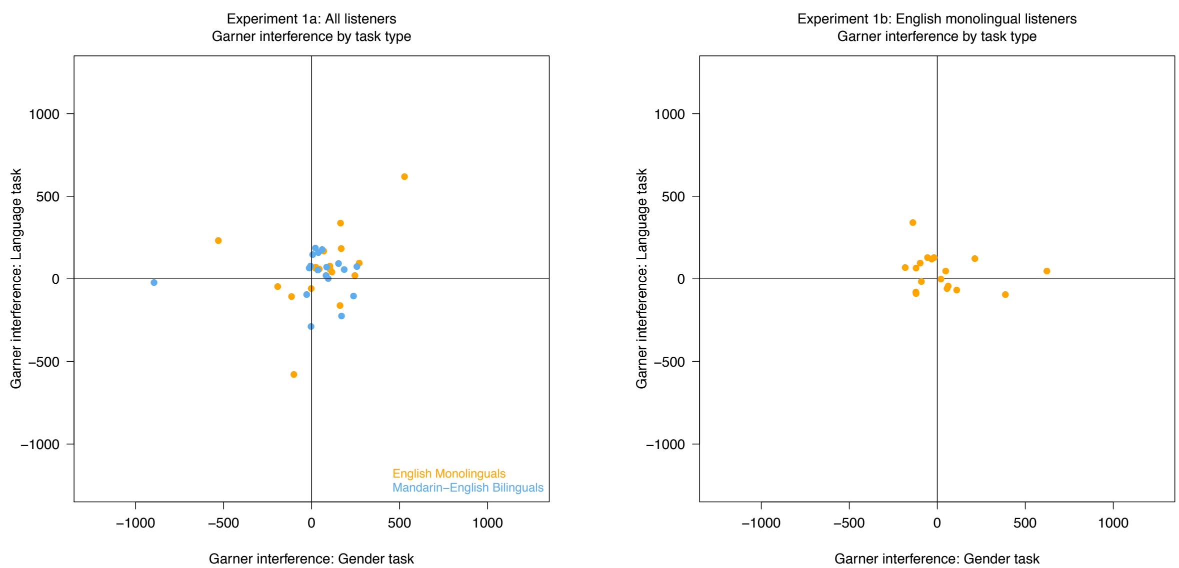


Figure 3.7 below visualizes individual participants' levels of Garner interference as a function of the task, for Experiments 1a and 1b. Participants are represented by dots, color coded by listener language group. The *x* axis gives the amount of Garner interference on the gender task (the farther to the right of the $x = 0$ line, the greater the amount of interference on that task),

and the *y* axis gives the amount of Garner interference on the language task (the farther above the $y = 0$ line, the greater the interference on that task). Thus, participants who show positive Garner interference (orthogonal > control) on both tasks will be shown in the upper right quadrant of the plot. Participants who show negative Garner interference (orthogonal < control) on both tasks will be shown in the lower left quadrant of the plot. Participants who show positive interference on the gender task (orthogonal > control) but negative interference on the language task (orthogonal < control) will be shown in the lower right quadrant of the plot. Participants who show positive interference on the language task (orthogonal > control) but negative interference on the gender task (orthogonal < control) will be shown in the upper left quadrant of the plot.

Figure 3.7. Experiments 1a and 1b. Garner interference for both tasks, by participant.



3.6.6.2.3 Experiment 1: Redundancy gain?

To determine whether participants exhibited a redundancy gain, mean log-transformed reaction times were calculated for each participant for each block (correlated and control only) in each task. These were submitted to a 2 x 2 x 2 repeated measures ANOVA containing the between-subjects factor of listener language group (English monolinguals and Mandarin-English bilinguals) and the within-subjects factors of task (gender and language) and block (correlated and control). In this test, the presence of a redundancy gain would be signaled by a significant main effect of block, and an asymmetry in redundancy gain would be signaled by a significant interaction between task and block. There was no main effect of listener language group, $F(1,33$[15]$) = 0.977$, $p = 0.33$. There was a main effect of task, $F(1,33) = 9.839$, $p = 0.004$, such that gender (M = 1398.01 ms) was performed faster than language (M = 1505.64 ms). There was a main effect of block, $F(1,33) = 5.965$, $p = 0.020$, such that correlated blocks (M = 1422.34 ms) were performed faster than control blocks (M = 1481.31 ms), indicating a redundancy gain across task and language background. There was no interaction between listener language group and task, $F(1,33) = 2.052$, $p = 0.161$; no interaction between listener language group and block, $F(1,33) = 0.379$, $p = 0.542$; no interaction between task and block, $F(1,33) = 3.086$, $p = 0.088$; and no interaction between listener language group, task, and block, $F(1,33) = 0.533$, $p = 0.471$. In sum, across listener language groups, the main effect of block indicates that there was redundancy gain. Moreover, the lack of interaction between task and block reveals that the redundancy gain was stable across blocks (symmetrical), just as the lack of interaction between listener language group and block reveals that the redundancy gain was stable across listener language groups. Lacking a significant main effect of listener language group or any significant interactions involving listener language group, listener language groups will not be analyzed

[15] As before, there is no data for one participant in the correlated block, reflected here in fewer degrees of freedom.

separately with respect to redundancy gain.

To determine whether English listeners had similar amounts of redundancy gain across the two experiments, by-participant mean log-transformed reaction times were submitted to a 2 x 2 x 2 repeated measures ANOVA containing the between-subject factor of experiment (Experiment 1a and Experiment 1b), and the within-subjects factors of task (gender and language) and block (correlated and control). There was no main effect of experiment, $F(1,33) = 1.319$, $p = 0.259$. There was a main effect of task, $F(1,33) = 31.360$, $p < 0.0001$, such that gender (M = 1349.48 ms) was faster than language (M = 1525.46 ms). There was a main effect of block, $F(1,33) = 13.833$, $p = 0.0007$, such that correlated (M = 1395.65 ms) was faster than control (M = 1479.29 ms), indicating that English listeners had an overall redundancy gain across tasks and experiments. There was no interaction between experiment and task, $F(1,33) = 0.014$, $p = 0.905$; no interaction between experiment and block, $F(1,33) = 0.261$, $p = 0.613$; no interaction between task and block, $F(1,33) = 3.719$, $p = 0.062$; and no interaction between experiment, task, and block, $F(1,33) = 0.973$, $p = 0.331$. The lack of a main effect of experiment, and of any significant interactions with experiment, indicate that the proficiency differences between talker pairs did not impact the amount of redundancy gain experienced by English listeners.

In sum, listeners demonstrated a redundancy gain across gender and language-being-spoken tasks across language backgrounds.

3.6.7 Experiment 1: Summary of results

Table 3.17, below, gives a summary of results of Experiment 1a. Section 3.9, at the end of this chapter, will assess how well the results of this experiment were predicted by each hypothesis raised above.

Table 3.17. Summary of results from Experiment 1a. Listener language group did not interact with task or block in any analysis. English listeners from Experiment 1b also follow the same pattern.

Experiment 1a: Results summary, all listener language groups	
Control block performance	Gender task (*M* = 1404.25 ms) faster than Language task (*M* = 1534.75 ms) $F(1,34) = 11.996$, $p = 0.002$
Garner interference? *(Orthogonal - Control)*	Yes. Control blocks (*M* = 1469.50 ms) faster than Orthogonal blocks (*M* = 1512.57 ms) $F(1,34) = 5.647$, $p = 0.023$
Symmetrical Garner interference?	Yes. No significant interaction between task and block $F(1,34) = 0.022$, $p = 0.882$
Redundancy gain? *(Control - Correlated)*	Yes. Correlated blocks (*M* = 1422.34 ms) faster than Control blocks (*M* = 1481.31 ms) $F(1,33) = 5.965$, $p = 0.020$

3.7 Experiment 2: Talker–Language-being-spoken

3.7.1 Experiment 2: Rationale

Experiment 2 was designed to test the integration of language-being-spoken and another indexical dimension, talker identity. Experiment 2 tested listeners from three language backgrounds, English monolingual listeners, and Mandarin-English bilingual listeners, and non-Mandarin-English bilingual listeners.

3.7.2 Experiment 2: Hypotheses and predictions

As described above, there are several predictions for the results of Experiment 2. First, in terms of processing speed at baseline, talker and language-being-spoken are predicted to be equal. Next, in terms of dependency relations, several predictions hold. Following the levels of processing hypothesis, language-being-spoken and talker will show mutual and symmetrical interference, if predictions about processing speed are confirmed. If talker and language-being-spoken turn out to be processed at different speeds in control blocks, the levels of processing hypothesis will be refined to reflect the observed levels of processing. In contrast to the levels of processing hypothesis, the relative language-specificity/talker-generality hypothesis (LS/TG) suggests that language-being-spoken and talker will show mutual and asymmetrical interference such that language-being-spoken interferes more with talker than vice versa. With regard to listener language backgrounds, Mandarin-English bilinguals are hypothesized to show more interference overall than both English monolinguals and non-Mandarin-English bilinguals following the language familiarity enhances dimensional representations hypothesis. Following the bilingualism enhances dimensional representations hypothesis, however, both Mandarin-English bilinguals and non-Mandarin-English bilinguals are hypothesized to show more interference than English monolinguals. Finally, following the bilingualism enhances selective attention hypothesis, the opposite should be true (Mandarin-English bilinguals and non-Mandarin-English bilinguals should show less interference than English monolinguals).

3.7.3 Experiment 2: Participants

The monolingual group of listeners in Experiment 2 consisted of 18 English-speaking Northwestern undergraduates. These participants were recruited using the Northwestern Linguistics Department subject pool, and were given partial course credit for their time. The monolingual group was comprised of 7 males and 11 females, with the mean age 20.1 (range 18-22). Participants in this group reported using English 97% of the time on average (range 92%-100%). Ten of these participants did not learn a language other than English before age 7. The eight participants who did report learning a language before age 7 currently use that language on average 2% of the time (range 0%-10%). No participant in this group reported knowing any Mandarin Chinese. Fourteen additional participants were run but their data were excluded unanalyzed because they either had knowledge of Mandarin (2), did not conform to the language background requirements of the monolingual group (i.e. were bilingual in languages other than Mandarin) (2), had a hearing impairment (1), experienced a technical glitch (1), or could not learn to successfully classify the talkers (8).[16]

The 18 Mandarin-English bilingual subjects who participated in Experiment 2 were recruited both from the International Summer Institute (ISI) at Northwestern in 2012 and from the greater Northwestern community. These participants received compensation in cash for their time (less than 1 hour). Of these 18 Mandarin-English bilingual participants, 13 were male and 5 were female, with a mean age of 22.9 (range 21-28). On average, the participants in this group reported learning English at age 10.2 (range 3-14), and report using English an average of 16% of the time (range 10%-80%). These participants have lived in an English speaking country for an average of 0 years (range 0-0 years); these participants were almost exclusively recruited from ISI, which is a program for international graduate students just entering Northwestern. Two additional participants were run but their data were excluded unanalyzed because they either did not answer enough of the language background questionnaire to determine their eligibility (1), or could not learn to successfully classify the talkers (1).

The 18 non-Mandarin-English bilingual subjects (bilinguals whose first language was something other than Mandarin Chinese) who participated in Experiment 2 were recruited from the greater Northwestern community, and received compensation in cash for their time (less than

[16] As was described in Footnote 11 (to Section 3.6.3.1), an initial goal of this project was to have 36 total monolingual English listeners participate in this experiment, and as such additional participants were run in an attempt to add an additional 18 to the existing 18 participants described above. In the course of attempting to add this second group of 18 participants, 26 additional participants were run, some of which had knowledge of Mandarin (4), did not conform to the language background requirements of the monolingual group (i.e. were bilingual in languages other than Mandarin) (7), or experienced technical glitches (1). When the decision was made to only include 18 participants in all listener groups of all Garner paradigm experiments of this dissertation, the data of all 26 of these participants were excluded, unanalyzed.

1 hour). Of these 18 non-Mandarin-English bilingual participants, 10 were male and 8 were female, with a mean age of 27.3 (range 19-36). On average, the participants in this group reported learning English at age 8.9 (range 0-17), and report using English an average of 62% of the time (range 30%-96%). These participants have lived in an English speaking country for an average of 2.4 years (range 0-9 years).[17] The native languages of participants in this group were: Spanish (4), Greek (2), Thai (2), Portugese (2), Korean (1), Croatian (1), Czech (1), Russian (1), Serbian (1), Farsi (1), Norwegian (1), and Kannada (1). No participant in this group reported knowing any Mandarin Chinese. Three additional participants were run but their data were excluded unanalyzed because they could not learn to successfully classify the talkers.

3.7.4 Experiment 2: Stimuli

3.7.4.1 Experiment 2: Stimulus materials

Sixty-four sentences of English and 64 sentences of Mandarin Chinese were used in this experiment, as described in Section 3.5.3.1.

3.7.4.2 Experiment 2: Stimulus talker characteristics

The sentences in Experiment 2 were read two male talkers (named Male 1 and Male 2 in Table 3.10). Male talker 1, named "Wei" in the experiment, was 25 years old at the time of recording and had an overall Versant score of 50 (and a Versant fluency score of 54 and a Versant pronunciation score of 49). Male talker 2, named "Li" in the experiment, was 23 years

[17] One participant declined to report when they had moved to an English speaking country (but learned English at age 12, and use it 40% of the time), so this calculation only includes information from 17 of the 18 participants in this group.

old at the time of recording had an overall Versant score of 59 (and a Versant fluency score of 77 and a Versant pronunciation score of 61). Refer back to Table 3.10 to compare characteristics of talkers used across experiments.

3.7.4.3 Experiment 2: Acoustic characteristics of stimuli

3.7.4.3.1 Duration of stimuli

As given in Table 3.13, in Experiment 2 the mean duration of all 128 stimuli produced by the male talker given the name "Wei" (including both Chinese and English sentences) was 1750.00 ms, and for the talker given the name "Li" was 1622.64 ms. Averaged across talkers, the mean duration for all 128 Chinese stimuli was 1811.84 ms and for the English stimuli was 1560.80 ms.

3.7.4.3.2 Amplitude of stimuli

As described in Section 3.5.3.3.2, for Experiment 2 all stimuli were amplitude-normalized to 70 dB SPL and were presented at a comfortable listening level.

3.7.4.4 Experiment 2: Stimulus selection and arrangement

In Experiment 2, stimuli were selected and arranged in the manner reported in Section 3.5.3.4.

3.7.5 Experiment 2: Procedure

The procedure outlined in Section 3.5.4 was followed in Experiment 2, with one modification. In the talker task, participants must learn which name each voice corresponds to. Thus, before participants began the practice trials of the talker task, the experimenter gave the following instructions verbally: “In this experiment you’ll be asked to indicate whether the voice you’re hearing is the voice of one person or another person. You’ll learn who is who by trial and error in the practice blocks. On the first trial, just guess, and it will tell you whether you are right or wrong. From there, you can figure out who is who.” As mentioned in Section 3.7.3, several listeners were not able to learn to distinguish between the talkers despite eventual success on the practice trials, and their data were excluded. For Experiment 2, the button labels on the button box for the talker task were “WEI” and “LI”, and for the language-being-spoken task were “CHINESE” AND “ENGLISH”. Instructions given to participants in Experiment 2 are identical to those pictured in Figure 3.3, but changed to reflect the use of the dimension talker rather than gender.

3.7.6 Experiment 2: Results

As in Experiment 1, any trial in which a participant did not respond (or any response that occurred later than 4000 ms after stimulus onset) was recorded as “no response” and was eliminated. In Experiment 2, no response trials represented 0.72% of trials for English listeners, 0.42% of trials for the Mandarin-English listeners, and 0.28% of the trials for the non-Mandarin-English listeners.

3.7.6.1 Experiment 2: Accuracy analysis

Responses from practice trials were discarded. Means and standard deviations of accuracy (in percent correct) for each language group as a factor of block and task are presented in the first part of Table 3.18, below. As in Experiment 1, accuracy analysis was conducted by first converting percent correct to empirical logit proportions. The accuracy results in Table 3.18 illustrate that performance was near ceiling for all listener language groups in all blocks in all tasks (above 95% correct).

To determine whether there were any differences in accuracy between blocks for each listener group, empirical logit-transformed proportions were calculated for each participant for each block in each task. These were submitted to a 3 x 2 x 3 repeated measures ANOVA containing the between-subject factor of listener language group (English monolinguals, Mandarin-English bilinguals, and non-Mandarin-English bilinguals), and within-subjects factors of task (talker and language), and block (correlated, control, and orthogonal). There was no main effect of listener language group, $F(2,51) = 1.543$, $p = 0.223$, and no main effect of task, $F(1,51) = 2.990$, $p = 0.09$. There was a significant main effect of block, $F(2,102) = 8.692$, $p = 0.0003$. Planned comparisons revealed that orthogonal blocks were slightly less accurate than both correlated and control blocks. This pattern is evidenced by a significant difference between correlated and orthogonal blocks, $t(107) = 3.339$, $p = 0.001$, such that correlated blocks (M = 97.73% correct), were performed more accurately than orthogonal blocks (M = 97.32% correct), an average difference of 0.41%; a significant difference between control and orthogonal blocks, $t(107) = 3.366$, $p = 0.001$, such that control blocks (M = 98.48% correct) were performed more accurately than orthogonal blocks (M = 97.32% correct), an average difference of 1.16%; but no significant difference between correlated blocks and control blocks, $t(107) = -0.1382$, $p = 0.890$. There was no interaction between listener language group and task, $F(2,51) = 2.844$, $p = 0.068$, and no interaction between listener language group and block, $F(4,102) = 1.125$, $p = 0.349$. There was a significant interaction between task and block, $F(2,102) = 4.125$, $p = 0.019$. Follow-up simple effects tests revealed that the correlated blocks were more accurate in the language task (M = 99.16% correct) than in the talker task (M = 96.30% correct), $t(53) = 2.663$, $p = 0.010$[18], but that there was no difference in accuracy across tasks for the control blocks, $t(53) = -0.847$, $p = 0.401$, or the orthogonal blocks, $t(53) = 1.209$, $p = 0.232$. There was no interaction between listener language group, task, and block, $F(4,102) = 1.296$, $p = 0.277$.

Despite near ceiling performance, a very small but reliable difference emerged in block (where orthogonal was slightly less accurate than both correlated and control blocks), and in the interaction between task and block (where there was a difference in performance in the correlated block between tasks). However, accuracy was very high across all conditions (greater than 95%), negating the need to investigate a possible speed-accuracy tradeoff. As such, the remainder of the analyses in this experiment will focus on reaction times.

[18] The source of this interaction may be traceable to the one Mandarin-English bilingual participant with extremely low accuracy scores in the talker task in the correlated block, as will be described in Footnote 19.

Table 3.18. Experiment 2. Mean by-participant accuracy and reaction times (RTs) both with standard error.

Experiment 2		Accuracy (Percent correct)						RT (ms)					
		English monolinguals		Mandarin-English bilinguals		Non-Mandarin-English bilinguals		English monolinguals		Mandarin-English bilinguals		Non-Mandarin-English bilinguals	
Task	Block	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Talker	Correlated	95.66	1.37	95.31[19]	3.26[19]	97.92	1.19	1643.60	124.65	1570.02	102.13	1403.75	95.37
	Control	97.40	0.78	98.96	0.36	99.22	0.32	1577.61	117.28	1540.61	79.25	1339.04	92.36
	Orthogonal	95.57	1.37	98.61	0.50	96.00	0.96	1698.073	110.51	1599.80	77.07	1548.29	112.50
Language	Correlated	99.22	0.26	99.13	0.36	99.13	0.29	1452.82	94.22	1522.27	80.88	1326.26	66.78
	Control	98.70	0.29	98.44	0.63	98.18	0.46	1522.01	83.93	1646.82	62.01	1400.71	72.47
	Orthogonal	98.26	0.47	97.92	0.51	97.57	0.62	1520.84	74.89	1692.64	75.79	1492.75	78.49

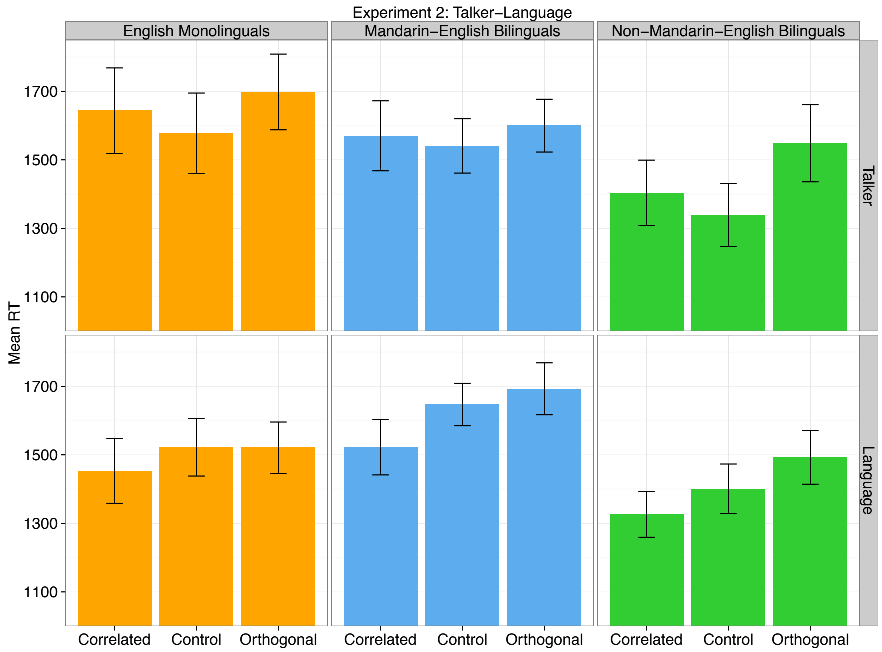
3.7.6.2 Experiment 2: Reaction time analysis

In order to analyze reaction time data, incorrect responses were discarded. Within the correct responses, data points greater than 2.5 standard deviations from each participant's mean for a particular block in a particular task were removed. For English monolingual listeners, this resulted in removing 1.94% of the data, for Mandarin-English bilinguals removing 2.05%, and for non-Mandarin-English bilinguals removing 2.60%. Means and standard errors of reaction times for each language group as a factor of block and task are presented in the second part of Table 3.18. Figure 3.8 plots raw reaction times by task and block separately for each listener group. Reaction times were then log-transformed before submission to ANOVA and analyzed as in Experiment 1.

---

[19] Note that one participant only got 40% correct in correlated block on talker task (though this participant got 100% correct in correlated block in language task), which accounts for the relatively low performance in this cell. This participant's performance on all other tasks seems to be normal.

Figure 3.8. By-participant means of reaction times by condition for Experiment 2. Error bars represent +/- 1 SE of the mean.



3.7.6.2.1 Experiment 2: Baseline performance (control block only)

In order to determine how participants performed in the baseline (control) block, mean log-transformed reaction times for the control block were calculated for each participant. These were then submitted to a 3 x 2 repeated measures ANOVA containing the between-subjects factor listener language group (English monolinguals, Mandarin-English bilinguals, and non-Mandarin-English bilinguals) and the within-subjects factor task (talker and language). There were no significant main effects or interactions. There was no main effect of listener language group, $F(2,51) = 2.349$, $p = 0.106$; no main effect of task, $F(1,51) = 3.673$, $p = 0.061$; and no interaction between listener language group and task, $F(2,51) = 1.413$, $p = 0.253$. Overall, there was no difference in reaction time between the talker task and language-being-spoken task in the control block across listener groups, as predicted. Thus, the prediction that talker and language are processed equally quickly was confirmed, and the levels of processing hypothesis does not need to be adjusted for these dimensions. Lacking a significant main effect of listener language group or a significant interaction involving listener language group, listener language groups will not be analyzed separately on control block performance.

3.7.6.2.2 Experiment 2: Garner interference?

To determine whether participants exhibited Garner interference, mean log-transformed reaction times were calculated for each participant for each block (orthogonal and control only) in each task. These were submitted to a 3 x 2 x 2 repeated measures ANOVA containing the between-subjects factor of listener language group (English monolinguals, Mandarin-English bilinguals, and non-Mandarin-English bilinguals) and the within-subjects factors of task (talker and language) and block (control and orthogonal). There was no main effect of listener language group, $F(2,51) = 1.571$, $p = 0.218$, and no main effect of task, $F(1,51) = 0.526$, $p = 0.472$. There was a significant main effect of block, $F(1,51) = 24.437$, $p < 0.0001$, such that control blocks (M = 1504.47 ms) were responded to more quickly than orthogonal blocks (M = 1592.07 ms), indicating the presence of Garner interference. There was no interaction between listener language group and task, $F(2,51) = 2.540$, $p = 0.089$, and no interaction between listener language group and block, $F(2,51) = 2.688$, $p = 0.078$. There was a significant interaction between task and block, $F(1,51) = 6.715$, $p = 0.012$, such that the difference between control and orthogonal blocks was larger for the talker task than for the language task. Follow-up simple effects revealed a significant difference between control blocks (M = 1485.75 ms) and orthogonal blocks (M = 1615.39 ms) for the talker task, an average difference of 129.64 ms, $t(53)$

$= -4.330$, $p < 0.0001$; and a significant difference between control blocks ($M = 1523.18$ ms) and orthogonal blocks ($M = 1568.74$ ms) for the language task, an average difference of 45.56 ms, $t(53) = -2.641$, $p = 0.011$. In other words, there was asymmetrical Garner interference, such that there was a greater amount of Garner interference for the talker task than there was for the language task. There was no interaction between listener language group, task, and block, $F(2,51) = 0.877$, $p = 0.422$. Lacking a significant main effect of listener language group or any significant interactions involving listener language group, listener language groups will not be analyzed separately with respect to Garner interference.

Overall, the main effect of block indicates the presence of Garner interference across listener groups. Furthermore, the interaction between task and block reveals that the interference was mutual and asymmetrical; it was harder for listeners to ignore language when processing talker (hence greater interference in the talker task) than it was for them to ignore talker when processing language. The lack of interaction between listener language group and block reveals that this interference effect was stable across listener language groups. This pattern of mutual and asymmetrical interference is evident in Figure 3.9, below. For all listener language backgrounds, it is evident that there is greater interference in the talker task than in the language task. The overall degree of interference for the non-Mandarin-English bilinguals appears to be greater overall than either of the other listener groups, but this trend is not statistically significant. The amount of asymmetry appears to be less for the Mandarin-English bilinguals than for the other two groups, but, again, this trend is not confirmed statistically.

Figure 3.9. Experiment 2, Garner interference. Bars represent group means of differences between orthogonal and control blocks, collapsed over participant. Dots represent individual participants' differences between the two blocks.

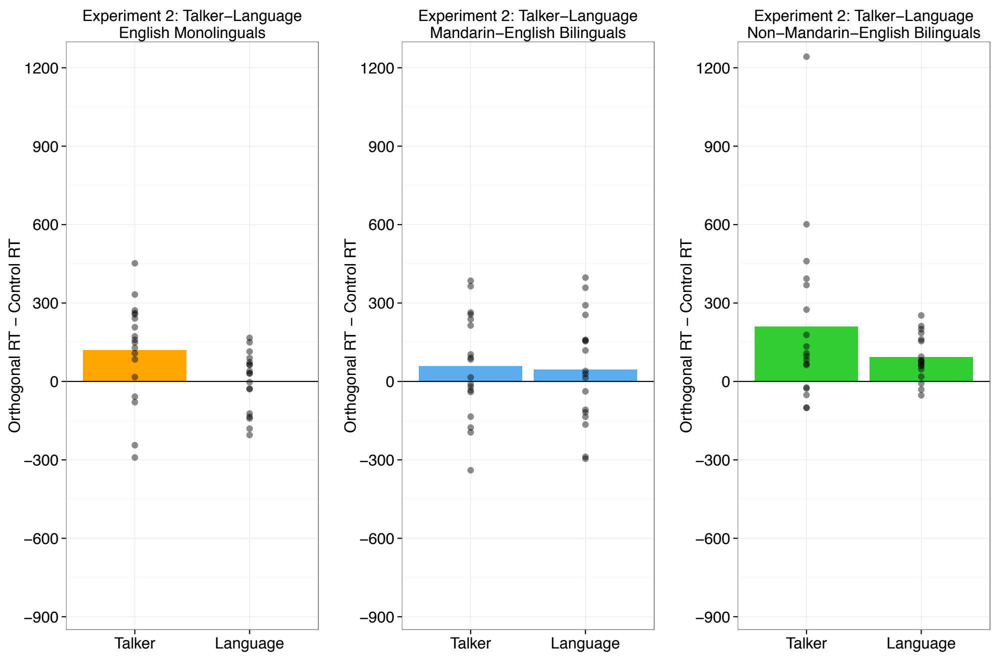


Figure 3.10 below plots participants' levels of Garner interference for the talker and language tasks individually, as shown in Figure 3.6 for Experiment 1.

Figure 3.10. Experiment 2, Garner interference for each task, by participant. Error bars represent +/- 1 SE of the mean.

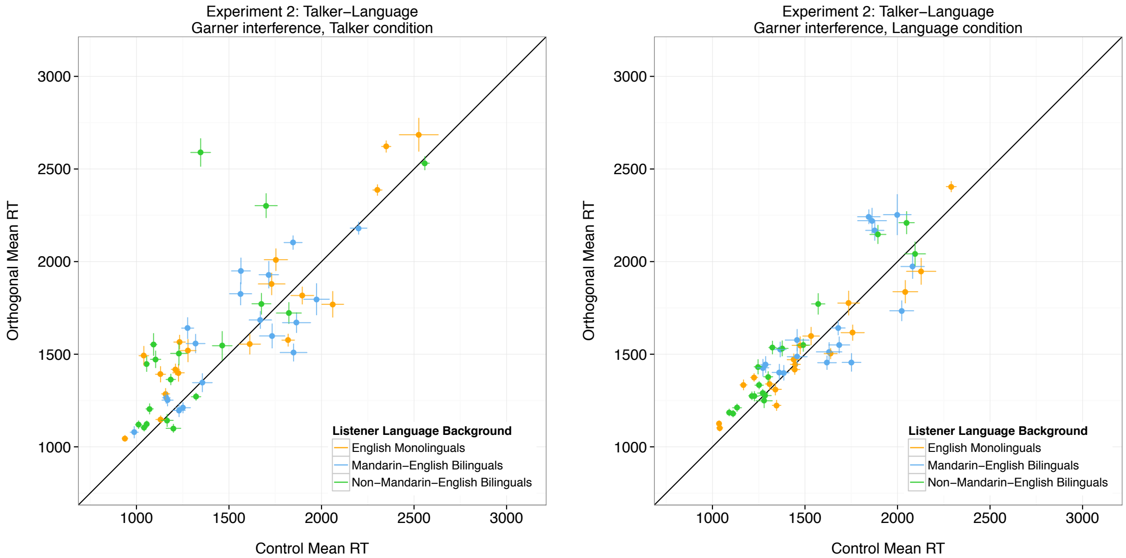
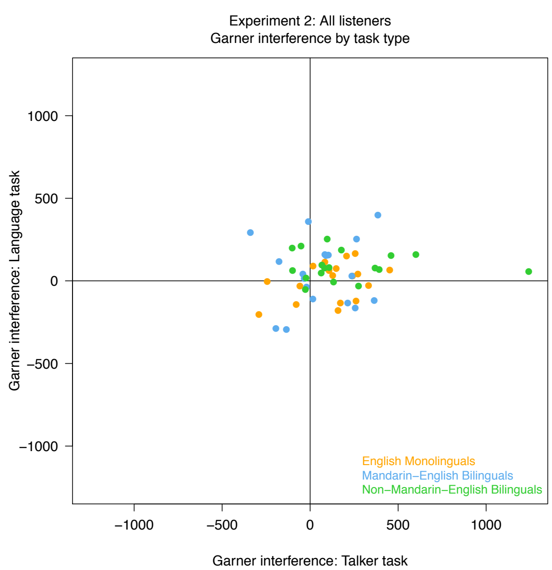


Figure 3.11 below visualizes individual participants' levels of Garner interference as a function of the task for Experiment 2, as shown for Experiment 1 in Figure 3.7.

Figure 3.11. Experiment 2. Garner interference for both tasks, by participant.



3.7.6.2.3 Experiment 2: Redundancy gain?

To determine whether participants exhibited a redundancy gain, mean log-transformed reaction times were calculated for each participant for each block (correlated and control only) in each task. These were submitted to a 3 x 2 x 2 repeated measures ANOVA containing the between-subjects factor of listener language group (English monolinguals, Mandarin-English bilinguals, and non-Mandarin-English bilinguals) and the within-subjects factors of task (talker and language) and block (correlated and control). There was no main effect of listener language group, $F(2,51) = 2.034$, $p = 0.141$; no main effect of task, $F(1,51) = 0.032$, $p = 0.859$. There was no main effect of block, $F(1,51) = 2.249$, $p = 0.140$, signaling that there was not a redundancy gain overall. There was no interaction between listener language group and task, $F(2,51) = 1.402$, $p = 0.255$, and no interaction between listener language group and block, $F(2,51) = 0.688$, $p = 0.507$. There was a significant interaction between task and block, $F(1,51) = 11.749$, $p = 0.001$. Follow-up simple effects testing revealed that in the language task, correlated blocks (M = 1433.78 ms) were performed significantly faster than control blocks (M = 1523.18 ms), an average difference of 89.4 ms, $t(53) = 4.479$, $p < 0.0001$, indicating redundancy gain. In the talker task, however, the opposite trend was found, though it was not significant; correlated blocks (M = 1539.13 ms) were performed numerically slower than control blocks (M = 1485.75 ms), an average difference of -53.38 ms, $t(53) = -1.390$, $p = 0.170$, accounting for the lack of main effect of block. There was no interaction between listener language group, task, and block, $F(2,51) = 0.003$, $p = 0.100$.

In sum, while there was no main effect of block (and thus no redundancy gain overall), follow-up testing on the significant interaction between task and block indicated that there was a redundancy gain for the language task. However, there was no difference between blocks for the talker task, though the means trend toward redundancy loss, where the correlated block was actually responded to more slowly than the control block. The lack of interaction between listener language group and block reveals that this pattern was stable across listener language groups. Lacking a significant main effect of listener language group or any significant interactions involving listener language group, listener language groups will not be analyzed separately with respect to redundancy gain.

3.7.7 Experiment 2: Summary of results

Table 3.19, below, gives a summary of results of Experiment 2. Section 3.9, at the end of this chapter, will assess how well the results of this experiment were predicted by each hypothesis raised above.

Table 3.19. Summary of results from Experiment 2. Listener language group did not interact with task or block in any analysis.

Experiment 2: Results summary, all listener language groups	
Control block performance	No difference between Talker and Language in Control blocks $F(1,51) = 3.673$, $p = 0.061$
Garner interference? *(Orthogonal - Control)*	Yes. Control blocks (M = 1504.47 ms) faster than Orthogonal blocks (M = 1592.07 ms) $F(1,51) = 24.437$, $p < 0.0001$
Symmetrical Garner interference?	No: asymmetrical. Greater interference in Talker (129.64 ms diff) than in Language task (45.56 ms diff) $F(1,51) = 6.715$, $p = 0.012$
Redundancy gain? *(Control - Correlated)*	Significant redundancy gain in Language task (89.4 ms diff). No difference in Talker task, but trend toward redundancy loss Language: $t(53) = 4.479$, $p = 0.0001$; Talker: $t(53) = -1.390$, $p = 0.170$

3.8 Experiment 3: Amplitude–Language-being-spoken

3.8.1 Experiment 3: Rationale

Experiment 3 paired the language-being-spoken dimension with amplitude, a non-linguistic dimension, which is likely to differ from gender and talker in a number of ways, detailed in Section 3.3, above. Experiment 3 was conducted using listeners from two language backgrounds, English monolingual listeners and Mandarin-English bilingual listeners.

3.8.2 Experiment 3: Hypotheses and predictions

As described above, there are several predictions for the results of Experiment 3. First, in terms of processing speed at baseline, amplitude is predicted to be faster than language-being-spoken. Next, in terms of dependency relations, several predictions hold. Following the levels of processing hypothesis, language-being-spoken and amplitude will show some type of asymmetrical interference such that amplitude interferes more with language-being-spoken than vice versa, if predictions about processing speed are confirmed. If amplitude turns out not to be faster than language-being-spoken in control blocks, the levels of processing hypothesis will be refined to reflect the observed levels of processing. With regard to listener language backgrounds, the hypotheses related to more detailed representations of dimensions are not expected to apply to amplitude, whose relationship to language-being-spoken likely does not vary by language background. The bilingualism enhances selective attention hypothesis, however, predicts less interference for Mandarin-English bilinguals relative to English monolinguals.

3.8.3 Experiment 3: Participants

The monolingual group of listeners in Experiment 3 consisted of 18 English-speaking Northwestern undergraduates. These participants were recruited using the Northwestern Linguistics Department subject pool, and were given partial course credit for their time. The monolingual group was comprised of 7 males and 11 females, with the mean age 19.3 (range 18-21). Participants in this group reported using English 93% of the time on average (range 65%-100%), with the exception of one participant described in Footnote 11 (to Section 3.5.2), who reported using English only 65% of the time. Thirteen of these participants did not learn a language other than English before age 7. The five participants who did report learning a language before age 7 currently use that language on average 3% of the time (range 1%-6%). No participant in this group reported knowing any Mandarin Chinese. Fourteen additional participants were run but their data were excluded unanalyzed because they either had knowledge of Mandarin (5), did not conform to the language background requirements of the monolingual group (i.e. were bilingual in languages other than Mandarin) (8), or experienced a technical glitch (1).

The 18 Mandarin-English bilingual participants who participated in Experiment 1a were recruited from the greater Northwestern community. These participants received compensation in cash for their time (less than 1 hour). Of these 18 Mandarin-English bilingual participants, 7 were male and 11 were female, with a mean age of 24.9 (range 22-30). On average, the participants in this group reported learning English at age 10.3 (range 6-14), and report using English an average of 33% of the time (range 10%-70%). These participants have lived in an English speaking country for an average of 1.4 years (range 0-4 years). Two additional participants were run but their data were excluded unanalyzed because they either did not conform to the language background requirements of the bilingual group (1), or had a hearing impairment (1).

3.8.4 Experiment 3: Stimuli

3.8.4.1 Experiment 3: Stimulus materials

Sixty-four sentences of English and 64 sentences of Mandarin Chinese were used in this experiment, as described in Section 3.5.3.1.

3.8.4.2 Experiment 3: Stimulus talker characteristics

The sentences in Experiment 3 all were read by a male talker (named Male 1 in Table 3.10). The male talker was 25 years old at the time of recording and had an overall Versant score of 50 (and a Versant fluency score of 54 and a Versant pronunciation score of 49). Refer back to Table 3.10 to compare characteristics of talkers used across experiments.

3.8.4.3 Experiment 3: Acoustic characteristics of stimuli

3.8.4.3.1 Duration of stimuli

As given in Table 3.14, in Experiment 3 the mean duration of all 128 stimuli produced by the male talker (including both Chinese and English sentences) was 1750.00 ms. The amplitude manipulation produces no difference in duration.

3.8.4.3.2 Amplitude of stimuli

The stimuli in Experiment 3 were amplitude-normalized to 45 dB SPL for the soft level and 75 dB SPL for the loud level, using Praat (Boersma & Weenink, 2011). A 30 dB difference is known to be salient for listeners with normal hearing (e.g. Bradlow et al., 1999, used 65 dB SPL and 35 db SPL, a difference of 30 dB), and initial pilot testing confirmed that listeners had no trouble distinguishing between the two levels. Pilot testing also confirmed that, once the levels were finalized, no participant complained of the loud stimuli being too loud or that the soft stimuli were too soft. High accuracy scores on the language-being-spoken task for "soft" stimuli give additional evidence that these stimuli were sufficiently audible.

3.8.4.4 Experiment 3: Stimulus selection and arrangement

In Experiment 3, stimuli were selected and arranged in the manner reported in Section 3.5.3.4.

3.8.5 Experiment 3: Procedure

The procedure outlined in Section 3.5.4 was followed in Experiment 3. For Experiment 3, the buttons labels on the button box for the amplitude task were "LOUD" and "SOFT", and for the language-being-spoken task were "CHINESE" AND "ENGLISH". As was explained in Section 3.8.4.3.2, the term "loud" is only used in contrast to "soft", and 75 dB SPL is a comfortable listening level which no participant ever complained was too loud. Instructions given to participants in Experiment 3 are identical to those pictured in Figure 3.3, but changed to reflect the use of the amplitude dimension rather than gender.

3.8.6 Experiment 3: Results

As in Experiments 1 and 2, any trial in which a participant did not respond (or any response that occurred later than 4000 ms after stimulus onset) was recorded as "no response" and was eliminated. In Experiment 3, these exclusions represented 0.26% of all trials for English listeners and 0.16% of all trials for the Mandarin-English listeners.

3.8.6.1 Experiment 3: Accuracy analysis

Responses from practice trials were discarded. Means and standard deviations of accuracy (in percent correct) for each language group as a factor of block and task are presented in the first part of Table 3.20. As in Experiments 1 and 2, accuracy analysis was conducted by first converting percent correct to empirical logit proportions. The accuracy results in Table 3.20 illustrate that performance was near ceiling for all listener language groups in all blocks in all tasks (above 97% correct).

To determine whether there were any differences in accuracy between blocks within each listener group, empirical logit-transformed proportions were calculated for each participant for each block in each task. These were submitted to a 2 x 2 x 3 repeated measures ANOVA containing the between-subject factor of listener language group (English monolinguals and Mandarin-English bilinguals), and within-subjects factors of task (amplitude and language), and block (correlated, control, and orthogonal). There was no main effect of listener language group, $F(1,34) = 0.005$, $p = 0.945$ and no main effect of task, $F(1,34) = 0.064$, $p = 0.802$. There was a significant main effect of block, $F(2,34) = 14.508$, $p < 0.0001$. Planned comparisons revealed that orthogonal blocks were slightly less accurate than both correlated and control blocks. This pattern is evidenced by a significant difference between correlated and orthogonal blocks, $t(71) = 5.251$, $p < 0.0001$, such that correlated blocks ($M = 99.28\%$ correct), were performed more accurately than orthogonal blocks ($M = 98.18\%$ correct), an average difference of 0.1%; a significant difference between control and orthogonal blocks, $t(71) = 4.204$, $p < 0.0001$, such that control blocks (*M* = 99.09% correct) were performed more accurately than orthogonal blocks (*M* = 98.18% correct), an average difference of 0.9%; but no significant difference between correlated blocks and control blocks, $t(71) = -1.169$, $p = 0.246$. There was no interaction between listener language group and task, $F(1,34) = 0.193$, $p = 0.663$; no interaction between listener language group and block, $F(2,34) = 0.174$, $p = 0.841$; no interaction between task and block, $F(2,34) = 0.464$, $p = 0.631$; and no interaction between listener language group, task, and block, $F(2,34) = 2.072$, $p = 0.134$.

Despite near ceiling performance, a very small but reliable difference emerged in block, where orthogonal was slightly less accurate than both correlated and control blocks. However, accuracy was very high across all conditions (greater than 97%), negating the need to investigate a possible speed-accuracy tradeoff. As such, the remainder of the analyses in this experiment will focus on reaction times.

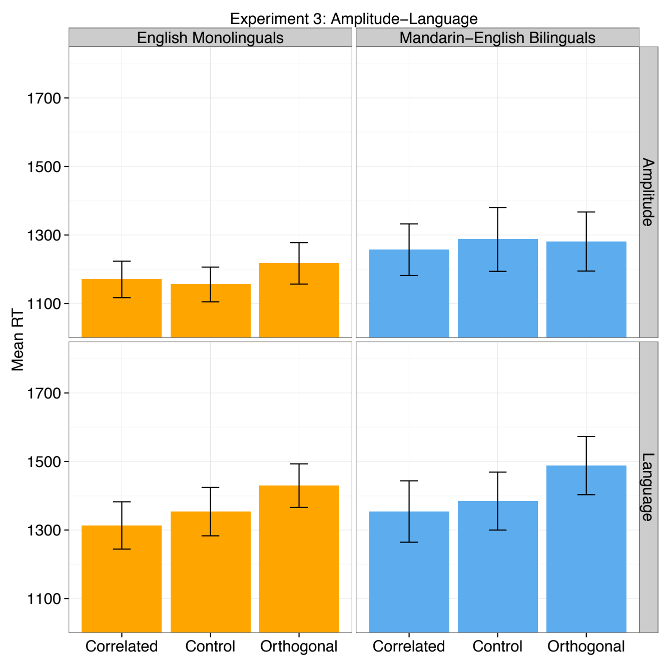
Table 3.20. Experiment 3. Mean by-participant accuracy and reaction times (RTs) both with standard error.

Experiment 3		*Error (%)*				*RT (ms)*			
		English monolinguals		*Mandarin-English bilinguals*		*English monolinguals*		*Mandarin-English bilinguals*	
Task	*Block*	*Mean*	*SE*	*Mean*	*SE*	*Mean*	*SE*	*Mean*	*SE*
Amplitude	*Correlated*	99.05	0.36	99.57	0.21	1170.37	53.17	1257.16	75.30
	Control	98.87	0.47	99.05	0.31	1155.66	50.57	1287.04	93.97
	Orthogonal	98.52	0.35	98.44	0.28	1217.22	60.56	1280.95	86.24
Language	*Correlated*	99.57	0.21	98.96	0.44	1313.39	69.04	1354.03	89.59
	Control	99.13	0.29	99.31	0.32	1353.75	70.63	1384.46	84.55
	Orthogonal	97.83	0.63	97.92	0.63	1429.56	63.58	1488.06	84.88

3.8.6.2 Experiment 3: Reaction time analysis

In order to analyze reaction time data, incorrect responses were discarded. Within the correct responses, data points greater than 2.5 standard deviations from each participant's mean for a particular block in a particular task were removed. For English monolingual listeners, this resulted in removing 2.56% of the data, and for Mandarin-English bilinguals removing 2.52%. Means and standard errors of reaction times for each language group as a factor of block and task are presented in the second part of Table 3.20. Figure 3.12 plots raw reaction times by task and block separately for each listener group. Reaction times were log-transformed before submission to ANOVA and analyzed as in Experiments 1 and 2.

Figure 3.12. By-participant means of reaction times by condition for Experiment 3. Error bars represent +/- 1 SE of the mean.



3.8.6.2.1 Experiment 3: Baseline performance (control block only)

In order to determine how participants performed in the baseline (control) block, mean log-transformed reaction times for the control block were calculated for each participant. These were then submitted to a 2 x 2 repeated measures ANOVA containing the between-subjects factor listener language group (English monolinguals and Mandarin-English bilinguals) and the within-subjects factor task (amplitude and language). There was no main effect of listener language group, $F(1,34) = 0.528$, $p = 0.472$. There was a main effect of task, $F(1,34) = 56.309$, $p < 0.0001$, such that the amplitude task ($M = 1221.35$ ms) was performed faster than the language task ($M = 1369.11$ ms), an average difference of 147.76 ms. There was a significant interaction between listener language group and task, $F(1,34) = 5.445$, $p = 0.026$, such that the difference in reaction times between the amplitude and language tasks in the control block was greater for English monolingual listeners than for Mandarin-English bilingual listeners. Follow-up simple effects testing revealed a significant difference between the amplitude task ($M = 1155.66$ ms) and the language task ($M = 1353.75$ ms) for English monolingual listeners, an average difference of 198.09 ms, $t(17) = -6.792$, $p < 0.0001$; and a significant difference between the amplitude task ($M = 1287.04$ ms) and the language task ($M = 1384.46$ ms) for Mandarin-English bilingual listeners, an average difference of 97.42 ms, $t(17) = -3.749$, $p = 0.002$.

The main effect of task revealed that listeners responded faster to control blocks in the amplitude task than in the language task, as predicted. Thus, the prediction that amplitude is processed more quickly than language was confirmed, and the levels of processing hypothesis did not need to be adjusted for these dimensions. This effect was mediated by the presence of an interaction between listener group and task, such that the task difference was greater for English monolingual listeners than it was for Mandarin-English bilinguals, an unpredicted result.

3.8.6.2.2 Experiment 3: Garner interference?

To determine whether participants exhibited Garner interference, mean log-transformed reaction times were calculated for each participant for each block (orthogonal and control only) in each task. These were submitted to a 2 x 2 x 2 repeated measures ANOVA containing the between-subjects factor of listener language group (English monolinguals and Mandarin-English bilinguals) and the within-subjects factors of task (amplitude and language) and block (control and orthogonal). There was no main effect of listener language group, $F(1,34) = 0.406$, $p = 0.528$. There was a main effect of task, $F(1,34) = 91.257$, $p < 0.0001$, such that the amplitude task ($M = 1235.22$ ms) was performed faster than the language task ($M = 1413.96$ ms), an average difference of 178.74 ms. There was a main effect of block, $F(1,34) = 25.620$, $p < 0.0001$, such that the control blocks ($M = 1295.23$ ms) were performed faster than the orthogonal blocks ($M = 1353.95$ ms), an average difference of 58.72 ms, indicating the presence of Garner interference. There was no interaction between listener language group and task, $F(1,34) = 2.348$, $p = 0.135$, and no interaction between listener language group and block, $F(1,34) = 0.753$, $p = 0.392$. There was an interaction (though marginal) between task and block, $F(1,34) = 4.119$, $p = 0.050$, such that the difference between control and orthogonal was significant for the language task, but not for the amplitude task. Follow-up simple effects revealed a significant difference between control blocks ($M = 1369.11$ ms) and orthogonal blocks ($M = 1458.81$ ms) for the language task, an average difference of 89.7 ms, $t(35) = -4.834$, $p < 0.0001$; but no difference between control blocks ($M = 1221.35$ ms) and orthogonal blocks ($M = 1249.09$ ms) for the amplitude task, an average difference of only 27.74 ms. In other words, there was asymmetrical Garner interference, such that there was interference for the language task but there was no evidence of interference for the amplitude task. There was no interaction between listener language group, task, and block, $F(1,34) = 2.328$, $p = 0.136$.

Across listener language groups and tasks, the main effect of block indicates the presence of Garner interference. The main effect of task indicates that the amplitude task was performed more quickly than the language task overall. These main effects were mediated by a borderline-significant interaction between task and block, which suggests that the Garner interference is asymmetrical and located in the language task (indicating that it was hard for listeners to ignore amplitude when attending to language, but that listeners could ignore language when attending to amplitude). The lack of interaction between listener language group and block reveals that the interference effect was stable across listener language groups. Lacking a significant main effect of listener language group or any significant interactions involving listener language group, listener language groups will not be analyzed separately with respect to Garner interference. The results as depicted Figure 3.13, below, demonstrate the asymmetrical dependency relation. Impressionistically, it also appears that there is a three-way interaction between listener language group, task, and block; for Mandarin-English bilinguals, there is no interference in the amplitude task, suggesting an asymmetrical dependency relation, while the English monolinguals appear to be less different in terms of the interaction of task and block. This view was not supported statistically, however, and the appearance of a three-way interaction is likely driven by outlier participants.

Figure 3.13. Experiment 3, Garner interference. Bars represent group means of differences between orthogonal and control blocks, collapsed over participant. Dots represent individual participants' differences between the two blocks.

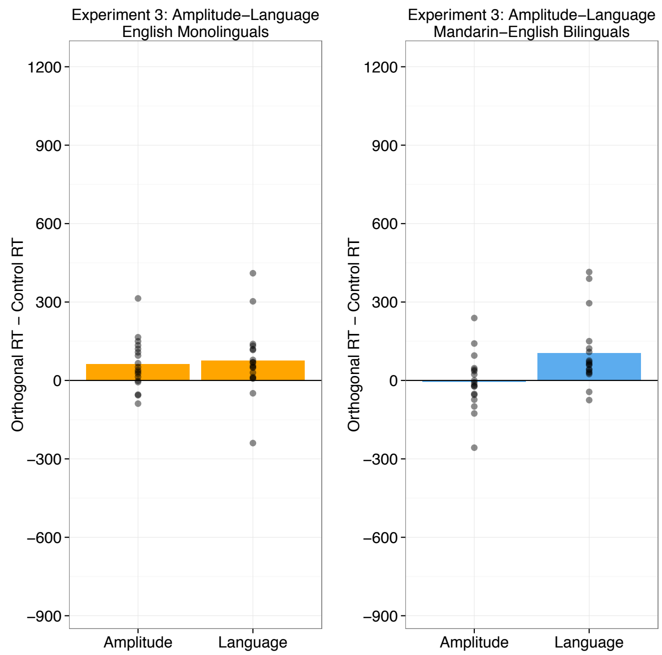


Figure 3.14 below plots participants' levels of Garner interference for the amplitude and language tasks individually, as shown for Experiments 1 and 2 in Figures 3.6 and 3.10.

Figure 3.14. Experiment 3, Garner interference for each task, by participant. Error bars represent +/- 1 SE of the mean.

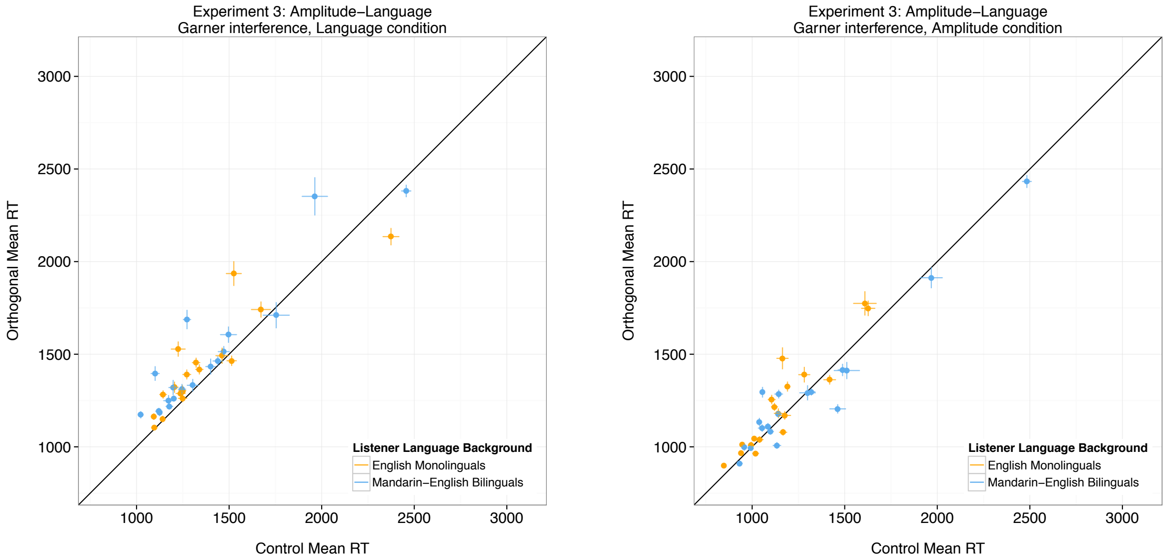
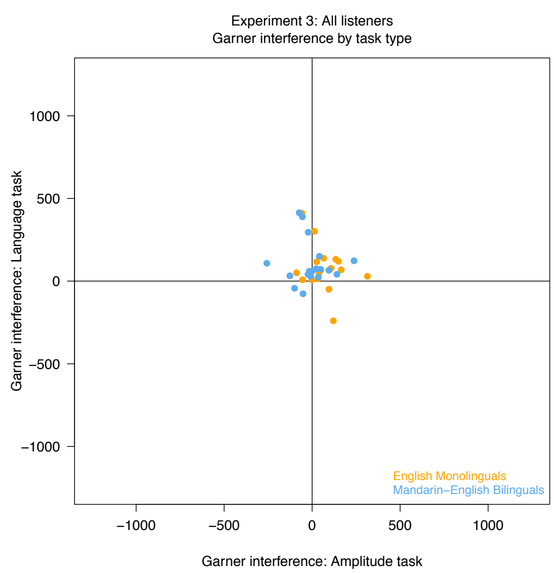


Figure 3.15 below visualizes individual participants' levels of Garner interference as a function of the task for Experiment 3, as shown for Experiments 1 and 2 in Figures 3.7 and 3.11.

Figure 3.15. Experiment 3. Garner interference for both tasks, by participant.



3.8.6.2.3 Experiment 3: Redundancy gain?

To determine whether participants exhibited a redundancy gain, mean log-transformed reaction times were calculated for each participant for each block (correlated and control only) in each task. These were submitted to a 2 x 2 x 2 repeated measures ANOVA containing the between-subjects factor of listener language group (English monolinguals and Mandarin-English bilinguals) and the within-subjects factors of task (amplitude and language) and block (correlated and control). There was no main effect of listener language group, $F(1,34) = 0.501$, $p = 0.484$. There was a main effect of task, $F(1,34) = 46.290$, $p < 0.0001$, such that the amplitude task (M = 1217.56 ms) was performed faster than language task (M = 1351.41 ms) overall, an average difference of 133.85 ms. There was no main effect of block, $F(1,34) = 2.436$, $p = 0.128$, and hence no redundancy gain overall. There were no interactions: there was no interaction between listener language group and task, $F(1,34) = 3.778$, $p = 0.060$; no interaction between listener language group and block, $F(1,34) = 0.107$, $p = 0.746$; no interaction between task and block, $F(1,34) = 1.571$, $p = 0.219$; and no interaction between listener language group, task, and block, $F(1,34) = 0.378$, $p = 0.543$.

In sum, the lack of main effect of block indicates that there was no redundancy gain overall. Lacking a significant main effect of listener language group or any significant interactions involving listener language group, listener language groups will not be analyzed separately with respect to redundancy gain.

3.8.7 Experiment 3: Summary of results

Table 3.21, below, gives a summary of results of Experiment 3. Section 3.9, below, will assess how well the results of this experiment were predicted by each hypothesis raised above.

Table 3.21. Summary of results from Experiment 3. Listener language group did not interact with task or block in any analysis except those listed in this table.

Experiment 3: Results summary, all listener language groups	
Control block performance	Amplitude (M = 1155.66 ms) faster than Language (M = 1353.75 ms), 147.76 ms diff $F(1,34) = 56.309$, $p < 0.0001$ Interaction between listener language group and task indicates greater difference for English monolingual than Mandarin-English bilingual listeners English Monolingual: Amplitude (M = 1155.66 ms); Language (M = 1353.75 ms), 198.09 ms diff, $t(17) = -6.792$, $p < 0.0001$ Mandarin-English Bilingual: Amplitude (M = 1287.04 ms) ; Language (M = 1384.46 ms),97.42 ms diff, $t(17) = -3.749$, $p = 0.002$
Garner interference? (Orthogonal - Control)	Yes. Control blocks (M = 1295.23 ms) faster than Orthogonal blocks (M = 1353.95 ms), 58.72 ms diff $F(1,34) = 25.620$, $p < 0.0001$
Symmetrical Garner interference?	No (though marginal). Interference in Language (89.7 ms diff) but not in Amplitude task $F(1,34) = 4.119$, $p = 0.050$
Redundancy gain? (Control - Correlated)	No. $F(1,34) = 2.436$, $p = 0.128$

3.9 Summary

The main results for all experiments are given in Table 3.22, below.

Table 3.22. Overall pattern of results in terms of control block performance, dependency relations, and dependency relations across listener language groups.

Experiment	*Reaction time at baseline*	*Dependency relation*	*Difference in dependency relation or magnitude of interference between listener groups*
Experiment 1: Gender–Language-being-spoken (ENG & MAN)	L-B-S > Gender	L-B-S ←→ Gender	ENG = MAN
Experiment 2: Talker–Language-being-spoken (ENG, MAN, & NMB)	L-B-S = Talker	L-B-S → Talker	ENG = MAN = NMB
Experiment 3: Amplitude–Language-being-spoken (ENG & MAN)	L-B-S > Amplitude	L-B-S ← Amplitude	ENG = MAN
Key	X > Y : X slower than Y X = Y : X same speed as Y X < Y : X faster than Y	X→Y : mutual and asymmetrical; X interferes with Y more than vice versa X←→Y : mutual and symmetrical X←Y : mutual and asymmetrical; Y interferes with X more than vice versa	X > Y : X has more interference than Y X = Y : X and Y have equal interference X < Y : X has less interference than Y
ENG = English monolinguals MAN = Mandarin-English bilinguals NMB = Non-Mandarin-English bilinguals			

In Experiment 1, gender was faster than language-being-spoken in control blocks. In Experiment 2, there was no difference in reaction time between talker and language-being-spoken in control blocks. In Experiment 3, amplitude was faster than language in control blocks. These results mirror the predictions given in Table 3.1, which were based on prior work on the processing of each dimension in isolation. This also represents the first results to demonstrate the relative speed of language classification with respect to the three other dimensions tested.

Despite the fact that predictions about control block performance were upheld, and thus the levels of processing hypothesis did not need to be adjusted, the use of the levels of processing hypothesis to predict dependency relations (given in Table 3.2) did not stand up, with the exception of the results from Experiment 3. In Experiment 3, language-being-spoken was performed slower (and thus its level of processing is deeper) than amplitude in the control block, and did suffer more interference from amplitude than vice versa (in an asymmetrical dependency relation), as predicted by the mismatch in levels of processing. However, language-being-spoken was also performed slower than gender (in Experiment 1), another mismatch in levels of processing, but this time the dependency relation was mutual and symmetrical, a result unpredicted by the levels of processing account. Further, there was no difference between Experiments 1a and 1b in terms of Garner interference (or any other measure tested), confirming that the relative proficiency of talker pairs did not affect performance on this task. In Experiment 2, the levels of processing match between language-being-spoken and talker did not result in a mutual and symmetrical relation between these dimensions, but instead showed a mutual and asymmetrical relation such that talker suffered more interference from language-being-spoken than vice versa.

The LS/TG hypothesis (Table 3.3) presented specific predictions with respect to the gender–language-being-spoken and talker–language-being-spoken pairings based on prior literature comparing talker-specific versus talker-general processing of language and language-specific versus language-general processing of talker. This prior work inspired the relative language-specificity/talker-generality hypothesis, which predicted that language-being-spoken and talker, and perhaps language-being-spoken and gender, would show a mutual and asymmetrical dependency relation where language-being-spoken interferes more with talker (and gender) than vice versa. This pattern was indeed found for the talker–language-being-spoken pairing. However, it was not demonstrated for the gender–language-being-spoken pairing.

The biggest surprise across the results of all three experiments was that listener language background did not appear to play a role. There were no differences in the amount of Garner interference experienced across listener language groups for any experiment. The only place where listener language group had an effect was in control block performance in Experiment 3 (English monolinguals had a greater difference between amplitude's speed of processing than language-being-spoken's speed of processing than non-Mandarin bilinguals), an unpredicted and puzzling, though minor, result. Three separate hypotheses were put forth (summarized in Tables 3.4, 3.5, and 3.6) predicting the ways in which bilingual listeners would perform differently from monolingual listeners in these tasks. However, none of these predictions were upheld, as there were no differences across listener language backgrounds.

The results from Experiments 1-3 will be discussed in depth as they relate to the various hypotheses put forth, and follow-up analyses will be examined in detail in Chapter 4. New considerations will be proposed that attempt to account for the pattern of results observed.

CHAPTER 4: DISCUSSION OF EXPERIMENTS 1-3

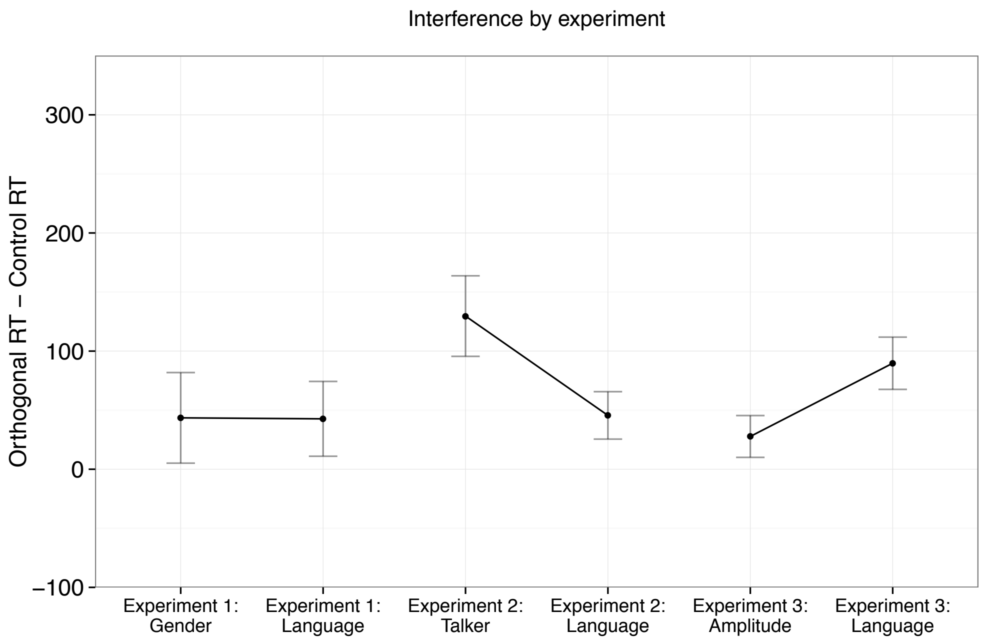
4.1 Chapter outline

This chapter begins with an overall summary of the results of Experiments 1-3. Each experiment is then reviewed in turn, summarizing the predictions of the levels of processing hypothesis (along with the relative language-specificity/talker-generality hypothesis for Experiments 1 and 2), leaving aside the language background hypotheses at first. The results of each experiment are evaluated with respect to each of these hypotheses, and the main findings are summarized. Next, attention is given to the relationship between language-being-spoken and other dimensions in terms of processing time in the control blocks, before a brief analysis attempts to replicate the language familiarity benefit for talker identification. From there the chapter moves on to the language background hypotheses and their inability to predict the results. I review each hypothesis in turn, beginning with the representation hypotheses and then moving onto the selective attention hypothesis. Speculations about how a hypothesis may have been misguided are offered, if relevant. Further, proposals are given suggesting that certain methodological and theoretical reasons may explain why hypotheses were not able to operate in the particular design of the current experiment. Finally, the chapter concludes with a summary.

4.2 Review of results of Experiments 1-3

Language-being-spoken showed some form of integration with each of the three dimensions with which it was paired, a finding that transcended differences in the language background of listeners. The pairings differed with respect to the precise dependency relation observed between dimensions. Experiment 1 (gender–language-being-spoken) showed mutual and symmetrical interference, Experiment 2 (talker–language-being-spoken) showed mutual and asymmetrical interference, and Experiment 3 (amplitude–language-being-spoken) showed asymmetrical interference. These results are depicted in Figure 4.1.

Figure 4.1. Amount of interference for each dimension collapsed across listener language groups, by participant. Lines connect the two dimensions used in each experiment, highlighting the symmetry (Experiment 1) or asymmetry (Experiments 2 and 3) found. Error bars represent +/- 1 SE of the mean. (Note that while Experiment 3 showed significant Garner interference overall, it was asymmetrical such that there was no significant interference in the amplitude task, though there was interference numerically, as shown below.)



In the following sections, these findings will be discussed experiment-by-experiment, and compared against the initial predictions of the levels of processing and LS/TG hypotheses, given in Table 4.1, below.

Table 4.1. Predictions for dependency relations across experiments. Within each hypothesis, the column on the left describes the evidence that the hypothesis uses to make the prediction. The prediction of the hypothesis is shown in the column on the right. The actual results observed in the experiments are given in the rightmost two columns.

	Levels of processing hypothesis		*Language-specificity/ Talker-generality hypothesis (LS/TG)*		*Observed results*	
Experiment	*Observed reaction time in control block*	*Predicted dependency relation*	*Dependency between dimensions as described in literature*	*Predicted dependency relation*	*Observed dependency relation*	*Does any hypothesis support the observed results?*
Experiment 1: L-B-S–Gender	L-B-S > Gender	L-B-S ← Gender	L-B-S processing may not be as dependent on gender processing as the reverse	L-B-S → Gender	L-B-S ←→ Gender	No
Experiment 2: L-B-S–Talker	L-B-S = Talker	L-B-S ←→ Talker	L-B-S processing may not be as dependent on talker processing as the reverse	L-B-S → Talker	L-B-S → Talker	Yes: LS/TG
Experiment 3: L-B-S–Amplitude	L-B-S > Amplitude	L-B-S ← Amplitude			L-B-S ← Amplitude	Yes: levels of processing
Key X→Y: mutual and asymmetrical; X interferes with Y more than vice versa X←→Y : mutual and symmetrical X←Y : mutual and asymmetrical; Y interferes with X more than vice versa						

4.3 Experiment 1

4.3.1 Predictions

In Experiment 1, gender was predicted to be performed faster than language-being-spoken in control blocks (implying a levels of processing mismatch) based on prior work examining the processing characteristics of the two dimensions independently. Given this expected relationship at baseline, the levels of processing hypothesis made the prediction that gender and language-being-spoken would display a some type of asymmetrical dependency relation such that it would be harder for participants to ignore gender when processing language-being-spoken than vice versa.

A competing hypothesis, based on empirical findings documenting the relative language-specificity of talker information and talker-generality of language information, was designed for the dimensions in Experiment 2 but extended to include the gender–language-being-spoken pairing, and predicted that there would be a mutual and asymmetrical dependency relation, such that language-being-spoken interferes more with gender than vice versa. With regard to listener language backgrounds, both of the hypotheses related to having more detailed representations of dimensions (whether stemming from language familiarity or bilingualism, the LF-DR and B-DR hypotheses) predicted that Mandarin-English bilinguals would exhibit more interference than English monolinguals. However, the bilingualism enhances selective attention hypothesis (B-SA) predicted that the opposite would occur (less interference for Mandarin-English bilinguals relative to English monolinguals).

4.3.2 Results

In Experiment 1, gender was classified faster than language-being-spoken at baseline, as predicted based on prior work examining the processing characteristics of the two dimensions independently. The levels of processing hypothesis, then, did not need to be adjusted to reflect the observed processing speed at baseline. The relationship between language-being-spoken and each of the other dimensions in terms of processing speed at baseline will be discussed in detail in Section 4.7 below.

A mutual and symmetrical dependency relation was found for gender and language-being-spoken, counter to the prediction made by the levels of processing hypothesis. According to the levels of processing hypothesis, the mutual and symmetrical dependency relation would be expected if the two dimensions had equal reaction times in control blocks.

As was raised in Section 2.9, whenever one dimension is classified faster than the other at baseline, the issue of discriminability emerges as a possible explanation for dependency relations observed between those dimensions. As previously described, it has been argued that differing amounts of discriminability between dimensions might affect dependency relations because it is harder to ignore irrelevant changes in a dimension that is easier to discriminate (cf. Carrell et al., 1981). This argument suggests that observed dependency relations may simply be due to the theoretically uninteresting difference in discriminability rather than from a "true" interaction between the dimensions caused by their levels of processing. However, such an argument is usually reserved for cases when two dimensions are differentially discriminable and result in an asymmetrical dependency relation. In the present case, however, a symmetrical dependency relation held despite differences in discriminability. Previous work that also found this pattern (symmetry despite difference in discriminability) reported that differences in discriminability do "not jeopardize our interpretation of the Garner interference because orthogonal variation affected the two dimensions equally, even though the dimensions were mismatched in discriminability. Thus, we may still conclude safely that spoken word and loudness interact symmetrically in speeded classification" (Melara & Marks, 1990a, p. 545). Nonetheless, it is logically possible that discriminability differences were a factor in these results.

The matter of discriminability could be tested by altering discriminability on one or both dimensions, making the dimensions more equal in terms of control block reaction time, and then determining whether that change in discriminability affects the dependency relation. In this case, since the gender dimension was performed faster than the language-being-spoken dimension at baseline, the gender dimension would need to be made less discriminable and/or the language-being-spoken dimension would need to be made more discriminable. For example, choosing a male voice that is more female sounding and a female voice that is more male sounding would decrease the discriminability of the gender dimension, presumably resulting in reaction times in control blocks that are more similar to those of the language-being-spoken dimension (i.e., longer). The question, then, becomes whether more matched discriminability between dimensions created by decreasing the discriminability of the gender dimension would lead to a corresponding increase in interference for the gender task. If so, then the difference in discriminability was likely responsible for the symmetrical results observed in the present experiment. If the symmetry persists even after equalizing discriminability, then the symmetry observed in the present experiment may be called a “true” symmetry.

It should be noted that two pairs of talkers were indeed tested for the gender–language-being-spoken pairing (those in Experiment 1a and Experiment 1b), but the results for the control block in a cross-experiment analysis did not show an interaction between task and experiment. This means that any difference in discriminability between the two sets of talkers was not enough to change the difference between gender and language-being-spoken at baseline, thus not providing the appropriate data to test the discriminability hypothesis.

A symmetrical redundancy gain was also present in this experiment, which can be taken as additional evidence for dimensional integration.

The LS/TG hypothesis did not account for the results of Experiment 1. Although it was thought that previous studies investigating the relationship between talker and language-being-spoken might apply to gender and language-being-spoken, this did not turn out to be true. This may reflect differences between the gender and talker dimensions previously unnoted. Thus, both of the hypotheses put forth failed to account for the data in this experiment, which will be discussed in greater detail in Section 4.6.1, below.

None of these patterns differed by listener language background, a result not predicted by any language background hypothesis. Reasons why this may be the case are discussed in detail in Section 4.9, below.

The overall pattern of results was the same across two pairs of talkers (in Experiments 1a and 1b), indicating that the proficiency of the talkers did not affect their discriminability, and generalizing the finding of gender–language-being-spoken integrality beyond one pair of talkers.

4.4 Experiment 2

4.4.1 Predictions

In Experiment 2, talker and language-being-spoken were predicted to be equivalent in processing time in control blocks (implying a levels of processing match) based on prior work examining the processing characteristics of the two dimensions independently. Given this predicted relationship at baseline, the levels of processing hypothesis made the prediction that the two would have a mutual and symmetrical dependency relation. However, based on empirical findings documenting the relative language-specificity of talker information and talker-generality of language information, a competing hypothesis (LS/TG) predicted that there would be a mutual and asymmetrical dependency relation, such that language-being-spoken interferes more with talker than vice versa.

With regard to listener language backgrounds, the language familiarity enhances dimensional representations hypothesis (LF-DR) predicted that Mandarin-English bilinguals would show more interference overall than both English monolinguals and non-Mandarin-English bilinguals. The bilingualism enhances dimensional representations hypothesis (B-DR), however, predicted that both Mandarin-English bilinguals and non-Mandarin-English bilinguals would show more interference than English monolinguals. Finally, the bilingualism enhances selective attention hypothesis (B-SA) predicted that the opposite would be true (Mandarin-English bilinguals and non-Mandarin-English bilinguals should show less interference than English monolinguals).

4.4.2 Results

Talker and language-being-spoken did not differ in reaction time at baseline, as predicted. The dimensions were found to have a mutual and asymmetrical dependency relation, such that language-being-spoken interfered more with talker than vice versa. This result was not predicted by the levels of processing hypothesis, but it was predicted by the relative language-specific/talker-general hypothesis. Further, given that the two dimensions were equally processed at baseline, the discriminability argument supports the finding of mutual and asymmetrical interference as being the “true” dependency relation between the dimensions. It seems, then, that the LS/TG hypothesis, grounded in empirically supported claims about how talker and language-being-spoken might be related, trumped the task-specific hypothesis (levels of processing) in this case, in contrast to the results of Experiment 1, where neither hypothesis was confirmed. This finding both clarifies and complicates the picture with regard to using task-based hypotheses to predict results of Garner experiments. On one hand, it serves as a reminder that dependency relations between dimensions may be best predicted by a close investigation of the properties of those particular dimensions (as in the LS/TG hypothesis). On the other hand, it leaves unanswered the question of when task-based factors (such as levels of processing) might override such dimension-specific factors. In other words, it remains unclear under what circumstances task-based factors may trump specific dimension-based factors, or vice versa.

There was a redundancy gain for the language-being-spoken task, but not for the talker task, providing further evidence for the asymmetry of the relationship between the two dimensions, though the redundancy gain was present in the task with less interference rather than more interference.

These patterns did not differ by listener language background, a result not predicted by any language background hypothesis. Reasons why this may be the case are discussed in detail in Section 4.9, below. As an aside, the fact that there were no differences across listener language backgrounds for the talker task specifically could be taken as evidence that the use of language-general cues were sufficient in accomplishing talker classification here; English monolinguals and non-Mandarin-English bilinguals, who only have access to language-general cues, did not show differences from Mandarin-English bilinguals, who also have access to language-specific cues, on any metric in this experiment. Language-being-spoken interfered with talker regardless of whether listeners were using only language-general cues or also using language-specific cues.

4.5 Experiment 3

4.5.1 Predictions

In Experiment 3, amplitude was predicted to be performed faster than language-being-spoken in control blocks (implying a levels of processing mismatch) based on prior work examining the processing characteristics of the two dimensions independently. Given this expected relationship at baseline, the levels of processing hypothesis made the prediction that amplitude and language-being-spoken would display some type of asymmetrical dependency relation such that it would be harder for participants to ignore amplitude when processing language-being-spoken than vice versa. This hypothesis further speculated that language-being-spoken may not interfere at all in the amplitude task due to amplitude's more peripheral status as a non-linguistic dimension, and that it has been shown to have different processing characteristics than indexical dimensions.

With regard to listener language backgrounds, both of the hypotheses related to more detailed representations of dimensions (whether stemming from language familiarity or bilingualism, the LF-DR and B-DR hypotheses) did not make predictions about this experiment, as amplitude was not thought to be represented language-specifically. The bilingualism enhances selective attention hypothesis (B-SA), however, predicted that Mandarin-English bilinguals would show less interference relative to English monolinguals for the amplitude–language-being-spoken pairing, as it predicted a bilingual advantage across the board.

4.5.2 Results

Amplitude was performed faster than language-being-spoken in the control block, the predicted result based on prior work examining the processing characteristics of the two dimensions independently. The levels of processing hypothesis, then, did not need to be adjusted to reflect the observed processing speed at baseline. In this experiment, the difference in reaction times between the two dimensions in the control blocks was greater for English monolingual listeners than for Mandarin-English bilingual listeners, which was not predicted by any hypothesis. This disparity represents the only comparison in which listener language groups differed significantly, across all measures for all experiments. The reason why language background interacted with dimensional processing in control blocks is not known.

As predicted by the levels of processing hypothesis, amplitude and language-being-spoken showed an asymmetrical dependency relation (though the statistical significance of the asymmetry was borderline), such that there was interference in the language-being-spoken task but not in the amplitude task. The fact that the interference was not mutual was also predicted by the levels of processing hypothesis, since amplitude operates on a more peripheral level than does language-being-spoken, and may not be represented in memory. Though the interference was not mutual, it is notable that any form of integration was found between amplitude and language-being-spoken, the dimensions most different from each other of all those tested. There was no redundancy gain.

As discussed in the results of Experiment 1, above, discriminability may have played a role in the dependency relations observed between amplitude and language-being-spoken, as the two dimensions were not equally processed in control blocks. It is possible that the fact that amplitude was more discriminable than language-being-spoken at baseline might account for the observed asymmetry. However, to make such an argument would effectively discount the likelihood that amplitude and language-being-spoken operate on different levels of processing, attributing the asymmetry instead to the particular values chosen in each dimension. Given that there are compelling a priori reasons to believe that amplitude is processed at a different level than language-being-spoken (cf. Section 3.3), the discriminability argument seems less plausible.

Nonetheless, the possibility that discriminability, not levels of processing, accounts for these results, could be tested in the manner outlined for Experiment 1, above. That is, since amplitude was performed faster than language-being-spoken at baseline, the amplitude dimension would be made less discriminable and/or the language-being-spoken dimension would be made more discriminable. Then, the dependency relations between the now more closely matched dimensions in terms of discriminability could be compared and the results reevaluated.

These patterns did not differ by listener language background, a result not predicted by the B-SA hypothesis. Reasons why this may be the case are discussed in detail in Section 4.9, below. Although technically the results of Experiment 3 do line up with the predictions of the LF-DR and B-DR hypotheses, those hypotheses were designed to predict language-being-spoken's relationship with gender and talker. The language groups were assumed to behave equally in the amplitude and language-being-spoken pairing not from a principled reason but only due to the fact that amplitude was not expected to behave language-specifically and thus the null hypothesis was defaulted to.

4.6 Summary discussion

In sum, language-being-spoken showed some pattern of interference with all dimensions tested. The levels of processing hypothesis accounted for the results of Experiment 3, but not of Experiments 1 and 2. Gender and language-being-spoken showed mutual and symmetrical interference despite differences in levels of processing, while talker and language-being-spoken showed mutual and asymmetrical interference despite equal levels of processing. The failure of levels of processing to account for dependency relations between these dimensions adds to the long line of previous experiments for which that hypothesis was not supported (refer back to

Table 2.1 in Section 2.8). It is notable that in these experiments the levels of processing account fared best for the pair where the two dimensions were most different from each other.

The relative language-specificity/talker-generality hypothesis accounted for the results of Experiment 2. This result is consistent with the idea that while language processing is affected by talker processing (since talker interfered with language-being-spoken), it is less affected by talker processing than talker processing is by language-being-spoken processing (since talker interfered less with language-being-spoken than vice versa). In other words, talker classification is more language-specific than language-being-spoken processing is talker-specific. Despite the thought that this hypothesis may also extend to the gender−language-being-spoken pairing, it did not, in fact, account for the results of Experiment 1.

Again, none of the language background hypotheses accounted for the results of these experiments, a finding which will be discussed at length in the second half of this chapter.

4.6.1 What may account for the gender−language-being-spoken results?

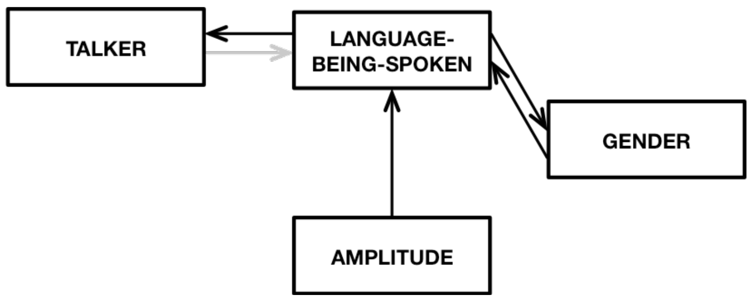
There appear to be explanations for the talker−language-being-spoken pairing (the LS/TG hypothesis) and for the amplitude−language-being-spoken pairing (the levels of processing hypothesis). What may account, then, for the gender−language-being-spoken results? One speculation is that language-being-spoken and gender may be more similar to each other than language-being-spoken and talker in the extent to which the specific values of each dimension have been previously encountered by participants, which may have served to equalize interference between language-being-spoken and gender. Before the task began, it is safe to say that all listeners would have had familiarity with both male and female talkers, and that all listeners were familiar with English. Further, while the Mandarin-English bilinguals were certainly familiar with Mandarin, even the English monolingual listeners had likely heard Mandarin before, or, at least English listeners would have been able to identify Mandarin as "not English" in a two-alternative forced choice task. Thus, there was at least some familiarity on the part of listeners with the values of both the gender and the language-being-spoken dimension. None of the listeners, on the other hand, would have been familiar with the specific talkers they were asked to classify; the talker task represented a choice between two previously unknown entities. While listeners can and are able to learn the difference between the two talkers, this choice is a fundamentally different one than the choice between two genders and two languages. Thus, it is possible that the type of decision involved in processing unfamiliar talkers, though performed as quickly as language-being-spoken decisions at baseline, was harder to selectively attend to than decisions between familiar entities such as languages, resulting in asymmetry. Whether this explanation drove the talker–language-being-spoken results, or served as a supplement to the factors involved in the LS/TG hypothesis, is unknown. Gender and language-being-spoken are more matched along this metric, which perhaps underlies the mutual and symmetrical interference. Thus, it could be that dimensional matches or mismatches in familiarity may mediate dependency relations in the same way that matches or mismatches in contrastiveness was found to do (e.g., Lee & Nusbaum, 1993; Repp & Lin, 1990, discussed in detail in Section 2.7.4). Though it is highly speculative, such an explanation aligns with the symmetry observed between language-being-spoken and gender and the asymmetry observed between language-being-spoken and talker. (A similar line of thinking is explored in Section

4.9.5, below, regarding English monolinguals' versus Mandarin-English bilinguals' familiarity with stimuli in Mandarin.) The role of dimensional value familiarity in these results could be assessed in a future experiment by using two talkers with which participants were familiar. If unfamiliarity with talkers was the source of asymmetry between talker and language-being-spoken (rather than the LS/TG hypothesis), then a version of the experiment with familiar talkers would result in symmetry between the two dimensions, and would account for the symmetry observed between gender and language-being-spoken in Experiment 1.

4.6.2 Hierarchy of dimensions tested

Having determined the dependency relations observed across all experiments, it is now possible to describe the dimensional hierarchy of interference that participants exhibited in processing these dimensions. Dependency relations between dimensions are depicted in Figure 4.2, below. Of course, the schematic would be complete with the addition of the amplitude–gender, amplitude–talker, and gender–talker comparisons, none of which were made here or exist in previous literature. This processing hierarchy appears to be the same for monolinguals and bilinguals, regardless of familiarity with the languages tested. This hierarchy will be used to make predictions for the free classification experiment of this dissertation, Experiment 4, which will be described in Chapter 5. The free classification paradigm will be used to gather further insight into the interactions between these dimensions.

Figure 4.2. Schematic of interference between all dimensions tested in this dissertation. The direction that the arrow is facing indicates the direction of interference. For example, the arrow pointing toward talker from language-being-spoken means that language-being-spoken interferes with talker. The color of the arrow indicates the degree of interference. Light arrows indicate less interference with respect to the dark arrow between the same pair of dimensions. The vertical distance between dimensions indicates their relative speed of processing at baseline as compared to language-being-spoken, from fastest at the bottom to slowest at the top.



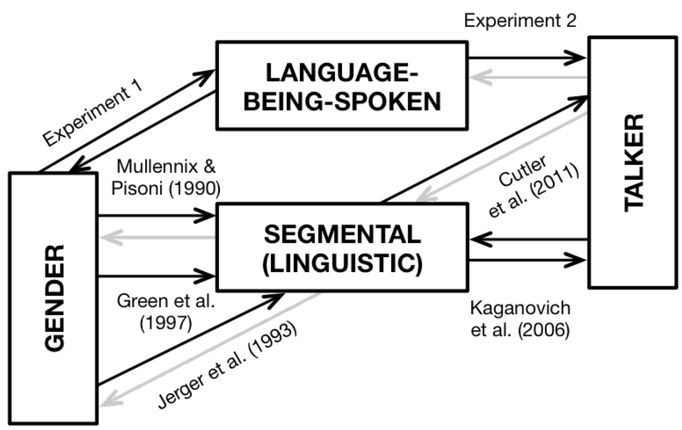
All together, though the pattern of results is clear, it is still difficult to resolve the relative contributions of levels of processing, discriminability, and dimension-specific theoretical hypotheses to these results. Others have previously remarked on the difficulty in pinpointing the root causes of interference in the Garner task: "…patterns of performance observed in various tasks of speeded classification may resist simple theoretical accounts because the causes of Garner interference… arise from a variety of sources" (Ben-Artzi & Marks, 1999, p. 594). Accordingly, the patterns of Garner interference observed here are clear, yet attributing these results to specific mechanisms remains a challenge for this paradigm.

4.6.3 Hierarchy of indexical and linguistic dimensions

Having used the results of Experiments 1-3 to determine listeners' hierarchies of dimensions, I now situate the indexical hierarchies within previous work comparing indexical and linguistic dimensions. As already noted, linguistic dimensions (specifically, the segmental dimensions consonant and/or vowel) and indexical dimensions have previously been found to show integration. In segmental–gender pairings, most results showed asymmetry such that it was harder to ignore the indexical dimension than the segmental dimension (Jerger et al., 1993; Mullennix & Pisoni, 1990 for mutual asymmetry; Green et al., 1997 for asymmetry). The segmental–talker pairings found either mutual asymmetry such that was harder to ignore the segmental dimension than the indexical dimension (Cutler et al., 2011) or symmetry (Kaganovich et al., 2006) between dimensions. To these results pairing indexical dimensions with linguistic dimensions, this dissertation adds the results of indexical dimensions paired with the indexical dimension language-being-spoken. The present findings further enrich the understanding of listeners' hierarchies of indexical dimensions. In order to schematize what is known about indexical dimensions' interactions in processing with linguistic dimensions and with the indexical dimension language-being-spoken, the results of Experiments 1 and 2 are presented alongside previous indexical–segmental results in Figure 4.3, below.

The difference between the results of Experiments 1 and 2, that different dependency relations were found between talker and language-being-spoken versus between gender and language-being-spoken, again underscores the importance of testing many examples of dimensions conveying the same type of information–here, indexical–before making claims about indexical dimensions more generally. For example, after only testing language-being-spoken as paired with gender, it might have been concluded that language-being-spoken has a mutual and symmetrical dependency relation with all other indexical dimensions, which was not found to be true after pairing it with talker. As can be seen in the figure below, the need for this caution is also evident in the varied results of different segmental–indexical pairings (and even the varied results within the same segmental–indexical pairing). Thus, it is important to approach the testing of each pairing of dimensions individually, taking into account the various properties of each dimension.

Figure 4.3. Schematic of interference between indexical dimensions tested in this dissertation alongside previous indexical–segmental results. The direction that the arrow is facing indicates the direction of interference. For example, the arrow pointing toward talker from language-being-spoken means that language-being-spoken interferes with talker. The color of the arrow indicates the degree of interference. Light arrows indicate less interference (with respect to the dark arrow between the same pair of dimensions). The vertical distance between dimensions indicates their relative speed of processing at baseline as compared to language-being-spoken, from fastest at the bottom to slowest at the top. As different studies have found different relative processing speeds between pairs of dimensions, the boxes for dimensions talker and gender are placed vertically on this plot to reflect this fact.



4.6.4 Intraclass variance and salience

The patterns found in Experiment 1-3 can be further examined by couching them in various descriptive schemas discussed in Chapter 2, namely intraclass variance and salience. First, the concept of intraclass variance, suggested as a possible source of Garner interference, is a useful construct in discussing these results. As discussed in Section 2.5, what Melara & Marks (1990b) referred to as stimulus-level processing occurs when dimensions are integral, and involves the existence of intraclass variation within a dimension. Though intraclass variance was not initially applied to dimensions with asymmetry, it is extended here and applied to such a case. In Experiment 2, for example, according to Melara & Marks the integrality observed between talker and language-being-spoken meant that for participants, there were effectively four different types of stimuli: Wei in Mandarin, Wei in English, Li in Mandarin, and Li in English. When faced with a talker decision, participants had trouble with the intraclass variance; the fact that there were two different languages for each talker distracted from the ability to classify stimuli as one talker or the other. On the other hand, when faced with the language-being-spoken decision, participants had less trouble with the intraclass variance; it was not as important that there were two talkers speaking each language.

Likewise, the concept of salience is a useful way to think about these results. As described in Section 2.5, a dimension is salient if it captures attention. In the gender–language-being-spoken pairing, when doing the language-being-spoken task, gender is quite salient. However, it is no more salient than language-being-spoken, since language-being-spoken likewise captures attention during the gender task. In the talker–language-being-spoken pairing, however, talker is less salient when doing the language-being-spoken task; it does not capture as much attention. Language-being-spoken, in contrast, does capture attention when performing the talker task. In the context of language-being-spoken, then, it may be said that gender is more salient than talker (this is also the case with amplitude). It is hard to ignore amplitude and gender when attending to language-being-spoken, but easier to ignore talker (though talker cannot be ignored completely).

Though these concepts may not offer much to the results in the way of explanatory power, it is helpful to cast them in a different frame. Salience, especially, will be useful in comparing the present results with those of Experiment 4, discussed in the next chapter.

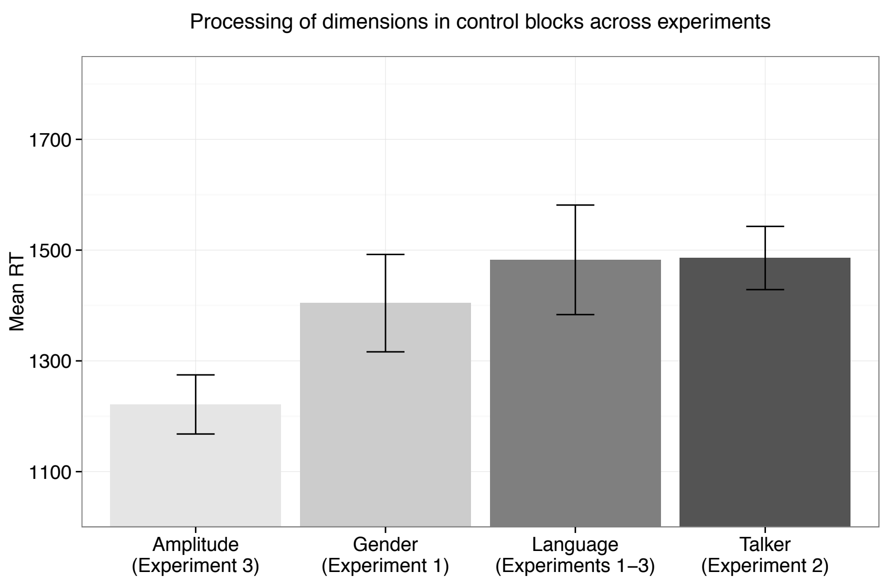
4.7 Language-being-spoken's position in processing at baseline

Before moving on to discuss the language background hypotheses, it is important to point out what has been learned about the processing time of language classification in relation to the other dimensions tested. While these experiments were not explicitly set up to assess how quickly listeners can classify which language is being spoken, the present data can be used to provide some insight into this question, which has not previously been addressed in the literature.

Figure 4.4, below, depicts the mean reaction time in the control block of each of the four dimensions used in these experiments, arranged in ascending order by RT. The amplitude, gender, and talker by-participant means come from the one experiment in which each dimension was included, and are collapsed across the listener language groups. As language-being-spoken was included in all three experiments, its by-participant mean is collapsed across experiments, and across listener groups within those experiments. From the figure it is evident that amplitude is performed much faster than the other dimensions (M = 1221.35 ms), followed by gender (M = 1404.26 ms), followed by language (M = 1482.46 ms) and talker (M = 1485.75 ms), whose mean RTs are nearly identical to each other. In order to assess this relationship statistically, a one-way ANOVA containing the between-subjects factor task (gender, talker, amplitude, and language) was performed on log-transformed reaction times. (As different listeners performed each experiment, and not all listener language groups were included in each experiment, a repeated measures ANOVA could not be conducted.) The ANOVA revealed a significant main effect of task, $F(3,248) = 6,189$, $p = 0.0005$. A post-hoc Tukey HSD test indicated that amplitude significantly differed from talker and from language at the 0.05 level of significance. No other comparisons were significant.

Though gender and language-being-spoken differed significantly at baseline in the results of Experiment 1, in this cross-experimental analysis, there is no significant difference in reaction time between the two. This is likely due to the fact that language-being-spoken, itself, tended to be responded to more quickly in other experiments (namely Experiment 3), bringing down its overall reaction time.[20]

Figure 4.4. Each dimension's mean reaction time in control blocks, collapsed across listener groups. The value for the language dimension is averaged across three experiments. Error bars represent +/- 1 SE of the mean.



[20] The fact that language-being-spoken appears to be processed faster in the control block for Experiment 3 than in Experiments 1 and 2 (compare the control block values across Experiments in Tables 3.15, 3.18, and 3.20 in Chapter 3) is likely caused by its pairing with amplitude in that experiment. Even though the control block of the language-being-spoken task isolates that dimension from variability from the other dimension, there are still ways in which performance could differ across experiments. For example, in Experiment 3, there was only one talker, as opposed to two talkers in Experiments 1 and 2. Further, the language-being-spoken control block in Experiment 3 was presented in either a loud or a soft amplitude, another difference from the other experiments. Finally, participants who had the amplitude task first, or who did not have the control block first in the language-being-spoken task, may have entered a more rapid response mode because of the ease of making amplitude judgments, which then bled into language-being-spoken responses. These are all reasons why future studies are necessary to assess the processing time of language classification in isolation, without the influence of these variables.

In order to assess whether these patterns differed by language background, a 2 x 4 ANOVA containing the between-subjects factors listener language (English monolinguals and Mandarin-English bilinguals) and task (gender, talker, amplitude, and language) was performed on log-transformed reaction times. Because non-Mandarin-English bilingual listeners did not participate in all tasks (only in talker and language-being-spoken), that listener group was not included in this analysis so that the interaction between listener language and task could be tested. There was again a main effect of task, $F(3,242) = 6.198$, $p = 0.0005$, but no main effect of listener language, $F(2,242) = 1.976$, $p = 0.141$, and no interaction between task and listener language, $F(4,242) = 0.595$, $p = 0.666$. The pattern of responses to dimensions at baseline observed above, then, appears to hold for both monolinguals and bilinguals.

Thus, when listening to sentence-length stimuli, listeners (across language backgrounds) process language-being-spoken significantly more slowly than amplitude, but at roughly the same speed as gender and talker (though language-being-spoken is processed more slowly than gender numerically, it is not significantly different in this overall comparison). These data represent the first report of the time course of language classification. The lack of previous work in this arena is surprising given the importance of language identification for bilingual speech processing, and also its relevance in providing benchmarks for speech and language technologies. Going forward, studies using paradigms with finer temporal resolution (using a methodology such as EEG) performed over more controlled stimuli (in order to see what types of linguistic cues are made use of and when) and without influence from other dimensions would provide great insight. More work is needed in this area to further refine our understanding of the process of language classification.

4.8 Language familiarity benefit for talker identification?

At this point, I will make a brief digression to explore whether the present data replicate the language familiarity effect for talker identification. Although this investigation was not built into the design of the experiments, it is possible to test for this effect in the current data. In order to do so, I will examine performance on control blocks in the talker task of Experiment 2.

As discussed previously, the language familiarity effect holds that listeners are better (in terms of accuracy) at identifying talkers in languages with which they are familiar, though accuracy is still above chance for languages with which they are unfamiliar. In the data for Experiment 2 specifically, the language familiarity effect would be confirmed if the English monolingual listeners and the non-Mandarin-English bilingual listeners performed the talker task more accurately for stimuli presented in English (the familiar language) than for stimuli presented in Chinese (the unfamiliar language). To test this, for the English monolingual listeners, a one-way repeated measures ANOVA was conducted on by-subject empirical logit-transformed proportions containing the between-subjects factor stimulus language (Mandarin and English). There was no main effect of stimulus language, $F(1,16) = 3.042$, $p = 0.1$, meaning that the language familiarity benefit for talker identification did not apply for these listeners. An identical ANOVA was run for non-Mandarin-English bilingual listeners, where there was also no main effect of stimulus language, $F(1,16) = 1.617$, $p = 0.222$, and thus no benefit of language familiarity.

Upon further examination of the data, it is evident that a ceiling effect limited the potential for observable differences; across English monolingual listeners, there were only 13 talker-classification errors for stimuli presented in Chinese, compared with 8 for stimuli presented in English (representing error rates of 2.29% for Chinese versus 1.39% English), and across non-Mandarin-English bilingual listeners, there were 6 Chinese errors versus 2 English errors (representing error rates of 1.04% versus 0.35%, respectively). It is hard to improve performance on a task where the error rate is so low to begin with. Previous studies demonstrating the talker familiarity benefit have used paradigms more challenging to listeners, which served to degrade performance enough that any benefit from language familiarity could be observed. For example, the number of possible talkers that listeners must identify in these studies is often high: Goggin et al. (1991) tested identification performance on six bilingual talkers (speaking in each of two languages); Winters et al. (2008) used ten bilingual talkers (speaking in each of two languages); and Perrachione & Wong (2007), and Perrachione et al. (2009) used five monolingual talkers from each of two languages; all as opposed to this study's two talkers. The lack of replication of the language familiarity benefit in this study as measured by accuracy is likely due to the fact that performing talker identification on only two talkers is an easy task for all listeners.

Given this limitation, I now examine whether a language familiarity effect may be evident in terms of participants' speed of talker classification. While previous studies have concentrated on measuring this effect in terms of accuracy, it may also be evident in reaction time: listeners may classify talkers more quickly when they are speaking in a familiar language. To test this, the analogous one-way repeated measures ANOVA was performed on by-subject means of log-transformed reaction times for each listener group. There was no main effect of stimulus language for either language background, $F(1,17) = 0.521$, $p = 0.48$ for English monolinguals and $F(1,17) = 0.716$, $p = 0.409$ for non-Mandarin-English bilinguals. This again demonstrates that the language familiarity benefit for talker identification did not apply for these listeners, even as measured by reaction time.

Both of these analyses, however, were conducted only on the subset of total participants run who successfully learned to distinguish between the talkers, and thus were included in the experimental analysis. As was discussed in Chapter 3, a number of participants were not able to perform the talker task accurately, and therefore their results were excluded and new participants were run as replacements. Eight English monolingual participants, one Mandarin-English bilingual participant, and three non-Mandarin-English bilingual participants were excluded for this reason, though different numbers of participants were run overall in each of these groups, making this an imbalanced comparison. Anecdotally, though, based on the number of participants excluded, it appears that the two groups unfamiliar with one of the languages being spoken had the hardest time with this cross-language talker identification task. Regardless of this putative tendency for certain language backgrounds to fail at the task more often than others, the point remains that the data available in these experiments for investigating the language familiarity benefit are unlikely to provide much insight.

4.9 Language background hypotheses

The hypotheses presented in Section 3.4.3 posit the various ways in which listener language groups might differ in their dependency relations between dimensions. Rather than differences, however, it was found that listener language groups showed similar dependency relations in each of the three experiments, as summarized in Table 4.2, below. In other words, there were no differences between listener groups in any experiment with respect to patterns of

Garner interference across dimensions; the interference observed was stable across participants, regardless of language background. Despite the fact that bilinguals can use the language-being-spoken dimension in mental representation to a greater extent than can monolinguals (as was discussed in Section 1.8), it appears that its place in the dimensional hierarchy is the same for both groups. Further, despite suggestions that bilinguals perform better on tasks of selective attention than do monolinguals (as was discussed in Section 1.9), there were no differences between groups on this task. From these results, a potential conclusion to draw is that language-being-spoken is integrated with gender, talker, and amplitude, and that this integration is unaffected by language familiarity and bilingualism. Before accepting this conclusion, however, it is worthwhile to entertain possible theoretical or methodological factors that may have obscured real language background differences. Therefore, in the following sections I will review each language background hypothesis and raise alternative explanations for why it may not have predicted results. If applicable, I discuss possible problems with the hypothesis itself that were initially unforeseen. Then, I describe reasons why the hypothesis might have in fact been at work even though it was not supported by the present experiments, and suggest follow up experiments that might clarify the hypothesis' role.

Table 4.2. Testing predictions made in Tables 3.4-3.6 regarding dependency relations based on listener language backgrounds. The predictions of three hypotheses as well as the observed results are given.

Experiment	*Predicted difference between listener languages*			*Observed difference between listener languages*	*Does any hypothesis support the observed results?*
	Language familiarity enhances dimensional representations hypothesis (LF-DR)	*Bilingualism enhances dimensional representations hypothesis (B-DR)*	*Bilingualism enhances selective attention hypothesis (B-SA)*		
Experiment 1: Gender–L-B-S (ENG & MAN)	MAN > ENG	MAN > ENG	MAN < ENG	MAN = ENG	No
Experiment 2: Talker– L-B-S (ENG, MAN, & NMB)	MAN > ENG NMB	MAN > ENG NMB	MAN < ENG NMB	MAN = ENG	No
Experiment 3: Amplitude–L-B-S (ENG & MAN)	MAN = ENG	MAN = ENG	MAN < ENG	MAN = ENG	No[21]

Key > : More interference than = : Equal interference < : Less interference than
ENG = English monolinguals MAN = Mandarin-English bilinguals NMB = Non-Mandarin-English bilinguals

4.9.1 Language familiarity enhances dimensional representations hypothesis (LF-DR)

The language familiarity enhances dimensional representations hypothesis predicted that listeners familiar with both languages being spoken (Mandarin-English bilinguals) would show more interference than the listener groups only familiar with one of the languages (English monolinguals, and non-Mandarin-English bilinguals in Experiment 2) for both the gender–language-being-spoken and the talker–language-being-spoken pairings. This was predicted based on the rationale that experience with a language enables listeners to develop more detailed linguistic and indexical representations in memory, resulting in increased interference between dimensions on the Garner task. This prediction was not confirmed, and instead listeners with different levels of familiarity with the languages being tested all showed

[21] As explained in Section 4.5.2, the results of Experiment 3 do line up with the predictions of the first two hypotheses given in this table, but only incidentally. The predictions of these hypotheses for Experiment 3 amounted to the predictions of a null hypothesis, since amplitude was not expected to vary according to language background.

equal amounts of interference.

The hypothesis made a secondary prediction, that English monolinguals and non-Mandarin-English bilinguals would show more interference overall on stimuli spoken in English (the language with which they are familiar) versus stimuli spoken in Mandarin (the language with which they are not familiar), in both Experiment 1 and Experiment 2. Due to the nature of the Garner task design, this prediction could not be tested straightforwardly; conducting a 2 x 2 ANOVA with repeated measures on the within-subjects factors block (orthogonal and control) and stimulus language (English and Mandarin) is not possible because it would result in unequal cell sizes.[22] Consequently, the following approach was taken for each experiment. For each listener language group, by-participant means of log-transformed reaction times were submitted to a 2 x 2 x 2 ANOVA with the within-subjects factors block (orthogonal and control), task (gender/talker and language-being-spoken, depending on the experiment) and stimulus language (Mandarin and English). Again, no repeated measures were used due to the inequality of cell means. A significant interaction between block and stimulus language would signify that for that particular language group, there were differential amounts of interference for stimuli spoken in different languages, confirming the secondary prediction of this hypothesis. In Experiment 1, no listener group showed a significant interaction between block and stimulus language. For English monolingual listeners, the only significant result was the main effect of task, $F(1,118) = 6.921$, p

[22] This is the case for several reasons. In the task that is not language-being-spoken (either gender, talker, or amplitude, depending on the experiment), half of the participants were given the control block with English stimuli and the other half were given the control block with Mandarin stimuli. Further, in the task that is not language-being-spoken, the orthogonal block will always contain half as many stimuli of each language than the control blocks (see Figure 3.2 in Chapter 3). The result of these two design features is that it is not possible to conduct valid repeated measures ANOVAs or paired t-tests using these factors due to unequal cell sizes.

= 0.010, and the interaction between block and stimulus language was not significant, $F(1,118) = 0.598$, $p = 0.441$. For Mandarin-English bilingual listeners, there were no significant main effects or interactions, including the interaction between block and stimulus language, $F(1,118) = 0.153$, $p = 0.696$. The same was true for Experiment 2; no listener group showed a significant interaction between block and stimulus language. For English monolingual listeners, there were no significant main effects or interactions, including the interaction between block and stimulus language, $F(1,118) = 1.203$, $p = 0.274$. For Mandarin-English bilingual listeners, the only significant result was the main effect of task, $F(1,118) = 4.019$, $p = 0.047$, and the interaction between block and stimulus language was not significant, $F(1,118) = 0.014$, $p = 0.908$. For non-Mandarin-English bilingual listeners, the only significant result was the main effect of block, $F(1,118) = 4.396$, $p = 0.038$, and the interaction between block and stimulus language was not significant, $F(1,118) = 0.075$, $p = 0.785$. Thus, the secondary prediction of the LF-DR hypothesis was not confirmed. The language of the stimulus did not affect the amount of interference for listener groups, regardless of their familiarity with the languages being tested. As the primary predictions of the LF-DR hypothesis were also not confirmed (i.e. there were no differences in amount of interference between different language backgrounds), this is perhaps unsurprising. Nonetheless, it is valuable to note that the language of the stimulus was not a factor in the interference results.

4.9.1.1 Possible lack of linguistic processing may have obscured language familiarity differences

Having established that neither of the predictions of this hypothesis were confirmed, I now speculate about why this hypothesis did not hold up. One possibility is that a circumstance particular to the experiments prevented the LF-DR hypothesis from operating: participants may not have accessed linguistic levels in performing the task. If participants did not engage, or engaged very little, with the linguistic system during this task, then this hypothesis may not have been applicable. That is, performing the task without accessing the linguistic system would essentially equate listeners across language backgrounds, possibly explaining why no differences between language backgrounds were seen in the results. In this set of experiments, the task of the listener did not, of course, require lexical access; participants only needed to make judgments about indexical and non-linguistic dimensions. However, it was expected for a variety of reasons that listeners would, in fact, undergo linguistic processing in the course of making these judgments. First is the result—by now well-documented in this dissertation—that indexical and linguistic processing are perceptually integrated, indicating some level of linguistic processing when engaging with indexical dimensions (e.g., Mullennix & Pisoni, 1990). Also, the proposal that the familiar language benefit for talker identification is based on having access to the linguistic level (e.g. Perrachione et al., 2009; Perrachione & Wong, 2007) suggests that linguistic processing is recruited in at least one of the dimensions tested, talker classification. For these reasons, then, it was reasonable to make hypotheses about language backgrounds that rested on the assumption that lexical access would take place during these experiments, even when it was not strictly necessary for the task.

However, it is possible that in these experiments listeners adopted a task-based strategy wherein they did *not* engage with the stimuli on a linguistic level. This may be true for a variety of reasons. First, the acoustic cues necessary to perform classification of all dimensions were present from nearly the onset of every stimulus, which listeners may have realized and thus employed a more primitive, auditory-based strategy. In fact, Repp & Lin (1990) similarly speculated about how much the lexical level was implicated in their study of segments versus tones using the Garner paradigm, even though dimensions they tested were linguistic: "we do not know to what extent automatic lexical influences may have operated in the speeded classification task; certainly, the task did not encourage lexical strategies" (1990, p. 494). It is worth pointing out, however, that average reaction times across experiments in this dissertation ranged between 1150 and 1690 ms long, and average stimulus durations were between 1620 and 1860 ms long, meaning that listeners did tend to listen to much of a stimulus before registering a response (though this differed across tasks). Nonetheless, the potential use of a more auditory-based rather than linguistic-level strategy in this task would account for a lack of differences between language backgrounds if the LF-DR hypothesis was actually correct. Recall that this hypothesis was based on the rationale that listeners familiar with a language build up a detailed network of exemplars in that language, both pertaining to indexical and linguistic information. It gained further support from the idea that talker identification in known languages may implicate linguistic processing to some degree (Perrachione et al., 2009), meaning that indexical processing implicates linguistic processing if it is possible to access linguistic levels. If listeners across language backgrounds made use of a very shallow processing strategy whereby they did not access linguistic levels, then in effect all listeners would behave as if they did not have access to linguistic levels.

In order to determine whether this explanation accounts for the lack of differences between language backgrounds, a version of the present experiments could be run that did encourage lexical access, perhaps achieved by telling listeners that they would be tested at the end of each task on memory for sentence content. Presuming that this manipulation would make listeners engage with the linguistic level, the results of this follow-up would help determine whether lack of linguistic processing was a factor in the present results. In the follow-up experiment, if the listeners with familiarity with both languages do show more interference than the other groups, then it is likely that listeners in the present experiments simply were not accessing linguistic levels during the tasks. However, if there are still no differences between listener groups even when linguistic processing is encouraged, then it is unlikely that this explanation accounts for the results of the present experiment. Another approach could be taken with the results of the current experiment to investigate whether listeners might have used a more auditory-based processing strategy, involving an investigation of reaction times over the course of each experiment. If listeners did use such a strategy, it may have only been after performing some number of trials and realizing that lexical access was not necessary, thereby beginning a more rapid response pattern. However, for several reasons, this type of analysis will not be pursued here. First, if an analysis of reaction times throughout each experiment were to reveal that reaction times did get reliably shorter over time, this would not be conclusive evidence that a more auditory-based strategy was used, as it could also be due to a practice effect or some other factor. Further, if participants did switch strategies at a certain point during the experiment, this point would likely be different across participants, making generalizations difficult.

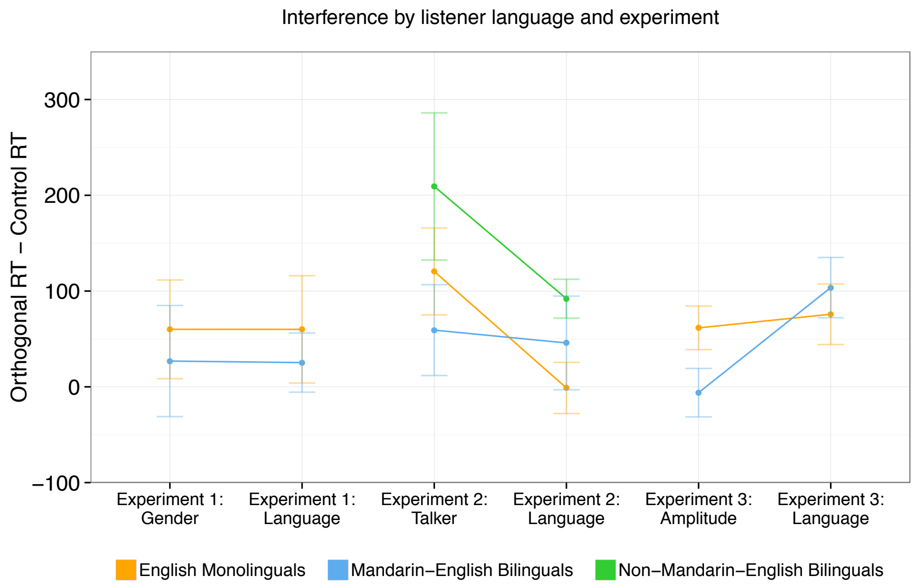
4.9.1.2 Examining interference across listener groups

Before moving on to the next representation-based hypothesis, it is worth examining Figure 4.5 below, which may hint at support of language familiarity as a force operating in at least one of the present experiments. This figure plots the amount of interference in each task across experiments, broken down by listener language background. It is suggestive of the fact that this hypothesis may be on the right track, at least with respect to the talker–language-being-spoken pairing. In this figure, it is evident that the overall mutual and asymmetrical dependency relation shown between talker and language-being-spoken in Experiment 2 (as depicted in Figure 4.1, above) is made up of two listener groups that appear to behave similarly, showing a large asymmetry (English monolinguals and non-Mandarin-English bilinguals), and one listener group that behaves differently (Mandarin-English bilinguals show more balanced interference across dimensions).[23] This was not the precise prediction of the language familiarity enhances dimensional representations hypothesis; that hypothesis predicted more interference overall for English monolinguals and non-Mandarin-English bilinguals than for Mandarin-English bilinguals, which is not exactly what is seen here. Here, the two listener groups unfamiliar with English pattern in a parallel manner, each with greater interference for the talker task than for the language-being-spoken task, while the group familiar with both languages patterns differently (less asymmetry). In short, the two listener groups predicted to pattern together in this hypothesis do so, though the pattern is not exactly as predicted. However, these observations are only trends, and since listener language was not found to be significant in the statistical analyses for interference, these tendencies should not be overly emphasized, especially since the appearance of the interaction between listener groups may be driven by outlier participants (see Figure 3.9). Moreover, this trend only holds for one of the two experiments in which the hypothesis was expected to act; English monolinguals and Mandarin-English bilinguals appear to behave

[23] These patterns can also be observed in the interference plots of Figure 3.9.

similarly in Experiment 1. Nonetheless it is interesting to note that there may be something to the language familiarity hypothesis. Whether the fact that its predictions were unsupported in these experiments is a matter of insufficient statistical power, methodological factors (such as some listeners adopting an acoustic-level processing strategy for the tasks), or an actual disconfirmation of the hypothesis, is a worthy subject of future investigations.

Figure 4.5. Interference across experiments as a factor of listener language background. This figure can be compared with Figure 4.1, above; the plot here is equivalent to Figure 4.1 broken down by language background.



4.9.2 Bilingualism enhances dimensional representations hypothesis (B-DR)

The bilingualism enhances dimensional representations hypothesis predicted that bilingual listeners (Mandarin-English bilinguals, and non-Mandarin-English bilinguals in Experiment 2) would show more interference than monolingual listeners (English monolinguals) for both the gender–language-being-spoken pairing and the talker–language-being-spoken pairing. This was predicted based on the rationale that bilingual listeners have more experience than monolinguals tuning into and making use of indexical representations (especially as related to which language is being spoken), resulting in increased interference between dimensions on the Garner task. This prediction did not hold up, instead finding equal amounts of interference across both bilingual and monolingual listeners.

4.9.2.1 Longer reaction times may have obscured differences in listener language groups based on representations

There are several possible explanations for these results that apply to both representation hypotheses, regardless of where the difference in representation originates. First, it could be the case that one of these representation hypotheses was in fact correct, but that the methodological decision to use sentence-length stimuli in these experiments obscured its effects. The use of longer stimuli was a deliberate choice in this study. For the language-being-spoken classification task, it was thought that participants would require more information than just a monosyllable in order to perform the classification. Further, providing more than a monosyllable would be particularly important in order for listeners to be able to perform the task accurately, since, despite different phonotactic and syllable structure constraints in Mandarin versus English, many monosyllables may be valid words of both English and Mandarin.

However, the use of longer stimuli may have encouraged slower responses, despite instructions to the contrary (as previously noted, verbal instructions were given to all participants encouraging them to respond as soon as they knew the answer, and emphasizing that they did not need to wait until the stimulus had ended before responding). The reaction times in this study do appear to be notably longer than those in other studies using shorter stimuli, even when classifying the same dimensions. For example, Cutler et al. (2011) reported reaction times in the 400-500 ms range for the talker task of their experiment, which used Dutch monosyllables as stimuli, as opposed to RTs of 1300-1600 ms for the talker task in Experiment 2 of this dissertation. Recall that Nakai & Turk (2011) used bisyllabic nonwords, which until this dissertation may have been the longest stimuli used in a Garner task on auditory speech stimuli. That study reports RTs in the 700-900 ms range for phoneme classification and phrase boundary classification tasks. Thus, the use of longer stimuli does appear to have resulted in longer reaction times in this study.

These longer reaction times may have obscured any observable differences between listeners who do and do not have prior representation of dimensions. The two representation hypotheses alleged that listeners with better representations of dimensions (whether they are created because of language familiarity or because of bilingualism) would have more interference than those without as detailed representations (in both hypotheses, the English monolingual listeners, and in the language familiarity version, the non-Mandarin-English bilinguals in Experiment 2). Presumably, according to this mechanism, listeners who already have stored representations of dimensions within a language would show greater interference early in processing as compared to those without stored representations in that language. However, the imbalance between groups may disappear as time elapses if those participants without stored representations are able to build up a representation for a dimension during the task, so then they, too, would show increased interference. In other words, even though English monolingual listeners did not have detailed representations of the various dimensions in

Mandarin, they may have been able to build up enough of a representation for each dimension during the course of each stimulus to match the Mandarin-English bilinguals.

In order to determine whether longer reaction times are responsible for the lack of differences between listener language groups, a follow-up experiment could be run that exactly replicates the current experiments but uses single words as stimuli rather than sentences. On the assumption that these shorter stimuli would induce shorter reaction times, the results of this follow up experiment could help determine whether the longer reaction times in the present study contributed to its results. If in follow-up experiments using words there are still no differences between listener language groups despite shorter reaction times overall, it is unlikely that longer reaction times played a role in the present results. If, however, differences between listener groups emerge in a follow-up experiment using shorter stimuli, it is likely that longer reaction times contributed to the lack of differences in the current study.

As an aside, another potential effect of longer stimuli and thus longer reaction times is that they may overshadow any potential differences in response times between blocks, thus obscuring any Garner interference which may have been present at earlier stages (as was pointed out in Section 3.5.3.1). However, as was also previously mentioned, the average length of stimuli should not change across blocks, so response tendencies for control blocks should not be any different than those for orthogonal blocks. Nonetheless, a Spearman's rank order correlation test was conducted for these experiments, comparing participants' control block reaction times with their level of interference (both log-transformed), in order to determine whether Garner interference may be related to a participant's reaction time in control blocks. Presuming that longer stimuli do cause longer reaction times at baseline, this test will determine whether those long reaction times at baseline are in turn related to different amounts of interference. The test revealed a significant correlation for Experiment 2, $r = -0.39$, $p < 0.0001$, such that Garner interference decreased as reaction times in the control block increased. There was no significant correlation for Experiment 1, $r = -0.20$, $p = 0.09$ or Experiment 3, $r = -0.16$, $p = 0.19$, though the direction of the correlation is the same as in Experiment 2. Therefore, longer reaction times do appear to decrease participants' susceptibility to interference from irrelevant variability, at least in the language-being-spoken–talker task. This may be taken as evidence that, at least for some pairings of dimensions, there might indeed be more interference overall in a version of the task inducing shorter rather than longer reaction times (such as one with shorter stimuli).

That said, the fact that Garner interference was found at all in these experiments can be taken as evidence that sentence-length stimuli are effective stimuli in the speeded classification task. However, the lack of language background differences in interference may call to question the use of longer stimuli anew, if it is indeed possible for listeners without prior representations of dimensions to create them on the fly in a longer stimulus environment. In order to speculate about whether using shorter stimuli may differentially affect the amount of interference for listeners from different language backgrounds, Spearman's rank order correlations were conducted (as above) on results from Experiment 2 (the only experiment in which the correlation between control RT and interference was significant). For this analysis, Spearman's rho was calculated separately for each listener group, again on log-transformed RTs from the control block and amount of interference. These tests revealed a significant correlation for English monolinguals, $r = -0.51$, $p = 0.002$. However, there was no correlation for Mandarin-English bilinguals, $r = -0.26$, $p = 0.120$, or for non-Mandarin-English bilinguals, $r = -0.24$, $p = 0.163$.

Though highly speculative, this suggests that longer reaction times may have caused English monolinguals to show less interference than they might have in a task that induced shorter reaction times, while the bilingual groups may show similar amounts of interference regardless of their baseline RTs. Therefore, it may indeed be the case that English monolinguals would show more interference than the bilingual groups in a follow-up study with shorter stimuli.[24]

4.9.2.2 If Garner interference measures working memory, representations are irrelevant

Going further with the proposal that experience with dimensions may emerge during the course of the task presents the possibility that representations are not implicated at all in the Garner paradigm (at least as it is instantiated here). In discussing the processing origins of Garner interference, Kaganovich et al. (2006) offered the proposal that Garner interference may in fact originate at the level of working memory. While their proposal had to do with the fact that the acoustic cues involved in talker and word identification are multidimensional and overlapping, their claim is relevant here. They explain: “[Garner interference] may be due to the cognitive effort required to create separate categorical representations of dimensions and maintain these representations in working memory while performing the task” (p. 162). Further support for this proposal comes from the fact that interference in the Garner task is based on variability between stimulus presentations, that is, on the block level (as opposed the presence of variability within a stimulus), so the presence of interference entails that working memory is

[24] Though note that, while the amount of interference in Experiment 2 is not significantly different across listener groups, it is greater for non-Mandarin-English bilinguals than for the other two groups, which are approximately equal (see Figure 3.9). Thus, if a follow-up study with shorter stimuli did increase interference for the English monolingual listeners, it remains to be seen how this would affect the pattern of results across listener groups; it may be that non-Mandarin-English bilinguals and English monolinguals would show more interference than Mandarin-English bilinguals, a pattern not predicted by any hypothesis.

being recruited (cf. Boenke, Ohi, Nikolaev, & Lachmann, 2009). This is in contrast to a task like the Stroop test, where interference comes within a trial, thus not requiring as much usage of working memory.

This proposal can be taken to mean that any sort of prior representations of dimensions held in memory by listeners may be irrelevant. Perhaps because it did not encourage deeper-level linguistic processing (as suggested in Section 4.9.1.1 above), this particular version of the Garner task may not have tapped into representations of dimensions, and instead dimensions were built up by listeners during the task, negating the effect of representational differences between language backgrounds. If working memory is indeed responsible for Garner interference, then a follow-up experiment using words instead of sentences (as described in Section 4.9.2.1 above) may still not show the effects of language background; representations would not be utilized in processing regardless of how long listeners spend processing. It is unlikely that this working memory explanation is completely correct, as others have found at least some effect of language background on interference (Lee & Nusbaum, 1993; Repp & Lin, 1990). However, future work comparing interference patterns on learned (and thus represented in long-term memory) versus perceptual dimensions would provide a better understanding of this provocative possibility.

4.9.3 Bilingualism enhances selective attention hypothesis (B-SA)

The bilingualism enhances selective attention hypothesis predicted that bilingual listeners (Mandarin-English bilinguals, and non-Mandarin-English bilinguals in Experiment 2) would show less interference than monolingual listeners (English monolinguals) on all tasks. This was predicted based on previous work showing that bilingual listeners have better executive control, a component of which is selective attention, which should result in decreased interference between dimensions on the Garner task. This prediction did not hold up, instead finding equal amounts of interference across both bilingual and monolingual listeners.

Of course, one possible outcome from this null result for language background is that the findings of the present experiment are added to the list of cases where a bilingual advantage in executive function was not found (cf. Hilchey & Klein, 2011; Paap & Greenberg, 2013). While this may certainly be the case, before coming to this conclusion, a number of possibilities are entertained below suggesting methodological reasons why this hypothesis may have failed.

4.9.3.1 Bilingual advantage may only be present in perceptual dimensions

I now consider possibilities that may account for why this hypothesis was not confirmed. First, it is possible that this hypothesis was not entirely accurate in the first place. In studies comparing bilingual and monolingual children's development, it has been suggested that the selective attention advantage only applies to dimensions deemed *perceptual* rather than *semantic* (Bialystok & Martin, 2004). For example, in Bialystok & Martin (2004), bilingual children showed an advantage versus monolingual children on a dimensional card-sorting task when sorting dimensions based on basic concepts like the color of simple shapes, but did not show an advantage when sorting by more complex concepts, like the color of complex objects, or sorting by objects' functions (e.g. "things you wear" versus "things you play with"). The dimensions under investigation in the current experiment would certainly fall under the category they call semantic (with the possible exception of amplitude), which may mean that there would not be differences between bilinguals and monolinguals when responding to these stimuli. That said, the results claiming an advantage only in perceptual stimuli are based on children and in the visual modality, and the extent to which the perceptual versus semantic distinction applies to adult cognition and the auditory modality, is unknown. Thus, the bilingualism enhances selective attention hypothesis was not an unreasonable one to make for the present experiments even given this knowledge.

4.9.3.2 Bilingual advantage may not apply to young adults

One major caveat to the bilingualism enhances selective attention hypothesis is that the population tested in the present experiments, young adults, has not always demonstrated such an advantage in previous literature (see Bialystok et al., 2012 for a review). Bialystok, Martin, & Viswanathan (2005) reported that for children (5 years old), middle-aged adults (age 30-59), and older adults (age 60-80), bilinguals show less Simon interference than monolinguals. However, no such difference was found for young adults (age 20-30; though the study did not report the average age of participants in this group, it did mention that the participants were undergraduates). The authors attribute the equivalent performance of both groups in this age range to the fact that young adults already are at peak performance in terms of inhibitory control, and therefore any advantage conferred by bilingualism is superfluous. Salvatierra & Roselli (2010) also found a lack of difference between monolingual and bilingual young adults in a simple Simon task (using 2 colors), but a bilingual advantage for older adults. However, there are also instances where bilingual young adults have demonstrated an advantage, namely in tasks that are more difficult. Bialystok (2006), for example, found a bilingual advantage for young adults only in conditions involving more switching and monitoring than in simple conditions.

There also appear to be exceptions to this tendency, however: the Salvatierra & Roselli (2010) study also included a complex Simon task (using 4 colors), where a bilingual advantage was not found for any age group, despite the finding of a bilingual advantage for older adults on the simple Simon task. It should be noted that the bilingual advantage on the Simon task in general has been a controversial finding, and one that has not always been replicated (cf. Colzato et al., 2008; Bialystok, 2006). Despite the lack of cohesion among these findings, there remains the possibility that the young adult age group in the present study is at its best performance in terms of selective attention, leaving little room for a bilingual advantage to emerge. A replication of this study with older adults could evaluate this possibility.

4.9.3.3 Bilingual population may not have been sufficiently balanced bilinguals

One possible difference between this study and previous results finding a bilingual advantage is the degree of bilingualism—more specifically, the degree of balance between the bilinguals' two languages—of the bilingual population examined in the present study. Prior work has found that differences between monolinguals and bilinguals in a variety of domains are more pronounced when the bilinguals are balanced (cf. Bialystok, 1988). In the present study, however, the bilingual groups tended toward L1-dominance rather than balance. Thus, it may be the case that the bilingual participants in this study were not sufficiently bilingual to register notable differences from monolinguals in terms of selective attention advantages, which would account for the lack of differences between listener language groups. In order to investigate this possibility, I first note the degree of balance present in bilingual groups from previous studies, and then assess the present results in light of differences between the present group of bilinguals and those from other studies.

The findings from Bialystok et al. (2004) were cited in support of the B-SA hypothesis in the present experiment (cf. Section 3.4.3.3). The group of younger (age 30-54) Tamil-English bilingual adults in Bialystok et al. (2004), who exhibited less Simon interference as compared with monolinguals, had quite an even balance between their two languages: Participants in this group reported on average using Tamil 56% of the time and English 44% of the time on a daily basis (p. 292).

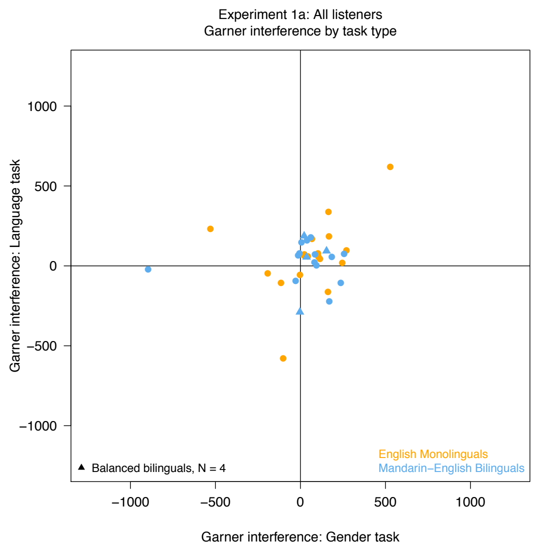
Another study supporting a bilingual advantage in selective attention is Bialystok et al. (2005). In this paper, the authors did not give explicit criteria for inclusion in their young adult bilingual sample, but give the impression that their participants were quite balanced bilinguals: “Bilingualism was determined by a strict set of questions regarding the participants’ language experiences in order to confirm that they had used two languages, essentially daily, since they began to speak… Typically the bilingual participants who met our rigorous standards were children or grandchildren of immigrants who were born into an English-speaking community but had always spoken, and continued to speak, their heritage language at home” (p. 110-111). However, even in such a balanced population, the young adult group did not show differences on the Simon task. This may indicate that the fact that the bilinguals in the present study were relatively L1-dominant may not have entirely accounted for the lack of differences between language groups. Nonetheless, a replication of the present experiments with a sample of bilinguals whose languages were more balanced would be necessary in order to completely rule out this possibility.

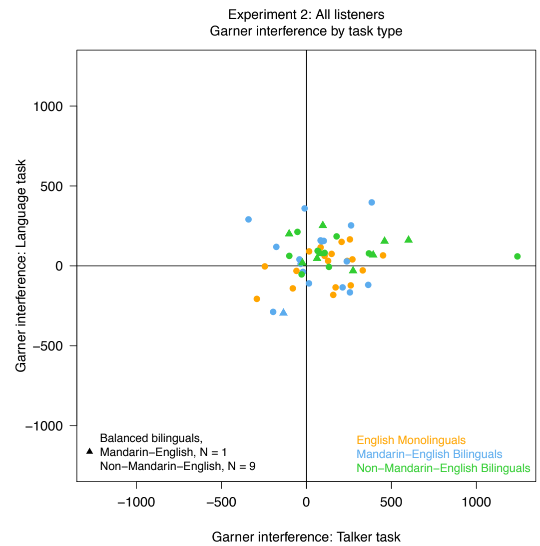
In light of these descriptions, the current data can be cursorily examined with respect to degree of balance of the bilingual participants. Roughly following the bilingualism profile of the younger adult participants in Bialystok et al. (2004), I subset the bilingual groups in these experiments by those who report using one of their languages no more than 60% of the time, and call this subset of the larger group "balanced bilinguals". Out of 18 total bilinguals in each group, in Experiment 1, there were four such Mandarin-English bilinguals; in Experiment 2, one Mandarin-English bilingual and nine non-Mandarin-English bilinguals were considered balanced by these criteria; and eight Mandarin-English bilinguals were considered balanced in Experiment 3. If the bilingual groups in this study were not balanced enough to show a selective attention advantage, then the subset of participants in this study who *are* balanced bilinguals should tend to demonstrate less interference than those who are less balanced. Figures 4.6-4.8, below, are reproductions of Figures 3.7, 3.11, and 3.15 in Chapter 3, which plot the amount of Garner interference for individual subjects on each task. In these new figures, participants falling into the balanced bilingual group are plotted as triangles, whereas the other participants are plotted as circles as before. Participants who show less interference on the language-being-spoken task should fall toward or to the left of the $x = 0$ line, and participants who show less interference on the other task (gender, talker, or amplitude) should fall toward or below the $y = 0$ line. From these figures, it is evident that, with few exceptions, the balanced bilinguals do not fall outside of the interference distribution for bilinguals overall in these experiments; they do not appear to show less interference overall than their more unbalanced bilingual counterparts. Thus, this analysis putatively suggests that restricting the bilinguals tested in these experiments to only balanced bilinguals would not have led to a difference between listener language backgrounds in terms of interference.

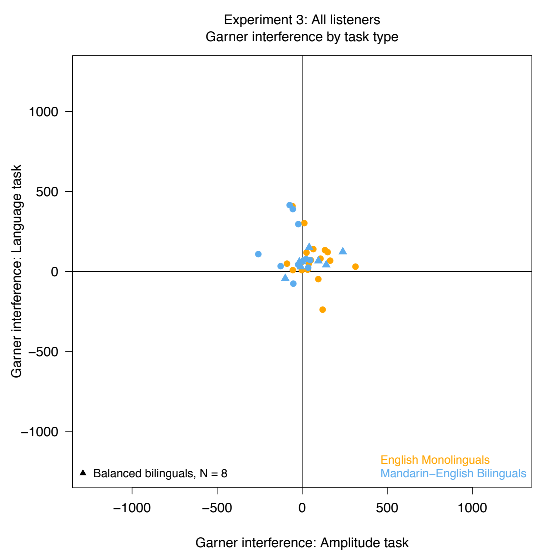
Figure 4.6. Garner interference for each task of Experiment 1a, by participant, comparing balanced and less balanced bilinguals (top left).

Figure 4.7. Garner interference for each task of Experiment 2, by participant, comparing balanced and less balanced bilinguals (top right).

Figure 4.8. Garner interference for each task of Experiment 3, by participant, comparing balanced and less balanced bilinguals (below). Triangles represent those bilingual participants who are more balanced (i.e. they do not use one of their languages more than 60% of the time).







Finally, a different approach could be taken in future work which would at least partially alleviate the need for such carefully controlled, binary groups of participants: The language experience of participants with knowledge of multiple languages could be quantified using a continuous measure. For example, degree of balance between languages could be assessed using a metric such as the Bilingual Dominance Scale (Dunn & Fox Tree, 2009). Then, in order to see whether Garner interference is, in fact, affected by language experience, future studies could test whether such a bilingualism metric correlates with the quantity of Garner interference exhibited. This is a promising possibility, as Bregman & Creel (2014) found that participants' position on the Bilingual Dominance Scale (along with several other measures of bilingualism such as age of acquisition) correlated with how quickly they were able to learn to classify talkers speaking in the listeners' L2.

4.9.3.4 The Garner task as an appropriate measure of selective attention?

A fundamental concern that should be addressed in relation to this hypothesis is the possibility that the Garner task is not appropriately suited to measuring selective attention abilities across language backgrounds. The Garner paradigm is not often used for this purpose; such studies tend to favor the Simon, Stroop, and flanker tasks. Furthermore, it is not known how the Garner task compares with these other tasks as a diagnostic of executive control abilities across language backgrounds. It appears that the only such study assessing bilinguals and monolinguals on the Garner task alongside another paradigm is an unpublished Master's thesis using both the Simon and Garner tasks, which finds no Garner interference in any group (Fiszer, 2008). Thus, successful direct comparisons between the Garner task and these other tasks in the context of bilingual research are scarce, making it difficult to define the difference between the tasks in the service of this purpose. Nonetheless, there have been speculations in the literature about how the Garner task might differ from these others. For example, Jerger et al. (1999, p. 47) write: “Whereas Garner interference derives from irrelevant stimulus variability over a series of trials, Simon interference derives from an initial tendency to respond toward a source in space.”[25] And from Fiszer (2009, p. 27): “The Garner task specifically tests for attentional control rather than the ability to inhibit a response [as in the Simon task].” A closer examination of what exactly makes up attentional control will help to determine exactly how Garner interference is related to these other measures.

Bialystok (1991) proposed that there are two components of performance that need to be coordinated for successful language use, *analysis of representations* and *control of attention*. She argued that bilinguals and monolinguals differ with respect to control of attention (but do not differ with respect to analysis of representations, which is not directly relevant here). Control of attention is defined as “the process by which attention is selectively directed to specific aspects of a representation, particularly in misleading situations... This selective attention is more difficult if a habitual or salient response contradicts the optimal one and must be overruled, making inhibition an essential component of control” (Bialystok & Martin, 2004, p. 325). This type of process seems to be precisely what the Garner task tests, and thus supports the basis of the bilingualism enhances selective attention hypothesis.

Further, within the realm of control of attention, there are two separable processes: *response inhibition* and *interference suppression* (cf. Bunge, Dudukovic, Thomason, Vaidya, & Gabrieli, 2002). Tasks that involve response inhibition are those which require a univalent stimulus to be categorized in a way that is incongruent with experience. For example, a version

---

[25] Though here it should be noted that contrary to their predictions, Jerger et al. (1999) found that similar patterns of performance held for the Garner task and the Simon task across the lifespan, which was the question of interest.

of the Stroop test where participants must respond “day” when shown pictures of moons and “night” when shown pictures of suns requires response inhibition. And, response inhibition is called on during the second phase of a dimensional card sort task, where the same stimuli once categorized by shape now must be categorized by color. On the other hand, tasks which require interference suppression are those where one dimension of a bivalent stimulus must be ignored in favor of the other dimension. Completing the Simon task involves interference suppression, as one dimension (usually color) must be attended to while ignoring another (position). The Garner task also appears to implicate interference suppression. It has been proposed that bilinguals may only show an advantage for interference suppression, but not for response inhibition, though under certain task demands bilinguals also show a response inhibition advantage (Martin-Rhee & Bialystok, 2008). This is likely because bilinguals have experience selecting between two competing representations (each of their two languages), which is similar to interference suppression, while inhibiting one response of a univalent stimulus as in response inhibition is not relevant to the bilingual experience. So far, this is a confirmation that the Garner task is the type of task in which bilinguals may be expected to have an advantage.

One difference between the Garner task and other types of bivalent designs involving interference suppression, however, was mentioned in Section 4.9.2.2, above. Interference in the Garner task comes from variability on the block-level, thus interference only occurs between trials (Boenke et al., 2009). A task like the Stroop or Simon test, however, contains trial-level distractors where interference is present within a trial. Whether the purported bilingual advantage extends to such a situation is unclear. However, the results of the present study (showing the same amount of interference for bilinguals and monolinguals across tasks), taken together with the results of previous studies comparing monolinguals and bilinguals on the Garner task (e.g., Lee & Nusbaum, 1993; Repp & Lin, 1990; with various amounts of interference for monolinguals versus bilinguals across tasks), might suggest that bilinguals may not show an advantage in this paradigm. That said, it has already been demonstrated that finding a certain dependency relation between dimensions is crucially dependent on the particular dimensions paired (cf. Section 4.6.3). Thus, in order to truly assess the viability of the Garner task as a measure of interference suppression in bilinguals and monolinguals, future studies could conduct Garner tasks on bilinguals and monolinguals using some of the same dimensions tested in Stroop and Simon tasks (such as color and shape). However, one would imagine that if bilinguals were going to show less interference on any dimensions, it would be on those dimensions involving a categorization they are uniquely familiar with, such as language-being-spoken. In sum, it is likely that a bilingual advantage for interference suppression could have been observed in the Garner task, though it did not appear in the current studies for reasons presently unknown. Future work with more controlled comparisons between paradigms would resolve such matters.

4.9.3.5 Longer reaction times may have obscured differences in listener language groups based on selective attention

The matter of long reaction times, raised in response to the representation hypotheses, above, also applies to the selective attention hypothesis. It could also be the case that the bilingualism enhances selective attention hypothesis was correct, but that the methodological decision to use sentence-length stimuli in these experiments obscured its effects. The mechanism involved in this hypothesis suggested that participants with superior selective attention

(Mandarin-English bilinguals, and non-Mandarin-English bilinguals in Experiment 2) would demonstrate less interference overall than other listeners (English monolinguals), an effect that would likely show up early on in processing. However, the reaction times in this study may have been long enough so that this advantage was curtailed by the time responses were made. Monolinguals may have had sufficient time to catch up to bilinguals given an overall bias toward longer response times. In other words, the overall longer reaction times caused by longer stimuli in this study were not sensitive enough to reveal any disadvantage in selective attention in monolinguals. The follow-up experiment described in Section 4.9.2.1, above, using words rather than sentences as stimuli, could also serve to determine whether long reaction times were responsible for the equal performance of bilinguals and monolinguals in the present experiments.

4.9.4 "Canceling out"

Two additional explanations may account for the lack of differences found between language groups. One possibility is that the mechanisms behind the competing hypotheses were additive, and canceled each other out. The two main mechanisms involved in the hypotheses, degree of representations of dimensions, and selective attention ability, were hypothesized to exert opposing influences. Listeners with more detailed representations of dimensions were hypothesized to exhibit more interference (Mandarin-English bilinguals in both versions of this hypothesis, with the addition of non-Mandarin-English bilinguals for the bilingualism version), while listeners with better selective attention capacity were hypothesized to exhibit less interference (both Mandarin-English bilingual and non-Mandarin-English bilingual listeners). One possible account of the lack of differences between listener groups, then, is that both mechanisms behind the hypotheses were at play, and in effect canceled each other out. This is doubtless an oversimplification of the nuanced behavior that each of the ostensible mechanisms is bound to display, but mutual neutralization does represent a possible account of the equal performance of listener groups in these experiments.

4.9.5 Cross-listener group decisional differences in the language-being-spoken task?

Another explanation that may account for the present results has to do with the fact that listener language groups differ in the type of decisions required of them by the language-being-spoken task. When choosing which language is being spoken, Mandarin-English bilingual participants are faced with a choice between two known entities, while the task for participants unfamiliar with Mandarin represents a decision between known (English stimuli) and "other" (Mandarin stimuli). What this difference in decision types represents, and how it may have influenced results, is best explained by analogy with a different type of experimental paradigm, the lexical decision test.

In a lexical decision task, participants must decide whether a string of letters is a word or a non-word in a given language. A robust pattern in lexical decision results is the so-called lexical status effect, which demonstrates that deciding that a "word" stimulus is a word is done more quickly than deciding that a "nonword" stimulus is a nonword (e.g. Rubenstein, Garfield, & Millikan, 1970). While a range of competing models have been proposed to describe the precise mechanisms involved in the lexical decision process, it is clear that the word decision ends with the participant finding the entry in his or her mental lexicon. In contrast, the nonword decision involves a more complicated series of comparisons between the incoming string and the mental lexicon when the entry for the nonword is not found, evidenced by the observation that more word-like (i.e. more phonotactically probable) nonwords are responded to more slowly than less word-like nonwords (e.g., James, 1975; Stone & Van Orden, 1993). Thus, whether there is a representation in memory matching the stimulus in question makes a difference in processing: It takes less time to respond to a known entity than to an unknown one.

Likewise, in the language-being-spoken task of the present experiments, listener groups differ in the types of responses they must make. For listeners familiar with both languages (Mandarin-English bilinguals), all language-being-spoken decisions, whether the stimulus is spoken in Mandarin or English, are similar to a word decision in lexical decision tasks. Both processes involve matching a stimulus with a known representation in memory. For participants unfamiliar with Mandarin (English monolinguals, and non-Mandarin-English bilinguals in Experiment 2), however, responding to a stimulus in English is similar to a word decision, but responding to a stimulus in Mandarin may more closely resemble responding to a nonword decision. For Mandarin decisions, there is no known entity represented in memory for those listeners unfamiliar with Mandarin, just as with nonwords. For these listeners, then, it is possible that their responses to English stimuli were faster than their responses to Mandarin stimuli. Consequently, when compared with the Mandarin-English bilinguals their reaction times overall may be slower because of the unknown language stimuli. Or, there may be an interaction between listener language and stimulus language such that there is a difference in response times for stimuli in different languages only for the listeners unfamiliar with Mandarin. (It is important to note the fact that Mandarin stimuli were on average longer than English stimuli in these experiments, which could serve to exaggerate any tendency for certain listeners to react more slowly to these stimuli.) This effect might interfere with the mechanisms behind the hypotheses in a number of ways. For example, if participants unfamiliar with Mandarin are being slowed down by processing an unknown entity, they may not be able to devote as much processing resources to building up a representation for a dimension. In order to check for this possibility, a comparison of overall response times between listener groups was conducted on each experiment separately, with particular attention to the language of the stimuli.

For Experiment 1, by-participant means of log-transformed reaction times were submitted to a 2 x 2 repeated measures ANOVA, across all blocks of the language task, with the between-subjects factor listener language (English monolinguals and Mandarin-English bilinguals) and the within-subjects factor stimulus language (English and Mandarin). There was no main effect of listener language, $F(1,34) = 2.364$, $p = 0.133$, indicating that reaction times did not differ significantly across listener groups overall. There was no main effect of stimulus language, $F(1,34) = 1.007$, $p = 0.323$, indicating that reaction times did not differ significantly depending on which stimulus language was heard. However, these effects were mediated by a significant interaction between listener language and stimulus language, $F(1,34) = 8.112$, $p = 0.007$. Follow up simple effects testing revealed a significant effect of stimulus language for English monolinguals but not for Mandarin-English bilinguals. English monolinguals showed faster RTs overall to English stimuli ($M = 1602.57$ ms) versus Mandarin stimuli ($M = 1678.03$ ms), t(70[26]) = 3.138, $p = 0.003$, as expected. For Mandarin-English bilinguals there was no significant difference between the two stimulus languages ($t(71) = -1.919$, $p = 0.059$) though the

[26] The degrees of freedom differ between language backgrounds because of the missing data in the correlated block from one English monolingual subject.

numerical trend was toward faster responses to Mandarin stimuli (English *M* = 1451.06 ms, Mandarin *M* = 1432.06 ms). In sum, this analysis confirms the proposal that English listeners respond more quickly to a stimulus in a familiar language (English) than to one in an unfamiliar language (Mandarin).

The picture is quite different for Experiment 2 (talker–language-being-spoken), however. By-participant means of log-transformed reaction times were submitted to a 3 x 2 ANOVA, across all blocks of the language task, with the between-subjects factor listener language (English monolinguals, Mandarin-English bilinguals, and non-Mandarin-English bilinguals) and the within-subjects factor stimulus language (English and Mandarin). There was no main effect of listener language, F(2,51) = 2.152, p = 0.127, indicating that reaction times did not differ significantly across listener groups overall. There was no main effect of stimulus language, F(1,51) = 1.303, p = 0.259, indicating that reaction times did not differ significantly depending on which stimulus language was heard. However, these effects were mediated by a significant interaction between listener language and stimulus language, F(2,51) = 5.289, p = 0.008. Follow up simple effects testing revealed a significant effect of stimulus language for Mandarin-English bilinguals and non-Mandarin-English bilinguals, but not for English monolinguals. Non-Mandarin-English bilinguals showed faster RTs overall to English stimuli (*M* = 1410.14 ms) versus Mandarin stimuli (*M* = 1446.12 ms), t(71) = 2.85, p = 0.006, as expected. English monolinguals did not follow this same pattern, as there was no significant difference between the two stimulus languages (t(71) = -1.103, p = 0.273) and even trended in the opposite direction (English *M* = 1510.19 ms, Mandarin *M* = 1498.67 ms). However, Mandarin-English bilinguals showed faster RTs overall to Mandarin stimuli (*M* = 1611.14 ms) versus English stimuli (*M* =

1665.41 ms), $t(71) = -3.437$, $p = 0.001$, counter to expectations. This puzzling result for the Mandarin-English bilinguals may indicate that the comparison with lexical decision results may not be as straightforward as originally proposed.

The results for Experiment 3 (amplitude–language-being-spoken) are different still. The ANOVA was conducted as for Experiments 1 and 2. There was no main effect of listener language, $F(1,34) = 0.112$, $p = 0.74$. However, there was a main effect of stimulus language, $F(1,34) = 4.881$, $p = 0.034$, where stimuli in Mandarin (M = 1396.42 ms) were responded to more quickly than stimuli in English (M = 1414.19 ms), $F(1,34) = 4.881$, $p = 0.034$. There was no significant interaction between listener language and stimulus language, $F(1,34) = 0.421$, $p = 0.521$. This result further complicates the matter.

Given the expected results of Experiment 1 and the non-Mandarin-English bilinguals in Experiment 2, it is worth investigating whether these variable responses concerning language of the stimulus may have affected Garner interference patterns across listener languages. Such an effect is unlikely, however, because an equal number of stimuli are presented in each language in both control and orthogonal blocks, so faster reaction times to one stimulus language should exert the same effect on both control and orthogonal blocks. Nonetheless, this possibility is investigated in results from the language task of Experiments 1 and 2, where there were effects of stimulus language on particular listener groups.

In Experiment 1, by-participant means of log-transformed reaction times were submitted to a 2 x 2 x 2 repeated measures ANOVA, with the between-subjects factor listener language (English monolinguals and Mandarin-English bilinguals) and the within-subjects factors stimulus language (English and Mandarin) and block (orthogonal and control only, as interference is the measure of interest). Only comparisons found to be significant will be reported in this analysis, unless otherwise noted. As before, the interaction between stimulus language and listener language was significant, $F(1,34) = 6.826$, $p = 0.013$. However, the interaction between stimulus language and block was not significant, $F(1,34) = 0.021$, $p = 0.885$, indicating that any effects of stimulus language on reaction times did not play a role in interference between blocks. Further, the three-way interaction between listener language, stimulus language, and block was not significant, $F(1,34) = 2.863$, $p = 0.10$, indicating that any relationship between listener language background and stimulus language did not have an effect on levels of interference.

The exact same pattern was found for Experiment 2 (significant interaction between stimulus language and listener language, $F(2,51) = 4.310$, $p = 0.020$; no significant interaction between stimulus language and block, $F(1,51) = 0.046$, $p = 0.832$; and no significant three-way interaction between listener language, stimulus language, and block, $F(2,51) = 1.324$, $p = 0.275$). Although the results comparing reaction times of different listener language groups to stimuli presented in different languages are inconsistent and intriguing, they do not affect the main measure of interest, Garner interference.

4.10 Conclusion

These experiments have demonstrated that language-being-spoken is integrated with gender, talker, and amplitude. Language-being-spoken has a mutual and symmetrical processing dependency relation with gender, a mutual and asymmetrical dependency relation with talker, and an asymmetrical dependency relation with amplitude. These results form a hierarchy between dimensions, depicted in Figure 4.3, which will be called upon in making predictions about the free classification experiment in the next chapter. The processing dependencies between language-being-spoken and these other dimensions were predicted by the task-based levels of processing hypothesis for some but not all experiments. Future work is needed in this area. The specific relationship between language-being-spoken and talker was predicted by the dimension-specific relative language-specific/talker-general hypothesis, and is an important contribution to the existing literature, but this hypothesis did not extend to the gender–language-being-spoken pairing. Another contribution of this work comes from reporting on the processing time of language-being-spoken at baseline as compared to the other dimensions.

While there were no differences in interference found between listener language groups in any of these experiments, this chapter points out important theoretical and methodological considerations that may have led this to be the case. However, as it stands, it appears that the processing hierarchy described by these experiments is present for all listeners tested in this dissertation, regardless of language background. Though these populations differed in their relationships to the target languages, and in their status as monolinguals or bilinguals, they behave in these tasks as though they belong to one population. The stability of such patterns across listener groups is an important finding.

CHAPTER 5: EXPERIMENT 4, MULTI-PHASE BINNED CLASSIFICATION

5.1 Chapter outline

This chapter presents a supplementary experiment designed to further examine the relationships between the dimensions under investigation, language-being-spoken, talker, gender, and amplitude. The chapter begins by giving a rationale for this experiment, before moving onto various predictions for the results. I then discuss the methodology of the current paradigm, pointing out the ways in which this instantiation of the task is different than the traditional auditory free classification task. After describing the experiment's participants, materials, and procedure, I present the results of the experiment using a variety of analysis methods. The chapter closes with a summary and conclusion.

5.2 Overview of Experiment 4

5.2.1 Rationale

Experiment 4 uses a different methodology, a modified version of the auditory free classification task (Clopper, 2008), in order to examine the relationship between multiple dimensions of interest at the same time, and the nature of that relationship when participants do not have time pressure to complete the task. While the speeded classification task used in Experiments 1-3 tested the relative interference of one dimension when attending to the other, the paradigm used in Experiment 4 tests what dimensions of a stimulus are salient to participants instructed to sort a set of stimuli into groups.

The free classification task asks participants to put stimuli into groups without giving explicit instruction as to the nature of the groups, so participants are effectively choosing the stimulus dimensions along which the stimuli should be classified. There are several advantages of this methodology for the present purposes. One advantage is that it allows the relationship between multiple dimensions to be examined simultaneously, while the Garner paradigm only allowed for testing two dimensions at one time. Another advantage of coupling this paradigm with the speeded classification task is that a different facet of the relationship between language-being-spoken and other dimensions can be examined: The free classification task assesses participants' conscious, explicit decisions about these dimensions, while the Garner paradigm measures more automatic reactions.

Both monolingual English and bilingual Mandarin-English participants were tested in Experiment 4. While there were no differences between language backgrounds in the amount of interference between dimensions in Experiments 1-3, any differences observed in free classification decisions between the two listener groups, then, may underscore differences between the processes involved in performing the two tasks.

The present study uses a modified version of the traditional auditory free classification paradigm, the details of which will be described in Section 5.3 below.

5.2.2 Repurposing the free classification task

The auditory free classification methodology has been used in speech research to examine the perceptual similarity between stimuli varying along a number of types of dimensions, both linguistic and indexical. It is often employed by researchers as a bottom-up method for determining similarity between stimuli when the relevant dimensions within those stimuli are unknown. For example, the paradigm has been used to examine how participants classify talkers by regional dialects (Clopper and Pisoni, 2007), foreign languages (Bradlow, Clopper, Smiljanic & Walter, 2010), and foreign accents (Atagi & Bent, 2011, 2013). Results from free classification tasks are typically submitted to clustering and multidimensional scaling analyses in order to produce hierarchies of similarity between emergent groups, which reveal the dimensions listeners use to classify stimuli.

The typical design of the task uses an audiovisual interface with audio stimuli presented as discrete, movable icons. Participants are asked to move the stimulus icons onto an unconstrained two-dimensional plane, putting them into as many groups as they wish based on any criteria they wish to use. Thus, the task does not impose experimenter-designated categories onto the grouping of stimuli by listeners. Instead, experimenters induce dimensions that are salient to participants via clustering or multidimensional scaling algorithms operating over participants' groupings. This dissertation, however, uses this task for a slightly different purpose, which is to explore the relationship between a pre-determined set of dimensions (language-being-spoken, talker, gender, and amplitude), rather than to induce a set of dimensions bottom-up. That is, the dimensions of interest are already known, and the task is used to determine how many participants use those dimensions and in which order. Therefore, a more top-down approach is taken, which analyzes how participants group stimuli into an analyst-determined number of bins over a series of task phases, within this constrained set of dimensions. While this specific instantiation of the paradigm is designed to look at a circumscribed set of dimensions, this information is not communicated to participants, whose task is still an open-ended one. As the results of this experiment will attest, participants used a wide variety of dimensions in their groupings, beyond those intended by the experimenter. However, because of the interest in only the predetermined set of dimensions, multidimensional scaling is not an appropriate form of analysis for this purpose, and different approaches to evaluating results are performed instead.

The multi-phase binned classification task, then, retains many of the benefits of an entirely “free” classification task, while somewhat simplifying the analysis process.

5.2.3 Overall predictions

Predictions about the results of the current experiment will be made based on a variety of sources, and are described in the following sections.

5.2.3.1 Predictions based on results of Experiments 1-3

The multi-phase binned classification task was conducted in order to gain more insight into the relationship between the dimensions of interest previously tested using the Garner paradigm. As such, I use those previous results to make predictions about the outcome of the multi-phase binned classification task. Before doing so, it is first important to discuss the ways in which the two tasks differ.

While the processing origins of free classification results have not been extensively studied, it is clear that this task differs in a number of ways from the speeded classification task used in Experiments 1-3 of this dissertation. First, participants performing the free classification task are under no time pressure, a characteristic emphasized in the instructions given to participants. This is counter to the speeded classification task, where participants are encouraged to respond as quickly as possible without sacrificing accuracy. Furthermore, the nature of the two tasks is different. In the speeded classification task, participants are instructed to focus on one dimension at a time. Though it is noted in the instructions that there may be variability in another dimension, the Garner task requires that participants give only one dimension their attention at a time. In the free classification task, however, the multidimensional nature of the stimulus is (implicitly) emphasized. Since the task of the participant is to place stimuli into groups, it is implied that there are multiple dimensions available over which to perform the grouping. Although it will be seen that participants frequently did use only one stimulus dimension per phase in performing the task, they are free to entertain using any number of dimensions in their groupings before settling on the one(s) they choose to use.

Despite the fact that the two tasks are quite different, the term salience is often used in accounting for participants' behavior in both tasks. In the free classification literature, the concept *salience* is often used in discussing dimensions utilized by participants. For example, Clopper (2008, p. 575) explains: "The participants' classifications are submitted to clustering or scaling analyses, and the results are interpreted as a reflection of the most salient perceptual dimension(s) of similarity across the stimulus items." Atagi & Bent (2013, p. 513) describe the results of a multidimensional scaling analysis of free classification decisions in terms of salience: "These results in which gender and degree of foreign accent are the two most salient dimensions replicate the findings in Atagi and Bent (2011)." In the free classification task, then, dimensions used by many participants are often referred to as salient dimensions.

As was discussed in Chapters 2 and 4, salience is also used in describing the origins of dependency relations between dimensions in the speeded classification task. In this literature, a salient dimension is one which captures attention, meaning that it is likely to not suffer from interference when it is the target dimension, and that it likely causes interference when it is the non-target dimension. Thus, dimensional salience will be used as a way to make predictions for the current experiment based on the results of the Garner task experiments.

Despite the differences between these tasks, I invoke the similar notion of salience across tasks to make the prediction that behavior on the binned classification task will mirror performance in the Garner task in terms of which dimensions are salient. Specifically, I will use the dimensional hierarchy schematized in Figure 4.3 of Chapter 4 as a basis to make predictions about how participants will perform on the two versions of the binned classification task. First, for Experiment 4a, the gender−language-being-spoken−amplitude task, I draw on the previous results that gender was found to have a mutual and symmetrical relationship with language-being-spoken, while amplitude was found to interfere with language-being-spoken and the reverse case was not true. Thus, out of the pairwise comparisons conducted between these three dimensions (note that one possible pairing, gender−amplitude, was not performed), amplitude was found to be the most salient, and is therefore predicted to be used by participants first in binning the stimuli. In the next phases, it is predicted that participants will be equally likely to use gender as they are to use language-being-spoken. For Experiment 4b, the talker−language-being-spoken−amplitude task, I draw on the previous results that language-being-spoken was found to interfere more with talker than vice versa, and, again, that amplitude was found to interfere with language-being-spoken but the reverse was not true. The results of the pairwise comparisons conducted between these three dimensions (again, note that one possible pairing, talker−amplitude, was not performed) set up the prediction that amplitude will be used first, then language-being-spoken, then talker.

Two other predictions fall out from the dimensional hierarchies determined by Experiments 1-3. First is the prediction that, comparing across the two experiments, participants will be more likely to use gender than talker in early phases of the task, at least in relation to language-being-spoken. This is due to the dimensions' respective relationships with language-being-spoken in the Garner task (symmetrical for gender and asymmetrical for talker), such that gender and language-being-spoken are predicted to be equally salient, while talker is likely to be less salient than language-being-spoken. Next is the prediction that language-being-spoken may be used in early phases more frequently in the talker version than in the gender version. This is again because language-being-spoken is expected to be more salient than talker, but as salient as gender.

5.2.3.2 Predictions based on previous free classification results

It may be, however, that the Garner task and the free classification task are different enough that what is considered salient in one task may not be so in the other task. The free classification task is likely to involve a level of awareness that the speeded classification task does not. In the free classification task, participants are free to choose the dimension they feel is most important (for whatever reason), rather than simply the dimension that is hardest for them to ignore. As such, a prediction about which dimensions will be used by participants in this experiment will be made based on previous free classification experiments.

Several free classification experiments have reported that gender is a salient feature used by participants in various classification tasks, despite differences in stimuli and instructions (see review in Atagi & Bent, 2013). For example, Atagi & Bent (2011) presented listeners with 24 speech samples (using a different sentence for each talker) from two male and two female non-native English talkers from each of six different language backgrounds speaking in English. Participants were asked to group the talkers based on general similarity. A multidimensional scaling analysis found that gender was the first dimension used by participants in this task (and that listeners did not use native language background as a dimension in their groupings, though they did use degree of accentedness). Experiment 2 of Clopper & Pisoni (2007) used 48 stimuli, (one stimulus per talker), from four male and four female talkers from six different regions of the United States. Participants were tasked with grouping sound clips by the region of origin of the talkers, but gender was frequently used by participants as a grouping criterion in performing the task, despite being explicitly told to ignore it.

Thus, although amplitude and language-being-spoken have not been manipulated in previous free classification experiments, based on these results it might be expected that gender will be a salient dimension to participants, and may be among those used most frequently in Phase 1.

5.2.3.3 Predictions based on listener language backgrounds

As in previous experiments of this dissertation, this experiment was conducted on two populations of listeners, English monolinguals and Mandarin-English bilinguals. Based on the lack of differences between these two groups in the speeded classification task, one hypothesis based on language background is that the hierarchy of dimensions derived from the binned classification results in this experiment will be the same for these two groups. However, it may also be the case that this task is sufficiently different than the Garner task, and that language background may have an effect on the results of this task. In that case, one possibility is that bilinguals and monolinguals will treat the language-being-spoken dimension differently based on bilinguals' history of experience with the dimension. There are two possible options here. It could be that bilinguals would be more tuned into the language-being-spoken dimension, and thus use it in classification at an earlier phase than do monolinguals. However, it could also be that bilinguals are more used to speech in both languages, and thus overlook the language-being-spoken dimension in favor one of the other dimensions. In terms of monolingual task behavior, they may have a harder time classifying Mandarin stimuli than English stimuli beyond the language-being-spoken level because they are not as familiar with Mandarin speech, and thus it is predicted that monolinguals may use the language-being-spoken dimension first.

5.3 Methods

5.3.1 The multi-phase binned classification task

As was discussed in Section 5.2.2 above, this experiment uses a modified version of the free classification paradigm. It differs from typical free classification experiments in two main ways: stimuli are grouped into bins rather than in a completely "free" manner, and participants complete several phases of bin classification. As such, this methodology will be referred to as a *multi-phase binned classification task*.

The decision to ask participants to group stimuli into a set number of bins is not a new task manipulation. For example, in a study exploring the perceptual dimensions involved in grouping talkers by voice quality across languages, Esposito (2006) had participants bin stimuli into only two groups. As indicated in Section 5.2.2 above, using a set number of discrete bins in this instantiation of the task is motivated by the fact that the task is being used for a slightly different purpose than most free classification studies. The intention of this task was to specifically evaluate the relationship between language-being-spoken, talker, gender, and amplitude. Thus, the analysis of extra information gained from conducting a purely free classification with no bins (e.g. the location in space of groups, the location of stimuli within groups) is not necessary for determining whether and how participants used these specific dimensions. The use of bins, then, is meant to streamline the analysis and allow for the other variation used in this design: the use of multiple phases.

The multi-phase approach taken in this experiment appears to be an innovation to the free classification paradigm. The purpose of this manipulation here is also related to the goal of the task, which is to determine whether and how the aforementioned dimensions are used by participants. Having participants bin the stimuli based on one dimension at a time sets up a sort of salience hierarchy among the dimensions they use. The dimension participants use in the first phase to sort stimuli into the initial two bins is likely the one that was the most salient to them. Then, participants use the dimension that is the next most salient in the second phase, followed by the dimension used in the third phase. Thus, the binned version of this paradigm allows the dimensions to be ranked by salience as measured by the order in which they are used in the task.

Participants completed two versions of the multi-phase binned classification task, each involving a different configuration of dimensions. One version, Experiment 4a, tested the dimensions gender, language-being-spoken, and amplitude. The other version, Experiment 4b, tested the dimensions talker, language-being-spoken, and amplitude. Multiple versions were necessary because gender and talker could not be combined into one task without creating an imbalance in the number of values within each dimension, as talker is nested within gender; in order to differentiate talker from gender, multiple within-gender talkers would need to be used. Having more values of one dimension versus other dimensions could bias participants to pay more or less attention to that dimension as compared to others, or could affect decisions on that dimension more generally (cf. Pisoni, 1992). For example, if two talkers were used per gender, this would necessitate using four talkers as compared with two genders (and two languages and amplitude values). Then, if talker appeared to be most salient to participants, it would be impossible to tell whether this was due to the presence of more talkers or if it reflected the fact that talker was indeed the most salient dimension. Thus, two different versions of the task were run, one including gender with the other two dimensions, and the other including talker with the other two dimensions. The order in which the tasks were completed was counterbalanced across participants. Since they were not included in the same version, no direct comparison between gender and talker can be made, though comparisons will be made indirectly.

5.3.2 Participants

As in Experiments 1-3, participants in this experiment had no history of uncorrected hearing or language impairment and had normal or corrected-to-normal vision. No participant in this experiment also participated in another experiment of this dissertation.

The monolingual and bilingual groups in this experiment were delimited using the same criteria as in Experiments 1-3, described in detail in Section 3.5.2. The monolingual participants in Experiment 4 consisted of 19 English-speaking Northwestern undergraduates. These participants were recruited using the Northwestern Linguistics Department subject pool, and were given partial course credit for their time. The monolingual group was comprised of 7 males and 12 females, with the mean age 20.1 (range 18-23). Participants in this group reported using English 97% of the time on average (range 80%-100%). Eighteen of these participants did not learn a language other than English before age 7. The two participants who did report learning a language before age 7 reported currently using that language on average 0% of the time (range 0%-0%). No participant in this group reported knowing any Mandarin Chinese. Five additional participants were run but their data were excluded unanalyzed because either they did not conform to the language background requirements of the monolingual group (1), or because of a technical glitch (4). See Table 3.7 in Chapter 3 for a comparison between these participants with their counterparts in Experiments 1-3.

The 21 Mandarin-English bilingual subjects who participated in Experiment 4 were recruited both from the greater Northwestern community (who received compensation in cash for their time; less than 1 hour) and from the Northwestern Linguistics Department subject pool (who were given partial course credit for their time). Of these 21 Mandarin-English bilingual participants, 10 were male and 11 were female, with a mean age of 23.4 (range 18-31). On average, the participants in this group reported learning English at age 8.6 (range 3-15), and report using English an average of 42% of the time (range 10%-90%). These participants have lived in an English speaking country for an average of 1.7 years (range 0-6 years). Three additional participants were run but their data were excluded unanalyzed because of a technical glitch (3). See Table 3.8 in Chapter 3 for a comparison between these participants with their counterparts in Experiments 1-3.

5.3.3 Stimuli

5.3.3.1 Stimulus materials, selection, and arrangement

Stimuli used in Experiment 4a and 4b were a subset of the stimuli used in Experiments 1-3. Out of the 64 sentences used from each language in the earlier experiments, 16 sentences of English and 16 sentences of Mandarin Chinese were selected for use in this experiment. The 16 sentences of each language were arranged in the following manner for Experiments 4a and 4b. No item was presented more than once across both versions of the task (gender and talker).

Table 5.1. Stimulus arrangement for Experiment 4a (gender–language-being-spoken–amplitude) and 4b (talker–language-being-spoken–amplitude).

Experiment 4a (gender–language-being-spoken–amplitude)			
Stimulus pair	*Gender*	*Language*	*Amplitude*
1	Male	Mandarin	Loud
2	Male	Mandarin	Soft
3	Male	English	Loud
4	Male	English	Soft
5	Female	Mandarin	Loud
6	Female	Mandarin	Soft
7	Female	English	Loud
8	Female	English	Soft
8 stimulus pairs x 2 sentences for each pair = 16 total stimuli			

Experiment 4b (talker–language-being-spoken–amplitude)			
Stimulus pair	*Talker*	*Language*	*Amplitude*
1	Male 1	Mandarin	Loud
2	Male 1	Mandarin	Soft
3	Male 1	English	Loud
4	Male 1	English	Soft
5	Male 2	Mandarin	Loud
6	Male 2	Mandarin	Soft
7	Male 2	English	Loud
8	Male 2	English	Soft
8 stimulus pairs x 2 sentences for each pair = 16 total stimuli			

5.3.3.2 Stimulus talker characteristics

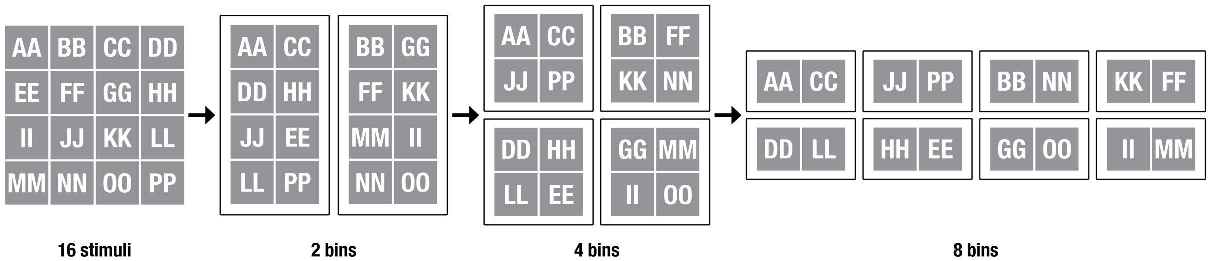
The sentences in Experiment 4a (gender–language-being-spoken–amplitude) were read by a male talker (named Male 1 in Table 5.1) and a female talker (Female in Table 5.1). The male talker was 25 years old at the time of recording and had an overall Versant score of 50 (and a Versant fluency score of 54 and a Versant pronunciation score of 49). The female talker was 22 years old at the time of recording had a Versant score of 65 (and a Versant fluency score of 80 and a Versant pronunciation score of 69). Refer back to Table 3.10 of Chapter 3 to compare characteristics of talkers used across experiments.

The sentences in Experiment 4b (talker−language-being-spoken−amplitude) were read by two male talkers (named Male 1 and Male 2 in Table 5.1). Male talker 1 ("Wei" in Experiment 2), was 25 years old at the time of recording and had an overall Versant score of 50 (and a Versant fluency score of 54 and a Versant pronunciation score of 49). Male talker 2 ("Li" in Experiment 2), was 23 years old at the time of recording had an overall Versant score of 59 (and a Versant fluency score of 77 and a Versant pronunciation score of 61).

5.3.4 Procedure

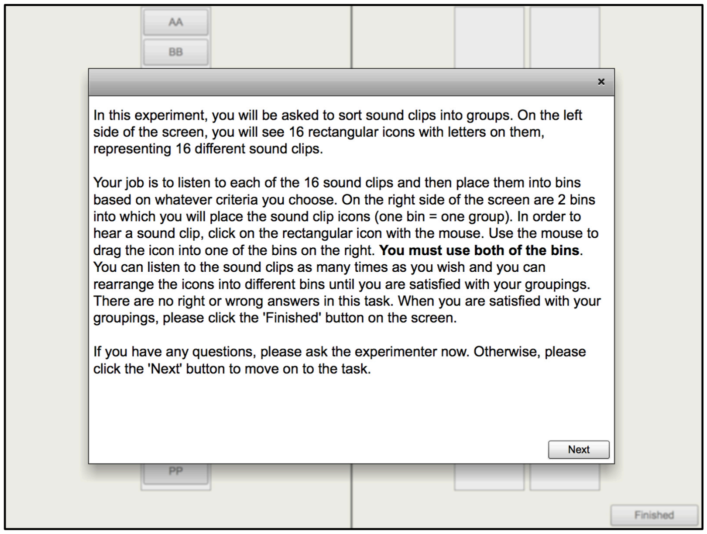
As in Experiments 1-3, participants were seated in a sound attenuated booth equipped with a Mac Mini. Before the start of the experiment, each participant completed a web-based demographic language background questionnaire (Northwestern University Subject Database, or NU-subDb). After the completion of the questionnaire, participants began the experiment, which was programmed in Flash and was presented in a web browser. Each participant completed both versions of the experiment (4a and 4b). The order in which versions were presented was counterbalanced across participants. Each version proceeded in three phases, schematized in Figure 5.1, below.

Figure 5.1. Schematic representing a hypothetical participant's possible set of responses to the three phases of the multi-phase binned classification task. Each participant completed all three phases for both versions of the task, Experiment 4a and Experiment 4b.



Before beginning, participants were not told that there would be two experiments, nor were they told that there were three phases of each version. Instructions given to participants before the first phase of the first version of the experiment are shown in Figure 5.2, below. In Phase 1 of each experiment, sixteen rectangular boxes were displayed on the left side of the screen, and two empty rectangles, or "bins" were displayed on the right side of the screen. Each of the sixteen boxes represented one sound clip, and was arbitrarily labeled with double capital letters. Stimuli were randomly associated with letter codes for each participant. Participants clicked on a box to hear the sound clip, presented via Sony MDR-V700 headphones. Their task was to sort the sound clips into the two bins using whichever criteria they wished, and were told explicitly that there were no correct or incorrect answers, though they did need to use both bins. They were able to listen to the stimuli as many times as they wished, and move the boxes around as much as they wished, in making their classifications. When satisfied with their grouping of items into the two bins, participants clicked a button labeled "Finished", at which point Phase 2 of the task began.

Figure 5.2. Instructions for the first phase of the first set of binned classification tasks.



In Phase 2, participants were given instructions to take the sound clips that they had grouped into two bins (which were then displayed on the left side of the screen exactly as participants had just binned them) and move them into four bins now displayed on the right side of the screen. Regarding this point, the instructions explicitly stated: "You will notice that there are now 4 bins, so you must make a decision about how to make your previous 2 groups into 4 groups." Participants were given the same instructions as before about their ability to listen to sound clips multiple times, and that they were required to use all four of the bins. Once participants were satisfied with their four groupings, Phase 3 began, which displayed their previous groupings of sound clips into four bins on the left side of the screen and asked them to further divide the sound clips into the eight bins on the right side of the screen. Instructions were the same as in previous phases. At the end of Phase 3, if the participant has binned the stimuli using the three expected dimensions (gender, language-being-spoken, and amplitude for Experiment 4a and talker, language-being-spoken, and amplitude for Experiment 4b), then each of the eight bins would contain a pair of stimuli. That pair would consist of the two sentences matching in all three dimensions. In other words, in Experiment 4a the two sentences produced in Chinese by the female talker at the loud amplitude would be paired together in one bin, the two sentences produced in Chinese by the female talker at the soft amplitude would be paired together in another bin, and so on.

After completing all three phases of the first version of the binned classification task, a prompt appeared on the screen, reading: “We’re interested in what your reasoning was in completing the task on the previous page. How did you go about grouping these sound clips into bins? Please tell us in one or two sentences.” There was a dialog box below this prompt where participants typed their rationales. Once they had responded, they began the second version of the task. Participants were given a new set of sound clips and bins, and were given the same instructions as in the first version of the task.

5.4 Results

5.4.1 Coding

In order to analyze the results of this experiment, participants’ classifications of stimuli into bins were first examined to determine which dimension was used to perform each classification. This is a straightforward procedure when, for example, in the first phase a participant chose to group all stimuli spoken in Chinese into one bin and all stimuli spoken in English in the second bin: in that case it is obvious that the dimension language-being-spoken was used to perform the grouping. However, quite frequently, participants chose to group stimuli based on dimensions not anticipated in the design of this experiment. For example, one English monolingual participant in Experiment 1b (talker–language-being-spoken–amplitude) grouped the stimuli as displayed in Table 5.2, below.

Table 5.2. Sample English monolingual participant's performance in Experiment 4b.

<table>
<tr><th>Phase 1</th><th>Dimension used</th><th>Phase 2</th><th>Dimension used</th><th>Phase 3</th><th>Dimension used</th></tr>
<tr><td rowspan="4">1 Mandarin-Male2-Loud
Mandarin-Male1-Loud
Mandarin-Male1-Soft
Mandarin-Male2-Soft
Mandarin-Male1-Soft
Mandarin-Male1-Loud
Mandarin-Male2-Loud
Mandarin-Male2-Soft</td><td rowspan="8">Language</td><td rowspan="2">1 Mandarin-Male1-Soft
Mandarin-Male1-Loud
Mandarin-Male1-Soft
Mandarin-Male2-Soft</td><td rowspan="8">"Other"</td><td>1 English-Male2-Loud
English-Male1-Soft</td><td rowspan="8">"Other"</td></tr>
<tr><td>2 English-Male1-Loud
English-Male2-Soft</td></tr>
<tr><td rowspan="2">2 English-Male1-Loud
English-Male2-Soft
English-Male2-Soft
English-Male2-Loud
English-Male1-Soft</td><td>3 English-Male2-Soft
English-Male2-Loud
English-Male1-Soft</td></tr>
<tr><td>4 English-Male1-Loud</td></tr>
<tr><td rowspan="4">2 English-Male1-Loud
English-Male2-Loud
English-Male2-Soft
English-Male2-Soft
English-Male2-Loud
English-Male1-Soft
English-Male1-Soft
English-Male1-Loud</td><td rowspan="2">3 Mandarin-Male2-Soft
Mandarin-Male2-Loud
Mandarin-Male1-Loud
Mandarin-Male2-Loud</td><td>5 Mandarin-Male2-Soft
Mandarin-Male1-Soft
Mandarin-Male1-Soft</td></tr>
<tr><td>6 Mandarin-Male1-Loud</td></tr>
<tr><td rowspan="2">4 English-Male2-Loud
English-Male1-Soft
English-Male1-Loud</td><td>7 Mandarin-Male2-Soft
Mandarin-Male2-Loud</td></tr>
<tr><td>8 Mandarin-Male1-Loud
Mandarin-Male2-Loud</td></tr>
</table>

As can be seen in the table, the stimuli grouped together in each bin in Phase 1 of the task share in common the language-being-spoken, and therefore it is assumed that the participant used the dimension language to perform the grouping, and Phase 1 for this participant is coded as "language" (as was the case in Experiments 1-3, the language-being-spoken dimension will be referred to as language in the data analysis for sake of brevity). However, in Phase 2, there is no discernable pattern in classification; both talkers and both amplitude values appear in the same groups. The same holds for Phase 3. Thus, for the purposes of the present analysis, both Phase 2 and Phase 3 are assigned the code "other" as the dimension used. Participants did give an account of their reasoning when grouping the stimuli in a free response section following each version of the experiment. (In this case, the participant gave the following response in explanation of the groupings given above: "I kept the sound clips that were English separate from those that weren't. For the English ones, I grouped by tense. For the other ones, I tried to arrange similar sounds together.") For the purposes of coding, however, these self-reported justifications were not consulted, and any phase binned in a way that did not use one of the three dimensions given was labeled “other.” Self-reported dimensions will be examined separately in Section 5.4.3, below.

The fact that “other” dimensions are used at all confirms that there were, in fact, dimensions besides language, gender, talker, and amplitude that existed in the stimuli for participants to key into. If such dimensions hadn’t been used, participants’ selections of the dimensions of interest would have been predetermined by the experimental setup, rather than selected as truly salient dimensions. Thus, the fact that some listeners did use unanticipated dimensions is an important confirmation that the stimuli and experimental setup were well-designed and did not limit participants to choosing dimensions of interest.

Besides language, gender/talker (depending on experiment), amplitude, and “other”, one other outcome was possible, which was labeled “all”. This code was given in cases where a participant used some non-discernable dimension to bin Phase 1 and/or 2, but in Phase 3 the stimuli were paired as if all three relevant dimensions had been used to perform the grouping.

In coding participants’ performance, no attention was given to the ordering of items within each bin; items within a bin were treated as one monolithic group. Further, no attention was given to the configuration of items across bins; if pairs of stimuli in English were placed in bins near each other on the screen, for example, this was not recorded.

5.4.2 Classification results

The results of each version of the experiment are displayed as trees in Figures 5.3 and 5.4, below, depicting classification decisions by participants throughout the course of the task.

The phases of the experiment progress from left to right on each figure. White numbers on the left side of each box depict the total number of participants choosing to use that particular dimension at that particular phase. On the right side of each box, English monolingual totals are displayed in orange and Mandarin-English bilingual totals are displayed in blue. I will discuss the results for the gender–language–amplitude task (Figure 5.3) in detail as an illustration of how to interpret the trees. In the first phase of this task (2 bins), 23 participants (out of the 40 total participants) grouped the stimuli by language, one participant grouped the stimuli by gender, seven by amplitude, and nine by other dimensions not anticipated in the experimental design. Twelve of these 23 participants were English monolinguals (out of 19 total English monolinguals) and 11 were Mandarin-English bilinguals (out of 21 total Mandarin-English bilinguals). The dimensions used by participants in the second phase (4 bins) are displayed in the next column to the right. For example, of the 23 participants who grouped by language first, 14 grouped by gender next, five grouped by amplitude next, and four grouped by an unanticipated dimension next. The dimensions used by participants in the final phase (8 bins) are displayed in the column on the far right. For example, of the 14 participants who grouped by language in the first phase followed by gender in the second phase, 13 of them grouped by amplitude in the third phase and one grouped by an unanticipated dimension. These results obtained via coding and displayed in the trees will be used as the basis of a variety of analyses in the following sections.

Figure 5.3. *Classification results of all participants (N = 40) in Experiment 4a. The white numbers in each box represent the number of total participants who grouped by that dimension in that phase. Orange values are the number of English monolinguals (out of 19) who chose that dimension, and blue values are the number of Mandarin-English bilinguals (out of 21) who chose that dimension.*

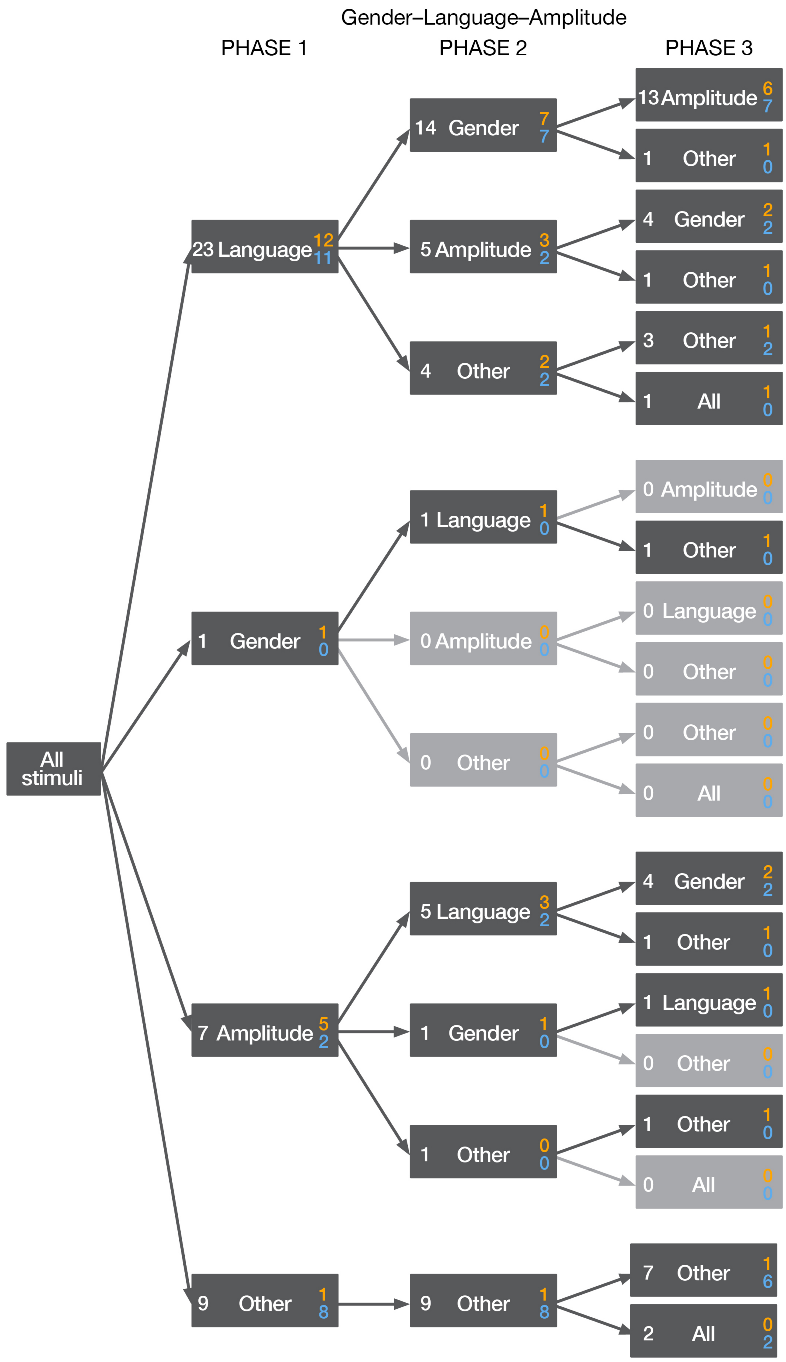
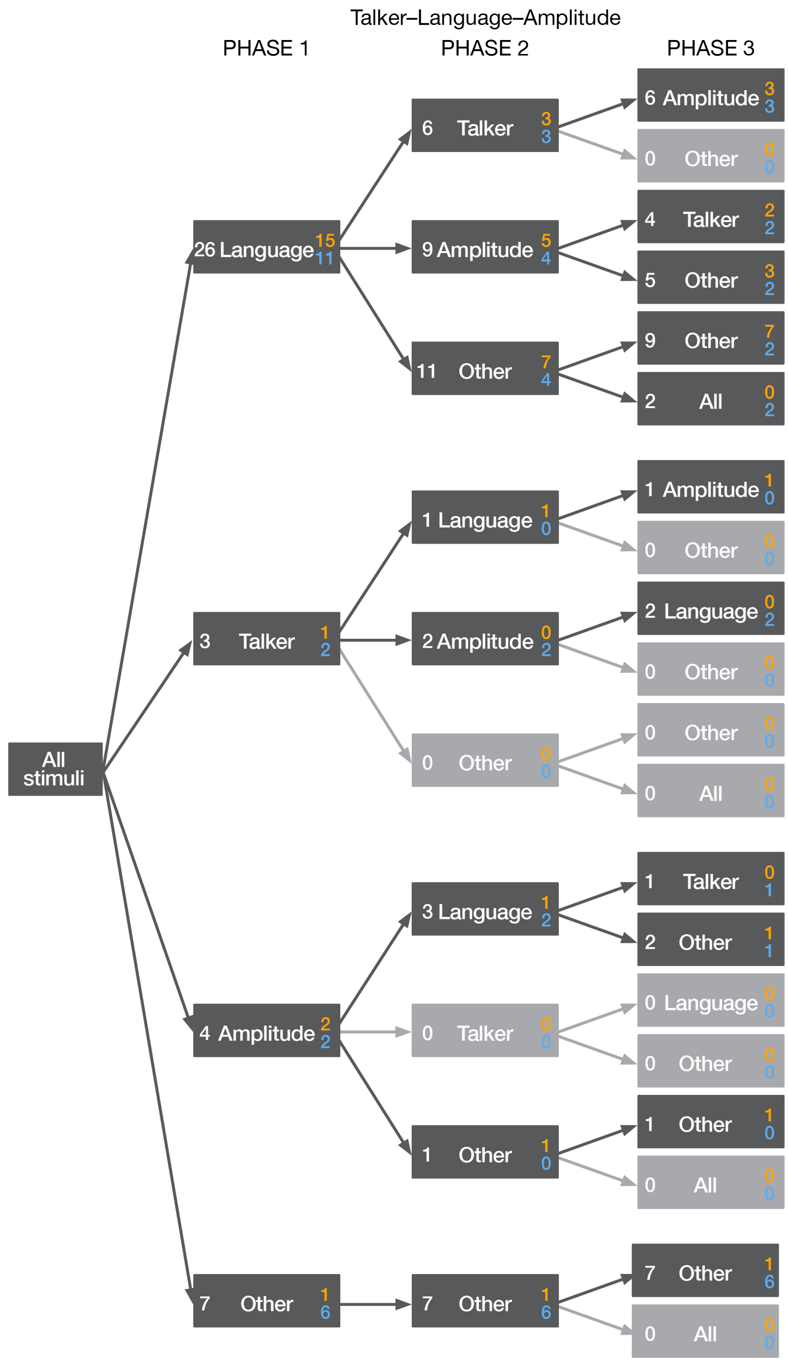


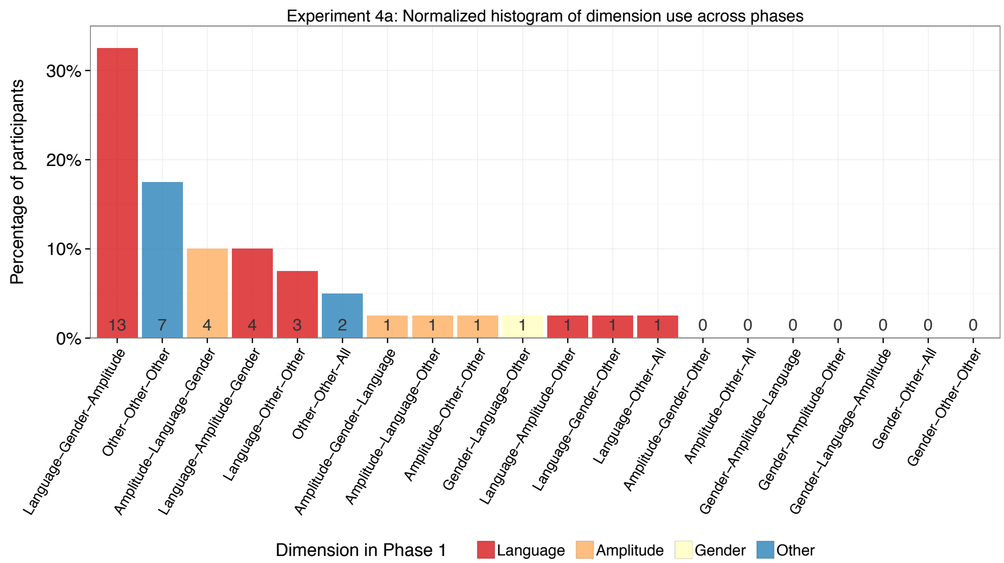
Figure 5.4. Classification results of all participants (N = 40) in Experiment 4b. The white numbers in each box represent the number of total participants who grouped by that dimension in that phase. Orange values are the number of English monolinguals (out of 19) who chose that dimension, and blue values are the number of Mandarin-English bilinguals (out of 21) who chose that dimension.

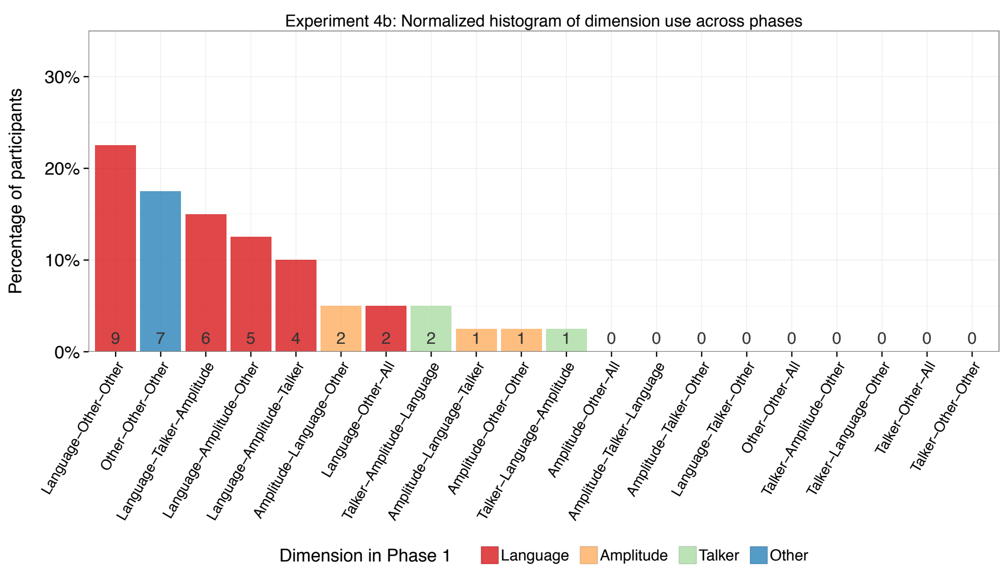


5.4.2.1 Ordering of dimensions

For each version of the task, I now present results illustrating the frequency with which the same progressive dimensional decisions across the three phases were made by participants. In order to do this, I represent participants' behavior as the order in which they used each of the three dimensions for classification. For example, in Experiment 4a (gender–language–amplitude), looking down the Phase 3 column of Figure 5.3 it can be seen that the ordering language-gender-amplitude was used most frequently by participants, employed by 13 of 40 participants (or 32.5%). In other words, 32.5% of participants chose to use gender in Phase 1, language in Phase 2, and amplitude in Phase 3. The next most frequent ordering was other-other-other, which was used by seven participants (17.5%). (Note that this does not mean that the same "other" dimension was used by a participant across phases, only that these participants did not use one of the expected dimensions in any of the three phases. It is also not necessarily the case that the same "other" dimension was used by seven different participants, only that seven participants did not use any of the expected dimensions.) For Experiment 4b (talker–language–amplitude), the language-other-other ordering was used most frequently by participants, employed by nine of 40 participants (or 22.5%). The next most frequent ordering was other-other-other, which was used by seven participants (17.5%). (Again, note the warning about the interpretation of "other" dimensions, above.) Histograms of dimensional orderings across dimensions normalized by the total number of participants are given in Figures 5.5 and 5.6 below for each experiment.

Figures 5.5 and 5.6. Figure 5.5. Normalized histograms showing ordering of dimensions used by participants across phases in the gender version (above).
Figure 5.6. Normalized histograms showing ordering of dimensions used by participants across phases in the talker version (below). Each possible configuration of orderings across phases is given on the x-axis. For example, in Figure 5.5, Language-Gender-Amplitude refers to participants who classified the dimensions by language in Phase 1, gender in Phase 2, and amplitude in Phase 3. Numbers at the bottom of each bar represent the raw frequency of use for that ordering (N = 40). Bars are color coded according to the dimension chosen in Phase 1.





From these histograms it is evident that many orderings of dimensions were used by participants, another indication that multiple classification strategies were employed by listeners. Although there is not much agreement among participants, it is clear that, for both versions of the experiment, language emerged as the most salient dimension, being used first most frequently of all possible dimensions. This is indicated in the figures above by the fact that many of the most frequent orderings are shaded in red, indicating that language was used as the dimension in Phase 1. In the next section I will examine whether language remains the most used dimension over the course of the subsequent phases. It is also notable that in Experiment 4a gender is only used in Phase 1 by one participant, and in Experiment 4b talker is only used in Phase 1 by three participants. This observation will also be examined further in subsequent sections.

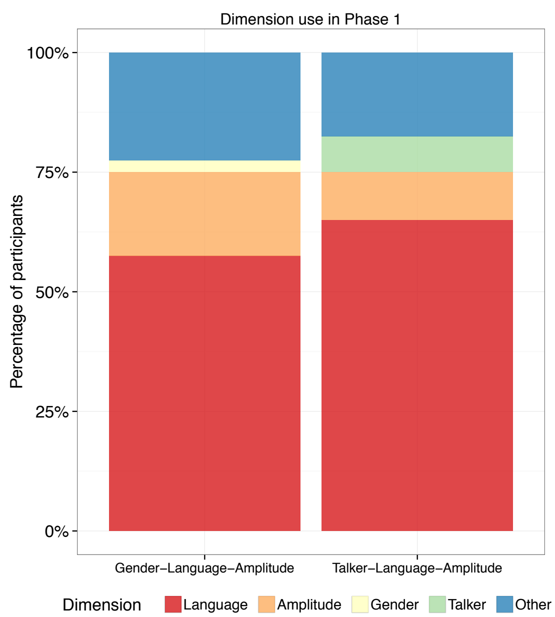
These results also disconfirm several predictions raised in the previous sections. First, the results were not consistent with predictions based on the speeded classification results for both experiments. The results of the Garner experiments suggested that for Experiment 4a, amplitude would be used most frequently in Phase 1, followed by gender or language with equal likelihood. In actuality, the ordering amplitude-language-gender was used by four participants and the ordering amplitude-gender-language was used by one participant. Even when combining both orderings, the dimension use of those five participants was eclipsed by the 13 participants who chose to use language-gender-amplitude and the seven participants who used other-other-other. In Experiment 4b, the Garner results predicted that amplitude would be used most frequently in Phase 1, followed by language in Phase 2 and talker in Phase 3. However, only one participant used that ordering. Thus, it appears that the two experimental paradigms tap into different phases of processing, each with its own hierarchy of dimensions. Further, these results do not support the suggestion by previous free classification studies that gender would be frequently used in early phases of the experiments. Instead, language was the most dominant dimension. However, as previously mentioned, no previous studies of this type that found gender to be a salient dimension also incorporated language-being-spoken or amplitude. Further, none of these studies pitted gender against other dimensions as was done in the design of the present experiment. This suggests that the hypothesis positing a salient role for gender may have been overly specific.

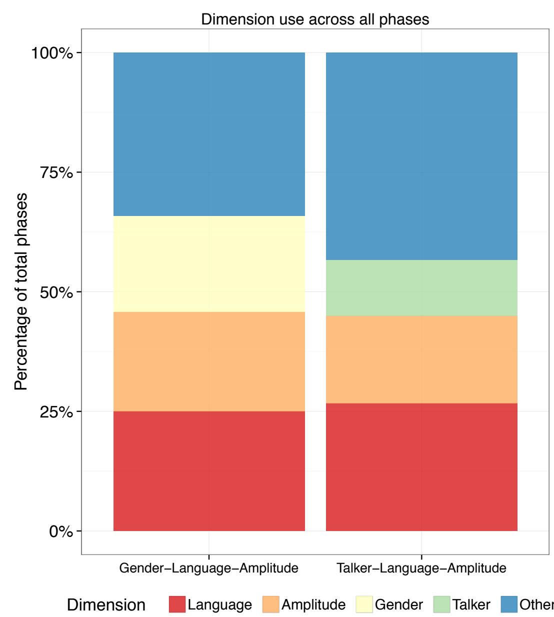
Measuring participants' agreement in dimension orderings would be instructive here, in order to gain a measure of consistency across participants' binning behavior. However, the presence of the "other" dimension prevents such an assessment of inter-participant agreement. Direct comparison of participants' orderings using a metric like Kendall's W is not possible when the dimensions used in orderings vary across participants. However, this sort of analysis will be conducted in Section 5.4.4.1 below on only the subset of participants who use the three expected dimensions for each experiment.

5.4.2.2 Individual dimension use

I now examine dimension use by participants throughout various phases of the experiments. As previously discussed, the dimension participants used to first group stimuli into two bins is an important metric, as that dimension can be thought of as the most salient dimension for that participant. However, the extent to which each dimension was used by participants at any phase is also indicative of its importance to listeners. Since this was an open-ended task, and participants could use any criteria they wished to classify stimuli, the use of a particular dimension at all signifies that it captured participants' attention to at least a certain extent. Thus, Phase 1 dimension use, and dimension use across all phases, are examined in parallel. Figure 5.7 below depicts the percentage of participants who chose to use each dimension in Phase 1 in each experiment, collapsing across language groups. Figure 5.8 below depicts the percentage of total phases in which a particular dimension was used, collapsing across language groups. This is measured across total phases rather than participants because the "other" dimension could be used in multiple phases by the same participant. In both of these figures, decisions marked as "all" in the previous figures are subsumed under the "other" dimension.

Figures 5.7 and 5.8. Figure 5.7. Dimension use by all participants (N= 40) in Phase 1 only (left). Figure 5.8. Dimension use by all participants (N = 40) across all three phases (right). Within each figure, Experiment 4a is on the left, 4b on the right.





First, from Figure 5.7 it is again notable that it is language that dominates Phase 1. This is in contrast to predictions made by the results of the speeded classification task that amplitude would be most salient, again confirming differences between these two experimental paradigms, and is also in contrast to suggestions made by previous free classification tasks that gender might be used frequently.

Next, in comparing Figure 5.7 and Figure 5.8, it is evident that the amount of use of each dimension is more even across all three phases than it is in Phase 1 only. Language is clearly dominant in Phase 1, but other dimensions are used more in later phases. For example, in Experiment 1a, gender is rarely used in Phase 1, but is used nearly as much as language after all three phases are taken into account. Because the dimensions gender and talker are only present in one experiment each, overall performance by participants on all dimensions cannot be compared across the two experiments. Thus, the degree of use of each dimension will be assessed separately for each experiment.

I now examine potential differences in dimension use across each experiment in order to determine whether dimension use varies as a function of which other dimensions are present. I begin with the language dimension. As described in Section 5.2.3.1, the results of previous experiments in this dissertation raise the possibility that language may be used differently in the presence of variability in the gender dimension versus in the presence of variability in the talker dimension. Namely, it was suggested that language may be used in earlier phases more frequently in the talker experiment than in the gender experiment. From Figure 5.7, it appears that language is indeed used by slightly more participants in the talker version than in the gender version. Fisher's exact test[27] can be used to determine whether the frequency of language use is

---

[27] As 20% of observed cells (in the talker–language–amplitude version) contained frequencies less than 5, a chi-square test is not reliable and Fisher's exact test is used instead.

dependent on what type of variability is present (i.e. gender and amplitude variability in Experiment 4a or talker and amplitude variability in Experiment 4b). A two-tailed Fisher's exact test[28] revealed that the two factors were independent for Phase 1 ($p = 0.647$), meaning that the frequency of language use was not related to experiment; language was not used statistically more frequently in Phase 1 of Experiment 4b than it was in Phase 1 of Experiment 4a. From Figure 5.8, usage of the language dimension appears to be similar across experiments when taking all phases into account. A two-tailed Fisher's exact test confirmed this trend ($p = 0.883$), indicating that the degree to which the language dimension was used was not related to which experiment was being performed and thus whether gender or talker variability was present.

Another way of examining the relationships between language and the indexical dimensions gender and talker is by directly comparing the frequency with which each dimension is used within an experiment. This analysis also directly tests a prediction made by the results of the Garner experiments, that gender would be used more than talker in earlier phases. A two-tailed Fisher's exact test revealed that the frequency of gender use in Phase 1 of Experiment 4a and the frequency of talker use in Phase 1 of Experiment 4b were not significantly different ($p = 0.615$), also disconfirming this hypothesis based on results from the Garner task, and again indicating that the two tasks tap into different phases in processing with different relative hierarchies of salience at each phase. Further, a two-tailed Fisher's exact test revealed that the frequency of gender use across all phases of Experiment 4a and the frequency of talker use across all phases of Experiment 4b also were not significantly different ($p = 0.111$).

[28] Statistical tests cannot be performed across experiments on overall dimension use because not all dimensions were present in both experiments; gender was only manipulated in Experiment 4a, and talker only in 4b.

There is a variation of the previous prediction regarding language, gender, and talker that can be tested across all phases, but was not raised in the predictions section above: I now compare the frequency of use of language with the frequency of use of gender in Experiment 4a, and the frequency of use of language with the frequency of use of talker in Experiment 4b, across all phases. It may be that the results of the Garner experiments did not accurately predict ordering of dimension use (as was indeed just demonstrated), but that the dependency relations instead reflect the amount of dimension use overall. Thus, based on the dependency relations observed in Experiments 1 and 2 it can be predicted that, across all phases, language and gender would be used equally frequently in Experiment 4a, but that language would be more frequently used than talker in Experiment 4b. This prediction was not initially considered because the amount that listeners would use unanticipated dimensions was underestimated. If listeners had only used expected dimensions, such a prediction could not be made because all dimensions would be equivalent after all phases had been performed. In Figure 5.8 above, the amount of use of language and gender across all phases (in Experiment 4a) appears to be equivalent, while language is used more than talker across all phases (in Experiment 4b), exactly as was just predicted. These trends are confirmed by two-tailed Fisher's exact tests, which reveal that there was no difference in the proportion of phases binned using gender versus using language in Experiment 4a across all phases, $p = 0.440$. However, there was a statistically significant difference between the proportion of phases binned using talker versus using language in Experiment 4b across all phases, $p = 0.005$. The direction of this difference is evident from the figure; language was used more than talker in Experiment 4b. Again, this tendency mirrors the results found in Experiments 1 and 2: symmetry for gender and language, and asymmetry for talker and language such that there is more interference from language. Thus, even in a task involving a very different process than the speeded classification task, it is again seen that language and gender are matched for listeners, while language and talker are not.

The two experiments can also be compared in the extent to which participants used the amplitude dimension. In Experiment 3, amplitude was found to be highly salient as compared to language-being-spoken, and as such, amplitude was predicted to be used in Phase 1 more than the other dimensions in both Experiment 4a and in Experiment 4b. As was already noted, this was not found. Comparing amplitude use across experiments does not have a precedent from Experiments 1-3, since amplitude was not directly tested against gender or talker. However, the fact that the non-linguistic amplitude dimension is likely to be processed at a different level than the indexical dimensions gender and talker suggests that the degree of amplitude use may not vary depending on the presence of variation in gender versus talker. Comparing the frequency of amplitude use across experiments, two-tailed Fisher's exact tests confirmed that there was, indeed, no association between degree of amplitude use and experiment, both for Phase 1, $p = 0.518$, and across all phases, $p = 0.745$.

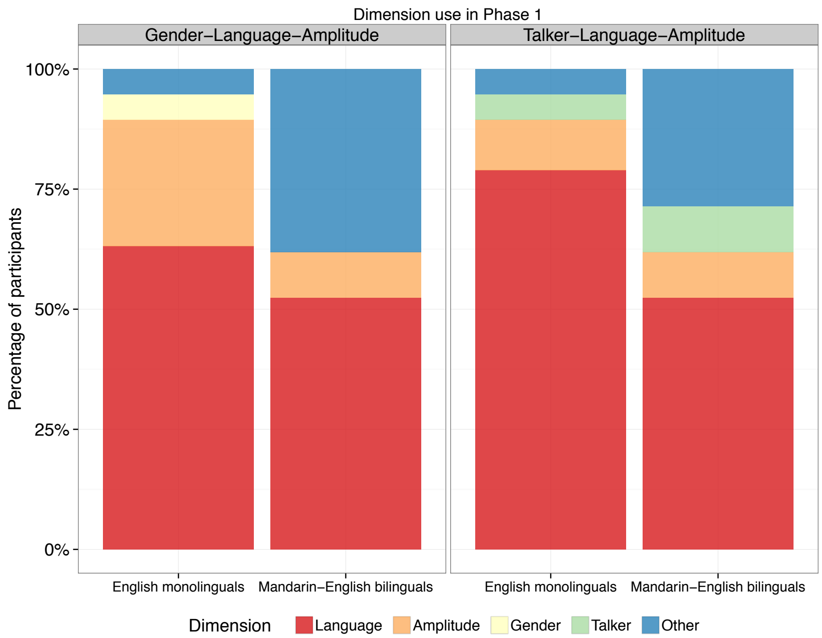
Finally, the extent to which participants used the three expected dimensions versus other unanticipated dimensions can be examined across the two experiments. In other words, the association between the number of participants who used unexpected dimensions (i.e. "other") and the types of dimensions present in each experiment (i.e. gender and amplitude in Experiment 4a versus talker and amplitude in Experiment 4b) can be compared. As with amplitude, there is not a principled reason to believe that the amount of "other" use would differ across experiments. Two-tailed Fisher's exact tests confirmed that these dimensions were also independent, both for

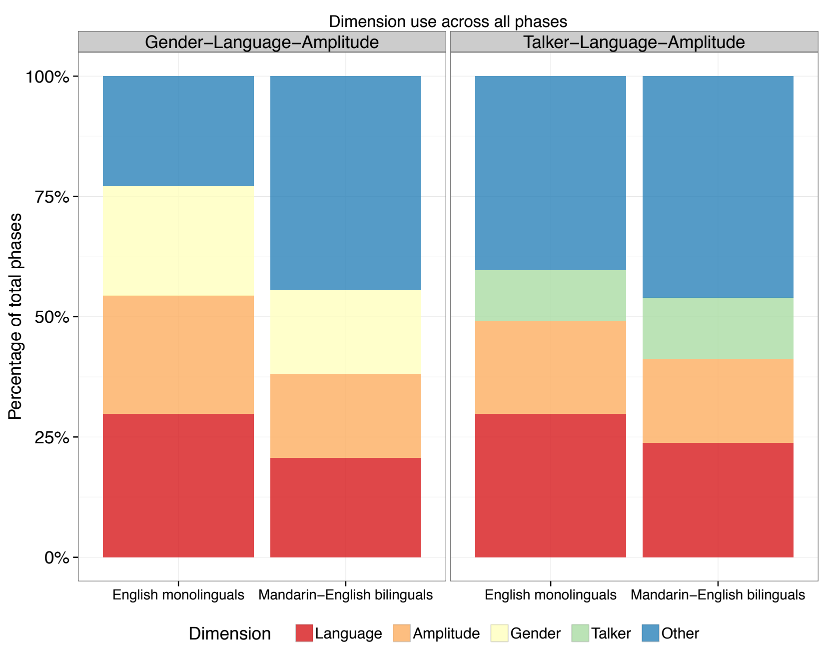
Phase 1, $p = 0.781$, and across all phases, $p = 0.185$. This indicates that the degree to which an unanticipated dimension was used was not related to which experiment was being performed.

5.4.2.3 Language background comparison

Next, I examine the use of dimensions as a function of the language background of participants. The amount that dimensions are used in Phase 1 is presented below in Figure 5.9, with participants separated by language background. Figure 5.10, below, depicts the amount that dimensions are used across all phases of both experiments, with participants separated by language background.

Figures 5.9 and 5.10. Figure 5.9. Dimension use in Phase 1 only, split by language background (above).
Figure 5.10. Dimension use across all three phases, split by language background (below). Within each figure, Experiment 4a is on the left, 4b on the right. English monolingual participants (N = 19) are to the left of Mandarin-English bilingual participants (N = 21) for each experiment.





First, the relationship between frequency of use of all dimensions and listener language background is compared for each experiment. In other words, I examine whether English monolinguals and Mandarin-English bilinguals performed similarly in each experiment. Recall that Section 5.2.3.3, above, detailed several predictions regarding language background. First, the results of the Garner paradigm experiments predicted that English monolinguals and Mandarin-English bilinguals would behave similarly in both experiments of this multi-phase binned classification task. However, based on the difference in familiarity with the language dimension and the fact that the two experimental paradigms tap into different phases of processing, a contrasting prediction was that bilinguals and monolinguals would differ in the frequency with which they use the language dimension in Phase 1.

In Phase 1, two-tailed Fisher's exact tests revealed that dimension use and listener language were related for Experiment 4a (gender–language–amplitude), $p = 0.030$, but not for Experiment 4b (talker–language–amplitude), $p = 0.221$. Across all phases, however, two-tailed Fisher's exact tests revealed that dimension use and listener language were not related for either experiment, $p = 0.101$ for Experiment 4a, and $p = 0.857$ for Experiment 4b. Thus, in Phase 1, there was a difference between the performance of English monolinguals and Mandarin-English bilinguals for the gender version, but not for the talker version. These findings partially confirm the prediction made by the results of the speeded classification task that the language backgrounds would perform equally, but, for Phase 1 of Experiment 4a, run counter to that prediction. In order to determine which dimension(s) were responsible for this difference between language backgrounds (and specifically whether it was the language dimension, as was predicted above), the frequency of use of each of the dimensions in Phase 1 by each language background will be compared for Experiment 4a, below.

It was already noted that language was used first most frequently in both versions of this experiment, and the figures above show that this was true for both English monolinguals and Mandarin-English bilinguals. Two-tailed Fisher's exact tests confirmed this trend, indicating that language use in Phase 1 was independent of participant language background for Experiment 4a, $p = 0.538$, and is therefore not a source of language background differences, thus negating the predictions above that language would be used differently by bilinguals and monolinguals in Phase 1.

The gender and amplitude dimensions might have also been responsible for the overall difference between language backgrounds on Phase 1 decisions in Experiment 4a. Two-tailed Fisher's exact tests revealed that both gender and amplitude use in Phase 1 were independent of participant language background in Experiment 4a, $p = 0.475$ for gender and $p = 0.226$ for amplitude. Neither of these dimensions accounts for the language background differences.

From the figures, it appears that Mandarin-English bilinguals used more unanticipated dimensions (labeled "other") than English monolinguals to perform the binned classification task, which may be the source of the difference between groups. Two-tailed Fisher's exact tests confirmed that the use of unexpected dimensions in Phase 1 was indeed related to participant language background for Experiment 4a, $p = 0.021$. Thus, the difference between listener language groups in their use of the "other" dimension in Phase 1 of the gender–language–amplitude version of the experiment is responsible for the groups' overall difference in performance in Phase 1. English monolinguals used one of the expected dimensions (as compared with "other" dimensions) to group stimuli more frequently than did Mandarin-

English bilinguals. Mandarin-English bilinguals instead relied on more unexpected dimensions this task. Speculation about why this may be is discussed in the next section.

5.4.3 Self-reporting of criteria used for classification

In order to determine which types of dimensions participants used beyond language-being-spoken, gender, talker, and amplitude, I report on their self-reported explanations for their task behavior. First, for reference, I present a few representative explanations given by listeners who only used expected dimensions for all three phases. Self-reported rationales by such participants tended to be fairly straightforward, as in the following examples. One Mandarin-English bilingual participant who used the three expected dimensions gave this rationale for how he/she binned stimuli in the gender–language–amplitude experiment, which is representative of the types of responses from this experiment: “First by language; second, the gender of the speaker; third, the sound volume of the clip.” One English monolingual participant who used the three expected dimensions gave this rationale for how he/she binned stimuli in the talker–language–amplitude experiment, which is representative of the types of responses from this experiment: “I organized the first two bins by english and non-english, then organized those two into loud and soft, and lastly split up those four bins into the two different male voices I heard.” These types of responses were similar across participants using the anticipated dimensions in bin classification.

For those participants who did not rely only on expected dimensions, I now present a summary of how participants reported that they performed the task. In order to do this, I first coded participants’ free responses into groups based on their described strategies. Participants’ self-reports were not always coherent or transparent, but no attempt was made to reconcile actual groups made by participants with their self-reported explanation of how they made the groups. Further, I did not distinguish between dimensions used in different phases, but instead present them here all together. The main reason for this is that in many cases it was not possible to understand from the self-reports which dimension a participant used in which phase. If a participant had been previously coded as using “other” dimensions in more than one phase, the part of the explanation that was most interpretable was given as the “other” code for that participant in the report below.

In the gender–language–amplitude experiment, the following unanticipated dimensions were reported by the eight English monolinguals who used such dimensions: characteristics of the subject of the sentence (e.g. animacy, N = 2), characteristics of the talker (e.g. speaking rate or “accent”, N = 2), a suprasegmental cue (e.g. intonation contour, N = 1), and the tense/aspect of sentence (N = 1). One additional participant in this group reported using a gender-language-amplitude ordering, but a mistake made in their classification of amplitude placed them into this group of “other” dimension users. The final participant in this group used different criteria for English stimuli than they did for Mandarin stimuli (in this case, tense/aspect for English stimuli and gender for Mandarin). For the 10 Mandarin-English bilinguals in the gender–language–amplitude experiment who used unanticipated dimensions, these dimensions were reported: whether the content of the sentence was “a subjective feeling or an objective fact” (N = 2), the tense/aspect of the sentence (N = 1), a characteristic of the subject of the sentence (N = 1), and miscellaneous grammatical criteria (including participants who listed more than three grammatical dimensions, e.g.: “I grouped them by whether they are multipal or singular, happens regularly or just one time, female or male, now or past, judgement or fact, could easily happen or hard.”, N = 4). One participant in this group reported grouping the sentences together so they told a story. The self-report for the final participant in this group did not contain anything that could be interpreted as a dimension used for grouping.

In the talker–language–amplitude experiment, the following unanticipated dimensions were reported by the 13 English monolinguals who used such dimensions: characteristics of the talker (e.g. “whether the speaker was native”, “emotion filled versus bored voices”, N = 2), and general sound similarity (N = 1). The self-report for one participant in this group did not contain anything that could be interpreted as a dimension used for grouping. Another participant in this group divided stimuli first by language, and subsequently according to the arbitrary letter codes on each box (e.g. “HH”). Eight participants in this group used a different classification strategy for the English stimuli than for the Mandarin stimuli. For seven of these eight participants, the Mandarin stimuli were categorized by the stimulus sound in some way, whether it was talker-related, or more generally sound-related (e.g. “I did my best to group sentences together that had similar sound/syllabic patterns.”), while the last participant in this group admitted to grouping Mandarin stimuli randomly. For seven of these eight participants, the English stimuli were grouped by the stimulus meaning in some way, ranging from use of verb tense to characteristics of the subject of the sentence, while the last participant in this group also used a sound-related dimension to group English stimuli (“accent”). For the 13 Mandarin-English bilinguals in the talker–language–amplitude experiment who used unanticipated dimensions, these dimensions were reported: characteristics of the subject of the sentence (N = 4), the tense/aspect of sentence (N = 3), characteristics of the talker (N = 1), a suprasegmental cue (N = 1), and miscellaneous grammatical criteria (including participants who listed more than three grammatical dimensions, N = 2). One additional participant in this group reported using a language-amplitude-talker ordering, but a mistake made in their classification of talker placed them into this group of "other" dimension users. The final participant in this group used different criteria for English stimuli than they did for Mandarin stimuli (in this case, two different grammatical criteria for English stimuli and one grammatical criterion for Mandarin).

From examining these self-reported rationales for classification, it can be seen that part of the reason that Mandarin-English bilinguals tended to use more unanticipated dimensions in grouping stimuli might be that Mandarin-English bilinguals had access to more of these dimensions across all stimuli because they can understand the content of both languages. Given that many of the groupings by "other" criteria were actually based on linguistic dimensions, Mandarin-English bilinguals were the only group who could have used such dimensions to group all stimuli. Linguistic dimensions were clearly salient to many English monolingual participants, as well, particularly in the talker–language–amplitude task. Most of the English monolinguals in this task utilized meaning-related criteria to group the stimuli they could understand (e.g. tense or subject), and used sound-related criteria for the stimuli they could not (e.g. indexical or sound). However, beyond these participants who applied different criteria for each language, many participants appeared to adopt the task strategy whereby the same criteria were applied to all stimuli at once for each dimension. This ruled out the use of linguistic dimensions for the English monolingual participants adopting that strategy. Thus, the ability to access the linguistic content of all stimuli may have resulted in Mandarin-English bilinguals using linguistic dimensions more of the time and indexical dimensions less. Further, as revealed in the previous section, there is a difference in use of "other" dimensions between language backgrounds for the gender version but not for the talker version; Mandarin-English bilinguals differed from English monolinguals in "other" dimension use more in Experiment 4a than in Experiment 4b. The reason why this should be the case is unclear. However, it should be noted that more Mandarin-English bilinguals used unanticipated dimensions than English monolinguals in both experiments, but the difference between groups only reached significance for the gender version.

It is possible that the propensity to use linguistic dimensions in this experiment could have been avoided if participants had been given an additional task instruction, as in Atagi & Bent (2013, p. 512): "Listeners were further instructed to pay no attention to the meaning of the sentences in making these groups." This bit of guidance may have been enough to steer participants in the present study away from relying on dimensions like grammar and animacy. However, the decision was made not to include such instructions in the event that participants might interpret the language dimension as having to do with the content of the stimuli. Since language-being-spoken is a crucial focus of these experiments, even potentially dissuading participants from engaging with it would have been undesirable. Three pilot participants were run without explicit instructions to ignore the content of the sentences in order to see whether they might rely on extraneous dimensions in the absence of such instructions. None of the three initial participants deviated from using the expected dimensions, so the remaining participants were run in the same way. However, as has been demonstrated, a number of later participants did use other criteria to perform the task.

5.4.4 Participants using only dimensions of interest

After having examined the participants who used unanticipated dimensions to sort stimuli, I now examine those participants who only used the dimensions of interest in the experiments. In this analysis, I isolate the subset of participants in Experiment 4a who only used gender, language, and amplitude to group stimuli, and the participants in Experiment 4b who only used talker, language, and amplitude to group stimuli. The number of participants across experiments and language groups that used only expected dimensions to bin stimuli are given in Table 5.3, below.

Table 5.3. Number of participants who used dimensions as expected in each experiment, divided by language background.

Listener language background	*Experiment 4a: Gender–Language–Amplitude*		*Experiment 4b: Talker–Language–Amplitude*	
	Number of participants who only used anticipated dimensions / N	*Percentage*	*Number of participants who only used anticipated dimensions / N*	*Percentage*
English monolinguals	11 / 19	58%	6 / 19	32%
Mandarin-English bilinguals	11 / 21	52%	8 / 21	38%
Total	22 / 40	55%	14 / 40	35%

In Experiment 4a, 55% of participants used the expected dimensions, and 35% of participants in Experiment 4b used the expected dimensions. Although proportionally more participants did rely on expected dimensions in the gender experiment than in the talker experiment, the results of McNemar's test[29] revealed that there was no statistically significant difference between the proportion of subjects only using expected dimensions across experiments, p = 0.597. Further, within each experiment, two-tailed Fisher's exact tests revealed

[29] McNemar's test was used here instead of Fisher's or chi-square because the same participants took part in each experiment.

that there was no statistically significant difference between the proportion of subjects only using expected dimensions and their language background, $p = 0.761$ for the gender experiment, and $p = 0.748$ for the talker experiment. Thus, neither experiment nor listener language appears to affect whether a participant used the expected dimensions to perform the binned classification task. However, it is notable that at least for the gender–language–amplitude task, over half of participants did use the expected dimensions even without being given any explicit instructions to do so, indicating that these dimensions were highly salient to at least some listeners.

Within the participants who did only use expected dimensions, I examine several measures of performance. In order to do so, I first present the classification decisions made by this subset of participants in tree form in Figures 5.11 and 5.12 below.

Figure 5.11. *Classification results of those participants in Experiment 4a who completed all three phases of the task using only the three expected dimensions (N = 22). The white numbers in each box represent the number of total participants who grouped by that dimension in that phase. Orange values are the English monolinguals (out of 11) who chose that dimension, and blue values are the Mandarin-English bilinguals (out of 11) who chose that dimension.*

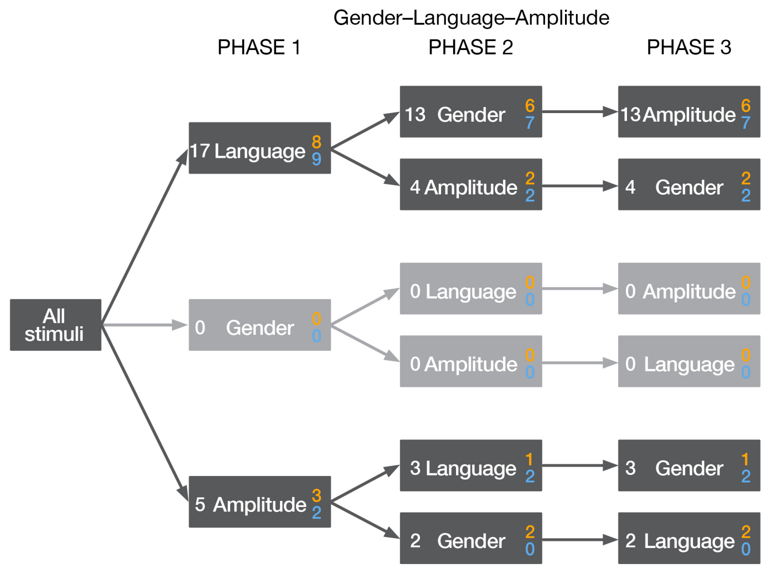
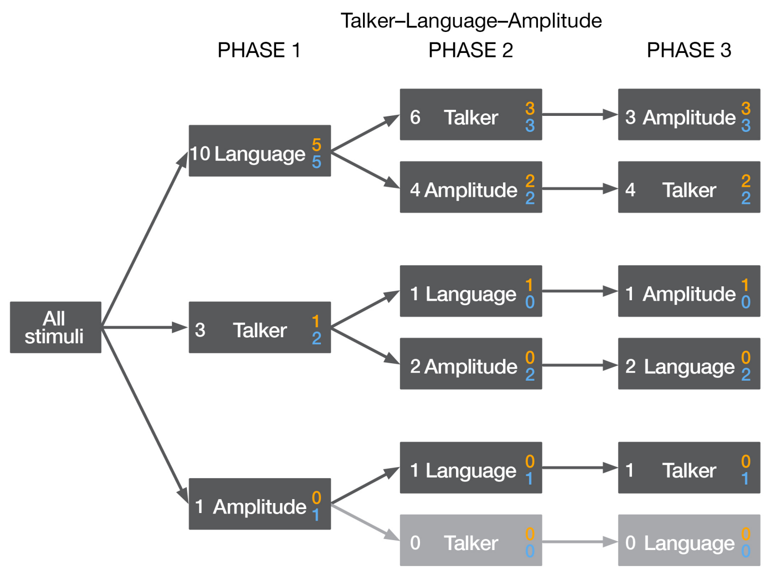


Figure 5.12. *Classification results of those participants in Experiment 4b who completed all three phases of the task using only the three expected dimensions (N = 14). Orange values are the English monolinguals (out of 6) who chose that dimension, and blue values are the Mandarin-English bilinguals (out of 8) who chose that dimension.*



5.4.4.1 Dimensional ordering agreement among participants only using anticipated dimensions

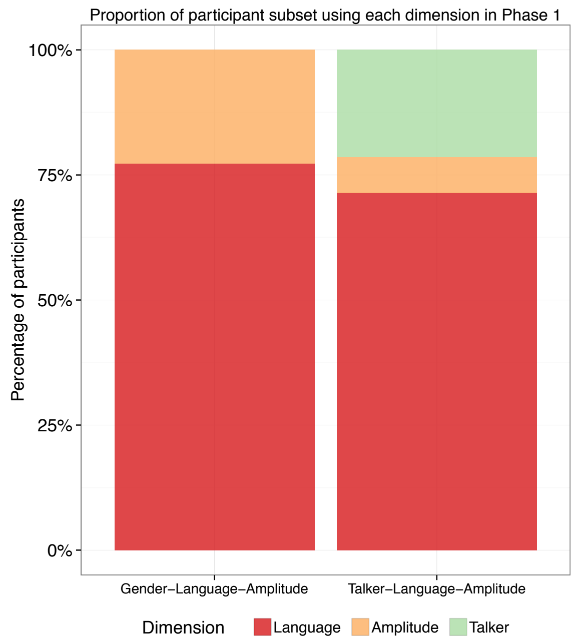
Within the subset of participants who only used anticipated dimensions, one way to assess whether participants approached the task differently in the presence of variability from gender versus variability from talker is to assess how consistent participants were within each experiment. To do so, the task behavior of each participant was summed up by the order in which they used each of the three dimensions for classification (e.g. language-gender-amplitude, amplitude-language-talker), as was done in order to create the histograms in Figures 5.5 and 5.6, above. Inter-participant agreement for these orderings is then used as a dependent measure. It could be that participants show agreement about the order in which dimensions were used for classification, or it could be that participants show dispersion in their orderings. If one version showed more agreement among participants than the other, the hierarchy of salience may be more defined for those three dimensions than for the three dimensions used in the other version.

In order to assess agreement, participants' orderings of dimensions across phases were first converted to a rank ordering matrix. Agreement among participants' rank ordering of dimensions was assessed via the nonparametric Kendall's coefficient-of-concordance (W) test. Kendall's W ranges from 0, indicating no agreement, to 1, indicating complete agreement. For the gender–language–amplitude experiment (Kendall's $W = 0.383$, $k = 3$, $N = 21$, $p = 0.0003$) the agreement between participants was stronger than for the talker–language–amplitude experiment (Kendall's $W = 0.231$, $k = 3$, $N = 13$, $p = 0.0498$). As the p-values for both experiments are below 0.05, the rank ordering among participants in both experiments was statistically concordant. Further, there was slightly more concordance among participants' rankings in Experiment 4a than for that same group of participants in 4b, indicated by larger Ws for Experiment 4a. In other words, the order in which dimensions were used by participants was more similar in the gender version than in talker version. However, the concordance strength is not very high for either experiment. Thus, it is possible that the salience hierarchy for gender, language, and amplitude is slightly more defined across participants than is the hierarchy for talker, language, and amplitude, though neither set of dimensions shows much agreement.

5.4.4.2 Individual dimension use among participants only using anticipated dimensions

Frequency of dimension use in Phase 1 for participants who only used anticipated dimensions is given in Figure 5.13 below. Frequency of dimension use across phases is not presented for this subset of participants, as frequencies would be equal for all dimensions by virtue of those participants' inclusion in the subset; all participants in this subset used all three anticipated dimensions, making all dimensions used equally frequently. Thus, only Phase 1 dimension use will be presented here.

Figure 5.13. Proportional dimension use by participants using anticipated dimensions only. Experiment 4a is on the left, 4b on the right.



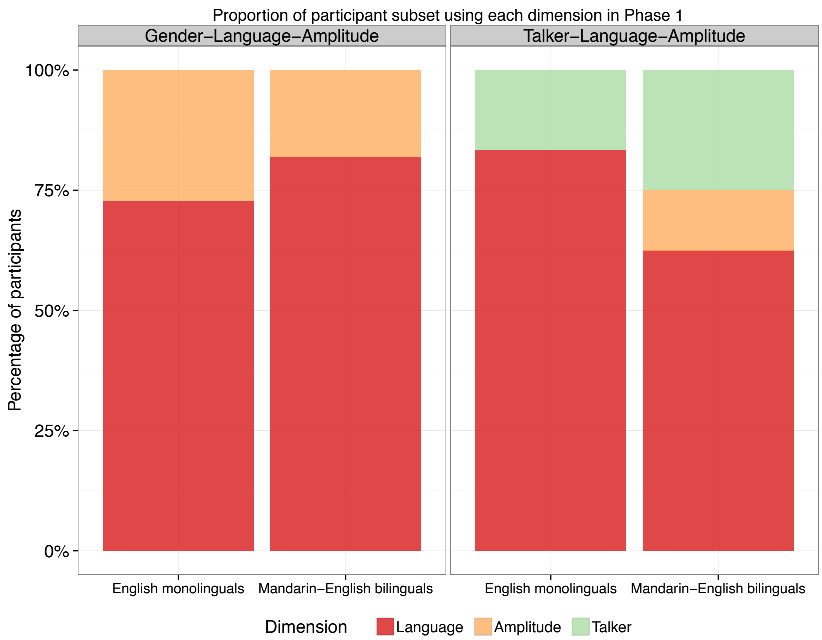
Analyzing results from the subset of participants who used only the dimensions intended (language, gender/talker, and amplitude) represents the most direct comparison to the speeded classification experiments, which only tested interference between those aforementioned dimensions. Thus, I examine how these participants used the dimensions with reference to the results from the speeded classification tasks in Experiments 1-3. For those participants only using anticipated dimensions, Figure 5.13 illustrates that, as was the case for the whole group, the language dimension was used most frequently in Phase 1, and was used equally between experiments (two-tailed Fisher's exact test $p = 0.712$). This again stands in contrast to predictions based on the results of Experiments 1-3, which predicted (a) that amplitude, not language, would be used first most frequently in both experiments, and (b) that language would be used in Phase 1 more frequently in the talker experiment than in the gender experiment. This disparity again underscores the difference between the two experimental paradigms, revealing the notable result that the two tasks do not make use of the same hierarchy of dimensions. Amplitude was not used in Phase 1 very often as compared with language, but it was used first more frequently in the gender version (25% of participants) than in the talker version (7% of participants), though not significantly so (two-tailed Fisher's exact test $p = 0.371$).

Language was so favored among this subset of participants that gender was not used at all in Phase 1 of Experiment 4a, and talker was only used first by 25% of participants in Experiment 4b. Again, this contrasts the prediction based on the speeded classification results that gender would be used first more frequently than talker, revealing differences between tasks in dimensional salience.

5.4.4.3 Language background comparison among participants only using anticipated dimensions

Next, I examine the use of dimensions in Phase 1 as a function of the language background of those participants using only expected dimensions. The amount that dimensions were used in Phase 1 across experiments is presented below in Figure 5.14, with participants separated by language background. There was no relationship between frequency of use of all dimensions in Phase 1 and listener language background for either experiment among this subset of participants, in line with the results observed in the speeded classification experiments. Two-tailed Fisher's exact tests revealed that dimension use and listener language were not related for Experiment 4a (gender–language–amplitude), $p = 1$, or for Experiment 4b (talker–language–amplitude), $p = 1$, meaning that English monolinguals and Mandarin-English bilinguals performed similarly in Phase 1 of each experiment.

Figure 5.14. Proportional dimension use by participants using conventional dimensions only, split by language background. Experiment 4a is on the left, 4b on the right. English monolingual participants (N = 11 for 4a, N = 6 for 4b) are to the left of Mandarin-English bilingual participants (N = 11 for 4a, N = 8 for 4b) for each experiment.



5.4.5 Item classification matrices

A different way to demonstrate classification performance in this task is to examine the extent to which each item was grouped with each other item. For each experiment, a matrix was created for each phase separately by calculating the number of times that a given item was grouped with each other item in that phase. Then, the individual matrices for each phase were added together, to give an overall matrix for that experiment. The counts were then converted to proportions. Tables 5.4 and 5.5 report item classification matrices for Experiments 4a and 4b, respectively.

Table 5.4. Item classification matrix for Experiment 4a (above).
Table 5.5. Item classification matrix for Experiment 4b (below). Each row represents an item. The proportion of times that an item was grouped in the same bin as another item is given in the cell corresponding to each item's column. Each matrix depicts the sum of separate matrices for Phase 1 (2 bin), Phase 2 (4 bin), and Phase 3 (8 bin). The matrix diagonal is in italics. Proportion corresponding to the item with which it was binned most frequently is given in bold.

Experiment 4a *Gender–Language–Amplitude*	English-Male-Loud-Sentence1	English-Male-Loud-Sentence2	English-Female-Loud-Sentence1	English-Female-Loud-Sentence2	English-Male-Soft-Sentence1	English-Male-Soft-Sentence2	English-Female-Soft-Sentence1	English-Female-Soft-Sentence2	Mandarin-Male-Loud-Sentence1	Mandarin-Male-Loud-Sentence2	Mandarin-Female-Loud-Sentence1	Mandarin-Female-Loud-Sentence2	Mandarin-Male-Soft-Sentence1	Mandarin-Male-Soft-Sentence2	Mandarin-Female-Soft-Sentence1	Mandarin-Female-Soft-Sentence2
English-Male-Loud-Sentence1	*1*	**0.7**	0.3	0.11	0.18	0.17	0.04	0	0.06	0.03	0.03	0.01	0.03	0	0.01	0.02
English-Male-Loud-Sentence2	**0.7**	*1*	0.12	0.17	0.32	0.28	0.03	0.04	0.05	0.06	0	0.02	0	0.02	0.03	0
English-Female-Loud-Sentence1	0.3	0.12	*1*	**0.73**	0.01	0.02	0.21	0.15	0.09	0.02	0.06	0.06	0.03	0	0.02	0.03
English-Female-Loud-Sentence2	0.11	0.17	**0.73**	*1*	0.08	0.15	0.22	0.32	0.01	0	0.02	0.05	0	0.01	0.04	0.01
English-Male-Soft-Sentence1	0.18	0.32	0.01	0.08	*1*	**0.76**	0.12	0.22	0	0.01	0.01	0	0.02	0.02	0	0.02
English-Male-Soft-Sentence2	0.17	0.28	0.02	0.15	**0.76**	*1*	0.18	0.23	0.02	0.01	0.01	0.04	0.01	0.06	0.09	0
English-Female-Soft-Sentence1	0.04	0.03	0.21	0.22	0.12	0.18	*1*	**0.79**	0.01	0.01	0.02	0.02	0.01	0.01	0.08	0.02
English-Female-Soft-Sentence2	0	0.04	0.15	0.32	0.22	0.23	**0.79**	*1*	0	0	0.02	0.02	0	0	0.05	0.02
Mandarin-Male-Loud-Sentence1	0.06	0.05	0.09	0.01	0	0.02	0.01	0	*1*	**0.78**	0.28	0.17	0.2	0.14	0.03	0.06
Mandarin-Male-Loud-Sentence2	0.03	0.06	0.02	0	0.01	0.01	0.01	0	**0.78**	*1*	0.12	0.15	0.21	0.28	0.03	0.08
Mandarin-Female-Loud-Sentence1	0.03	0	0.06	0.02	0.01	0.01	0.02	0.02	0.28	0.12	*1*	**0.78**	0.04	0	0.17	0.21
Mandarin-Female-Loud-Sentence2	0.01	0.02	0.06	0.05	0	0.04	0.02	0.02	0.17	0.15	**0.78**	*1*	0.07	0.02	0.22	0.18
Mandarin-Male-Soft-Sentence1	0.03	0	0.03	0	0.02	0.01	0.01	0	0.2	0.21	0.04	0.07	*1*	**0.78**	0.12	0.22
Mandarin-Male-Soft-Sentence2	0	0.02	0	0.01	0.02	0.06	0.01	0	0.14	0.28	0	0.02	**0.78**	*1*	0.16	0.15
Mandarin-Female-Soft-Sentence1	0.01	0.03	0.02	0.04	0	0.09	0.08	0.05	0.03	0.03	0.17	0.22	0.12	0.16	*1*	**0.76**
Mandarin-Female-Soft-Sentence2	0.02	0	0.03	0.01	0.02	0	0.02	0.02	0.06	0.08	0.21	0.18	0.22	0.15	**0.76**	1

Experiment 4b *Talker–Language–Amplitude*	English-Male1-Loud-Sentence1	English-Male1-Loud-Sentence2	English-Male2-Loud-Sentence1	English-Male2-Loud-Sentence2	English-Male1-Soft-Sentence1	English-Male1-Soft-Sentence2	English-Male2-Soft-Sentence1	English-Male2-Soft-Sentence2	Mandarin-Male1-Loud-Sentence1	Mandarin-Male1-Loud-Sentence2	Mandarin-Male2-Loud-Sentence1	Mandarin-Male2-Loud-Sentence2	Mandarin-Male1-Soft-Sentence1	Mandarin-Male1-Soft-Sentence2	Mandarin-Male2-Soft-Sentence1	Mandarin-Male2-Soft-Sentence2
English-Male1-Loud-Sentence1	*1*	**0.75**	0.18	0.21	0.08	0.09	0.04	0.14	0.06	0.08	0.03	0.02	0.01	0	0	0.01
English-Male1-Loud-Sentence2	**0.75**	*1*	0.14	0.21	0.11	0.09	0.04	0.15	0.1	0.02	0.02	0.02	0.03	0.01	0	0.02
English-Male2-Loud-Sentence1	0.18	0.14	*1*	**0.7**	0.14	0.19	0.23	0.17	0.02	0.06	0.09	0.07	0.03	0.08	0.01	0.08
English-Male2-Loud-Sentence2	0.21	0.21	**0.7**	*1*	0.12	0.03	0.09	0.19	0.02	0.03	0.07	0.05	0.01	0.02	0	0.02
English-Male1-Soft-Sentence1	0.08	0.11	0.14	0.12	*1*	**0.7**	0.3	0.22	0.02	0.02	0.08	0.03	0.08	0.1	0.01	0.05
English-Male1-Soft-Sentence2	0.09	0.09	0.19	0.03	**0.7**	*1*	0.38	0.19	0.01	0.04	0.06	0.02	0.06	0.1	0.02	0.11
English-Male2-Soft-Sentence1	0.04	0.04	0.23	0.09	0.3	0.38	*1*	**0.6**	0.01	0.06	0.03	0.02	0.01	0.02	0.07	0.18
English-Male2-Soft-Sentence2	0.14	0.15	0.17	0.19	0.22	0.19	**0.6**	*1*	0.03	0.02	0.02	0.01	0.01	0.03	0.05	0.08
Mandarin-Male1-Loud-Sentence1	0.06	0.1	0.02	0.02	0.02	0.01	0.01	0.03	*1*	**0.66**	0.22	0.2	0.12	0.12	0.03	0.04
Mandarin-Male1-Loud-Sentence2	0.08	0.02	0.06	0.03	0.02	0.04	0.06	0.02	**0.66**	*1*	0.21	0.24	0.11	0.12	0.12	0.07
Mandarin-Male2-Loud-Sentence1	0.03	0.02	0.09	0.07	0.08	0.06	0.03	0.02	0.22	0.21	*1*	**0.69**	0.12	0.06	0.11	0.22
Mandarin-Male2-Loud-Sentence2	0.02	0.02	0.07	0.05	0.03	0.02	0.02	0.01	0.2	0.24	**0.69**	*1*	0.13	0.07	0.12	0.12
Mandarin-Male1-Soft-Sentence1	0.01	0.03	0.03	0.01	0.08	0.06	0.01	0.01	0.12	0.11	0.12	0.13	*1*	**0.7**	0.26	0.26
Mandarin-Male1-Soft-Sentence2	0	0.01	0.08	0.02	0.1	0.1	0.02	0.03	0.12	0.12	0.06	0.07	**0.7**	*1*	0.31	0.18
Mandarin-Male2-Soft-Sentence1	0	0	0.01	0	0.01	0.02	0.07	0.05	0.03	0.12	0.11	0.12	0.26	0.31	*1*	**0.67**
Mandarin-Male2-Soft-Sentence2	0.01	0.02	0.08	0.02	0.05	0.11	0.18	0.08	0.04	0.07	0.22	0.12	0.26	0.18	**0.67**	*1*

There are several things to note from these matrices. First, the stimuli in each matrix are sorted by language, the English stimuli in the first half and the Mandarin stimuli in the second half. Thus, these matrices can be used to further demonstrate that language is a frequently used dimension across phases. Imagining that the matrices were divided into quadrants by stimulus language, the values in the quadrants where stimuli were paired with same-language stimuli (top-left quadrant for English-English and bottom-right quadrant for Mandarin-Mandarin) tend to be higher than the other two quadrants, illustrating the dominance of the language dimension.

These item-level matrices also allow us to make observations about the grouping together of individual stimuli that would be impossible to make from the previous analyses, which all were derived from an analyst-designated code for each of a participant's three phases, collapsing across stimuli. In Tables 5.4 and 5.5, looking across each row for each individual stimulus, the largest proportion is bolded. The stimulus whose column contains that bold value is, thus, the stimulus most frequently binned with the reference stimulus. It is observable in each matrix that the bold value for every stimulus is, in fact, its matched pair in terms of language, gender or talker, and amplitude values. In other words, the two stimuli that had the same values in all three expected dimensions were, indeed, considered to be most similar to each other. This observation underscores the salience of these dimensions to participants, despite their ability to use any potential criteria to bin stimuli.

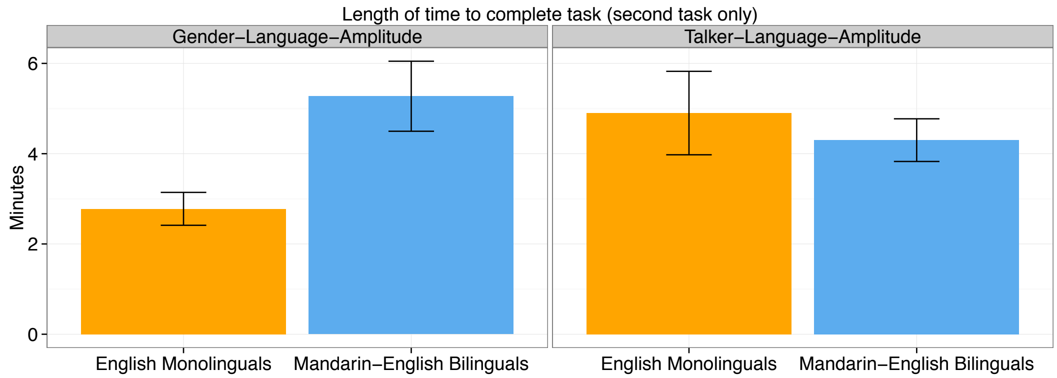
Such matrices could, of course, serve as input to a clustering or multidimensional scaling analysis. These analyses would offer additional information about similarities between individual items, which would be useful if the goal of this inquiry was to understand the classificatory role of particular characteristics of particular stimuli, beyond the dimensions of interest (e.g. whether there were specific properties of the phonetic or linguistic content of the items that caused certain samples to be grouped together more frequently). However, this type of analysis would not contribute to understanding the relationships between the dimensions of interest beyond what has already been determined, and therefore will not be conducted.

5.4.6 Time to task completion

Finally, one additional analysis compares participants' task performance across experiments and listener backgrounds using a different dependent measure: the length of time it took participants to complete their classifications. Though participants were allowed to use as much time as they wished to complete the task[30], time to task completion may still be a useful metric, as it may reflect that the ability to classify certain dimensions is more defined than for other dimensions, or that participants from certain language backgrounds show more defined dimensional hierarchies. Due to experimental design considerations, however, only the task that participants performed second can be analyzed in this manner.[31]

[30] While the experimenter was prepared to cut participants off at 30 minutes, no participant came close to taking this long.
[31] The experiment was programmed such that the "task start" timestamp was recorded as soon as the instruction page was displayed on the screen. Sometimes, the experimenter opened the experiment in the web browser before the participant arrived for the experiment, thereby recording task durations that were much longer than the actual time participants spent performing the first version of the experiment. However, the second task was always completed directly following viewing the secondary instructions, and therefore the duration recorded for this version of the experiment is a reliable measure of the time it took for participants to complete the second version of the task.

Figure 5.15. Task duration for each experiment, split by listener language. Experiment 4a is on the left, 4b on the right.

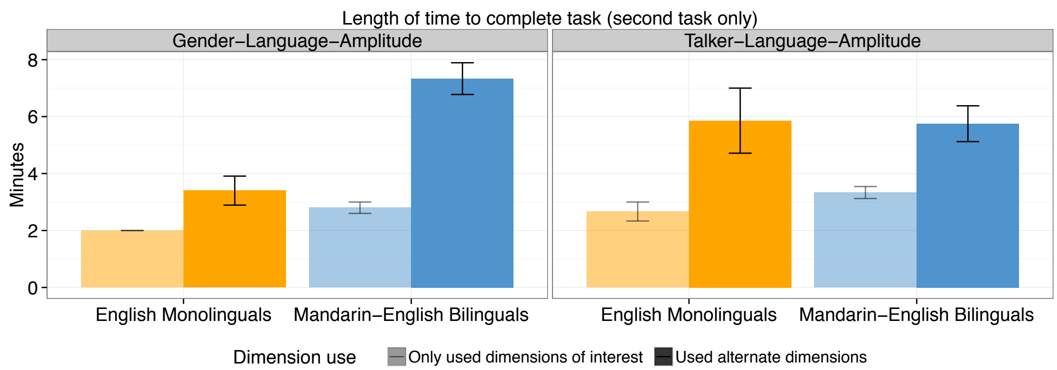


From Figure 5.15, it appears that Mandarin-English bilingual participants took longer than English monolingual participants to complete Experiment 4a, but that the listener groups were more matched in Experiment 4b. A two-way ANOVA[32] with the within-subjects factors of experiment (4a and 4b) and listener language background (English monolinguals and Mandarin-English bilinguals) found no main effect of experiment, $F(1,36) = 0.422$, $p = 0.520$, and no main effect of listener language, $F(1,36) = 8.79$, $p = 0.184$. This was mediated, however, by a significant interaction between experiment and listener language, $F(1,36) = 4.97$, $p = 0.032$. Follow-up simple effects testing revealed that there was a difference between listener groups in the gender–language–amplitude experiment, $t(18) = -2.706$, $p = 0.014$, such that English monolinguals ($M = 2.8$ minutes) were faster than Mandarin-English bilinguals ($M = 5.3$ minutes), but that there was no difference in task duration between listener groups in the talker–language–amplitude experiment, $t(18) = 0.578$, $p = 0.571$. This discrepancy is likely rooted in the fact that more Mandarin-English bilinguals used unexpected dimensions than

[32] Repeated measures was not used because the test was only conducted on the duration of one experiment (the one performed second) from each participant.

English monolinguals (as pointed out in Section 5.4.2.3 above), and that trend was significant for the gender version but not for the talker version. As can be seen in Figure 5.16, below, participants who sorted stimuli by alternative dimensions took longer to perform each task than participants who only used the expected dimensions. Given the variety of sometimes complex rationales reported by participants using alternate dimensions (and discussed in Section 5.4.3 above), it is no surprise that these participants took longer to complete the task.

Figure 5.16. Task duration for each experiment, split by listener language and whether participants used dimensions of interest. Experiment 4a is on the left, 4b on the right.



5.5 Summary and conclusion

The multi-phase binned classification task provided more information about the relationship between the language-being-spoken, talker, gender, and amplitude dimensions. To review, the language-being-spoken dimension was used most frequently by all participants in Phase 1, despite predictions otherwise. However, when all three phases were included, the frequency of use of all dimensions was more evenly distributed. Language-being-spoken, amplitude, and unanticipated dimensions were used equally frequently in both the gender version (Experiment 4a) and the talker version (Experiment 4b), when looking at Phase 1 only or at all three phases. There was no difference between the frequency that gender was used in Experiment 4a and the frequency that talker was used in Experiment 4b, either in Phase 1 only or across all three phases. Thus far, then, all of the predictions made by the results of the speeded classification task have been disconfirmed (as has the prediction made by previous free classification tasks); rather than providing converging evidence for the dimensional hierarchy revealed in the speeded classification tasks, the binned classification task reveals a dimensional hierarchy of its own, underscoring the differences between the nature of the two tasks and revealing that dimensional hierarchies are not necessarily stable for listeners across different task demands and different phases of processing.

However, in comparing the frequency of use of language-being-spoken versus gender in Experiment 4a across all phases and language-being-spoken versus talker in Experiment 4b across all phases, an unanticipated prediction based on the results of the speeded classification tasks is confirmed. Namely, across all phases, the frequency of gender use was equal to the frequency of language-being-spoken use in Experiment 4a, but language-being-spoken was used more frequently than talker across all phases of Experiment 4b. This result is consistent with previous findings from Experiments 1 and 2 of this dissertation: language-being-spoken and gender showed mutual and symmetrical interference, while language-being-spoken and talker showed mutual and asymmetrical interference such that language-being-spoken interfered more with talker than vice versa.

In terms of language background, frequency of dimension use did not depend on language background when including all phases of the task at once. However, when examining Phase 1 only, there was a difference in listener language backgrounds in the gender version but not in the talker version. This difference between language backgrounds is accounted for by the fact that Mandarin-English bilinguals used more unanticipated dimensions than English monolinguals in Phase 1 of Experiment 4a. The increased usage of unanticipated dimensions by Mandarin-English bilinguals resulted in a significant interaction between experiment and listener language in the length of time it took to complete the (second) task. Because of the fact that classifying "other" dimensions took participants longer than classifying the expected dimensions, English monolinguals completed the gender experiment faster than Mandarin-English bilinguals, but there was no difference between language backgrounds for the talker experiment in the amount of time taken to complete the task.

Most of the unanticipated dimensions used were related to grammatical features or the meaning of the sentences. As such, it is logical that Mandarin-English bilinguals used more of these content-based dimensions than did English monolinguals, because Mandarin-English bilinguals had access to the linguistic content of all stimuli, while English monolinguals could only use such dimensions for half the stimuli.

However, many participants did use only the anticipated dimensions to complete the task, indicating that these dimensions were indeed salient to listeners. In an item-level classification matrix summing across all phases, stimuli were most often binned with their exact matches along the anticipated dimensions, another indication that these dimensions were salient. There were no differences between experiments or between language backgrounds in the proportion of participants who only used expected dimensions across all phases. Among this subset of participants there was low agreement across participants in the dimensional orderings they used, but there was slightly more agreement among participants in the gender experiment than the talker experiment. There were no differences between experiments or between language backgrounds in the frequency of use of language-being-spoken, gender, talker, and amplitude in Phase 1 for this subset of participants.

In sum, the results of the multi-phase binned classification task shed more light on the relationship between language-being-spoken, talker, gender, and amplitude. However, the parallels with the previous experiments in this dissertation were found in unexpected places. Amplitude was not the most salient dimension in this experiment, as it was in Experiments 1-3. Instead, language-being-spoken was the most salient. This discrepancy likely reflects the fact that the two tasks require participants to use the dimensions quite differently. Amplitude is salient in a speeded task, where it interferes when another dimension is being attended to, while language-being-spoken is salient when participants can take more time to reflect on and compare the importance of multiple dimensions. This result brings to light the important finding that the relative salience of a dimension to a listener is dependent on the processing phase and task demands of the situation. Despite these differences, the results of Experiments 1-3 were echoed in this experiment in terms of the relative use of language-being-spoken as compared with gender and talker over the course of the experiment. Language-being-spoken and gender were used by participants equally frequently, while language-being-spoken was used more than talker. Thus, the hierarchy between language-being-spoken, gender, and talker established based on the speeded classification task appears to extend to the cognitive processes involved in the binned classification task, though these results are between-participants rather than within-participants, as was the case in the speeded classification task. However, more work remains to be done in delimiting the bounds of this hierarchy. For example, future work aimed at determining the cognitive basis of the binned classification task may help to explain why participants behave differently with respect to amplitude in the two tasks. Further, future studies comparing the relationship between these dimensions in yet another task domain may help to pinpoint how cognitively far these dimensional dependency relations extend, and what factors may influence the relative salience of dimensions at different phases of processing.

CHAPTER 6: CONCLUSION

6.1 Chapter outline

This final chapter begins by reviewing the main aims and goals of this dissertation. Then, I summarize the experimental design and results of Experiments 1-3, and then do the same for Experiment 4. Next, the insights gained from the use of multiple experimental paradigms are examined, before discussing what is to be made of the equivalent behavior by monolinguals and bilinguals across almost all metrics of these experiments. Finally, the chapter closes with remarks on the status of the language-being-spoken dimension.

6.2 Overall summary of the dissertation experiments

The goal of this dissertation was to explore how the processing of "which language is being spoken" is related to the processing of other speech dimensions in the minds of monolinguals and bilinguals. In this work, language-being-spoken was conceived of as an indexical dimension, as it does not, itself, convey linguistic information but rather gives information about the context of the utterance. While prior work has examined relationships between indexical and linguistic dimensions of speech in processing, no studies have, up until this dissertation, examined such relationships between indexical dimensions. Further, no previous work has investigated the relationships between indexical dimensions in bilingual processing. Thus, in this dissertation the relationship between language-being-spoken and other speech dimensions (two indexical, one non-linguistic) was examined at several points in processing for monolingual and bilingual listeners through the use of two experimental paradigms. The first paradigm, the speeded classification task (also called the Garner paradigm, Experiments 1-3), taps into an earlier point in processing than does the other paradigm, the multi-phase binned classification task (Experiment 4).

6.3 Speeded classification experiments (Experiments 1-3)

In Experiments 1-3, the language-being-spoken dimension, which in this dissertation always varied between English or Mandarin Chinese, was paired with a series of other dimensions in the speeded classification task. This paradigm was used to examine listeners' processing hierarchy of these dimensions by assessing whether they could selectively attend to one dimension at a time, while ignoring irrelevant variation in the other dimension. In each experiment, participants from different language backgrounds (English monolinguals and Mandarin-English bilinguals in Experiments 1-3, with the addition of participants who were bilingual in English and a language besides Mandarin in Experiment 2) listened to sentences and were asked to selectively attend to one dimension at a time. In the two tasks of Experiment 1, listeners classified stimuli by language-being-spoken (English or Chinese) and by another indexical dimension, gender (male or female). In Experiment 2, listeners classified stimuli based on language-being-spoken and the indexical dimension talker identity (two male talkers, named Wei or Li). And, in Experiment 3, listeners classified stimuli based on language-being-spoken and the non-linguistic dimension amplitude (loud or soft). Particular emphasis was placed on Experiment 2, as previous work has been devoted to examining talker classification across languages, most notably in documenting the language familiarity benefit for talker identification.

A series of hypotheses were presented to predict speeded classification results. First, after an extensive review of previous Garner studies, a hypothesis based on the levels of processing of each pair of dimensions was put forward to predict dependency relations between dimensions. Specifically, dimensions on the same level of processing were predicted to show mutual and symmetrical interference, while dimensions on different levels of processing were predicted to show asymmetrical interference, with the dimension thought to be on a shallower level of processing the dimension causing more interference than vice versa. Another hypothesis was formed to predict the results of the language-being-spoken–talker pairing based on prior work documenting the relative language-specificity of talker identification and the relative talker-generality of language identification (the LS/TG hypothesis). Specifically, from this work it was predicted that it would be harder to ignore variability in language-being-spoken when classifying talker than vice versa. Finally, a series of competing hypotheses were put forward to predict the ways in which listeners from different language backgrounds would differ in their dependency relations between dimensions. Two hypotheses were based on the idea that enhanced representations of dimensions would cause listeners to show greater Garner interference in Experiments 1 and 2 (where was thought possible for dimensions to be affected by language differences). Specifically, the language familiarity enhances dimensional representations hypothesis (LF-DR) predicted that listeners familiar with both languages (Mandarin-English bilinguals) would show more interference than listeners with familiarity with only one language (English monolinguals, and non-Mandarin-English bilinguals in Experiment 2). In contrast, the bilingualism enhances dimensional representations hypothesis (B-DR) predicted that bilingual listeners (Mandarin-English bilinguals, and non-Mandarin-English bilinguals in Experiment 2) would exhibit more interference than monolinguals (English monolinguals). Finally, a third hypothesis (the bilingualism enhances selective attention hypothesis, B-SA) predicted that bilinguals' purported advantage over monolinguals in selective attention tasks would result in bilingual listeners exhibiting less interference than monolingual listeners.

In Experiment 1, language-being-spoken was found to have a mutual and symmetrical pattern of interference with the indexical dimension gender (and gender was classified faster than language-being-spoken at baseline). Notably, these results did not change in a second version of this experiment using a different pair of talkers with different levels of intelligibility. In Experiment 2, language-being-spoken showed a mutual and asymmetrical pattern of interference with the indexical dimension talker, where it was harder for listeners to ignore language-being-spoken when attending to talker than the reverse (and language-being-spoken and talker were classified equally quickly at baseline). And, in Experiment 3, language-being-spoken showed an asymmetrical pattern of interference with the non-linguistic dimension amplitude, where listeners could not ignore amplitude when attending to language-being-spoken, but could ignore language-being-spoken when attending to amplitude (and amplitude was classified faster than language-being-spoken at baseline). In addition to revealing the dependency relations between language-being-spoken and these other dimensions, these experiments provided another novel result to the literature regarding the processing time of the language-being-spoken dimension at baseline as compared with these other dimensions. In control blocks, language-being-spoken was found to be processed equally as quickly as talker, but slower than gender and amplitude. Taken together, these results provide a sense of language-being-spoken's place within a dimensional processing hierarchy: in the Garner paradigm, language-being-spoken was equally as salient as gender, more salient than talker, and less salient than amplitude.

The levels of processing hypothesis did not predict the results of Experiments 1 and 2. In

Experiment 1, gender was predicted to be processed at a shallower level than language-being-spoken, and was indeed significantly faster than language-being-spoken in control blocks, but the two dimensions showed symmetrical interference. In Experiment 2, talker and language-being-spoken were predicted to operate at equal levels of processing, and indeed were equivalent in control blocks, but there was mutual and asymmetrical interference. However, the levels of processing hypothesis did account for the results of Experiment 3; the non-linguistic dimension amplitude interfered with language-being-spoken, which was assumed to be processed at a later stage, and there was no evidence of interference from language-being-spoken processing in the amplitude task. Future work should continue to investigate the relationship between levels of processing, discriminability, and dependency relations between dimensions in the Garner paradigm, a relationship which remains unclear.

Notably, across all three experiments there were no differences between listener groups from different language backgrounds with respect to patterns of Garner interference across dimensions; the interference observed was stable across participants, regardless of language background. Thus, all three hypotheses regarding listener language backgrounds were refuted, and instead the processing relationship between language-being-spoken and each of these dimensions is established as one independent of listeners' linguistic experience.

Since language-being-spoken was found to show some pattern of interference with all three dimensions with which it was paired, this indicates that the mechanisms involved in processing language-being-spoken are not independent of the mechanisms involved in processing gender, talker, or amplitude. Even though there were processing asymmetries between two pairings (language-being-spoken–talker and language-being-spoken–amplitude), across these experiments no dimension could be completely ignored when attempting to selectively attend to language-being-spoken; irrelevant variation from gender, talker, and amplitude complicated the listener's task of classifying a stimulus by language-being-spoken. In only one case, amplitude processing, could irrelevant variability in language-being-spoken be ignored; language-being-spoken could not be ignored when attending to gender or talker. This dissertation thus provides novel evidence about the patterns of integrality of language-being-spoken with gender, talker, and amplitude.

6.4 Multi-phase binned classification experiment (Experiment 4)

Experiment 4 investigated the relationship between language-being-spoken and other dimensions by using a modified version of the free classification paradigm, here called the multi-phase binned classification task. Two versions of this task were conducted, one using stimuli varying the dimensions language-being-spoken, gender, and amplitude, and the other using stimuli varying language-being-spoken, talker, and amplitude. In both versions, both English monolingual and Mandarin-English bilingual listeners grouped sixteen stimuli into bins using whatever criteria they chose, beginning with two bins (Phase 1), then four bins (Phase 2), followed by eight bins (Phase 3). The dimension used most frequently in Phase 1 was considered to be the most salient dimension.

One hypothesis was put forward which predicted that the same hierarchy of dimensions used in the speeded classification task would also be used in this task. Specifically, that hypothesis predicted: in the first version of the experiment, amplitude would be used first followed by an equally likely chance of using gender or language-being-spoken, and in the second version of the experiment, amplitude would be used first followed by talker followed by language-being-spoken. Another hypothesis predicted that gender would be salient in this experiment, as it had been used frequently in previous free classification studies. Finally, hypotheses based on listener language backgrounds suggested that either (following the results of the speeded classification experiments), the two groups would not differ, or (if the two tasks are sufficiently different and listener groups behave differently at different phases) that monolinguals and bilinguals would differ in the extent to which language-being-spoken was used.

The results showed that, for both versions of the experiment, the language-being-spoken dimension was used in Phase 1 most frequently, indicating that it was most salient to listeners. Across all phases, however, dimension use was more equally distributed. Further, across all phases, gender and language-being-spoken were used equally frequently, while talker was used less frequently than language-being-spoken, a result which mirrors the results of Experiments 1 and 2. There were no differences between language backgrounds in the amount that anticipated dimensions (language-being-spoken, gender, talker, and amplitude) were used, but Mandarin-English bilinguals used more unanticipated dimensions than English monolinguals, likely resulting from their ability to use content or grammar-based dimensions across stimuli in both languages. Thus, none of the hypotheses completely predicted the results. These findings, then, provide an important complement to the results of the speeded classification task, and suggest that the two tasks tap into different phases in processing, each phase having its own relative hierarchy of dimensional salience. This dissertation thus provides novel evidence about the relevance of language-being-spoken at different phases of processing.

6.5 Comparison of results across paradigms

As was just noted, the results of the speeded classification task in Experiments 1-3 and the multi-phase binned classification task in Experiment 4 mirror each other in some ways, but in other ways are quite different. In terms of similarities, the relationship of language-being-spoken with gender and talker are comparable across the two paradigms. In Experiment 4, across all phases, participants used gender and language-being-spoken to group stimuli with the same frequency, but used talker less frequently than language-being-spoken. Though it represents performance between- rather than within-participants, this asymmetry echoes the results of Experiments 1 and 2, finding that language-being-spoken and gender exhibited symmetrical interference, while there was an asymmetry between language-being-spoken and talker where language-being-spoken interfered more with talker than vice versa. That the binned classification task is in some way consistent with the dimensional relations between language-being-spoken and gender, and language-being-spoken and talker, lends further support to this pattern.

In terms of differences, the dimension found to be most salient was not the same for each paradigm. Across the pairwise comparisons tested in the three speeded classification experiments, amplitude was the only dimension that interfered with other dimensions more than it was interfered with by other dimensions, and in fact was the only dimension to not show mutual interference with language-being-spoken, indicating its status as a highly salient dimension according to the Garner paradigm. (It is important to note, however, that language-being-spoken was salient in certain pairings of the Garner paradigm; it was equally as salient as gender, and more salient than talker. However, in language-being-spoken's pairing with amplitude, amplitude captured more attention.) In the binned classification task, however, it was not amplitude but language-being-spoken which was most salient, being used in the first phase of the task most frequently, far more than any other dimension. (That said, across all phases of the experiment, the dominance of language-being-spoken use decreased.) In sum, different dimensions were salient in different tasks.

The divergence between the speeded classification task and the multi-phase binned classification task in dimensional salience reveals that the nature of relationships between dimensions in processing is not static over the course of processing. Each of these paradigms taps into a different stage of processing, and the results from each are reflective of the relative importance of dimensions at each stage. From these findings, then, we can make observations about the status of the central dimension of this dissertation, language-being-spoken, at multiple points in processing. It appears that while language-being-spoken is not necessarily the most salient dimension at an earlier phase of processing (as evidenced by the Garner task results), its relevance to the listener emerges later on in processing (as evidenced by the binned classification results). Moreover, language-being-spoken's relationship with talker is consistent at both phases of processing: language-being-spoken captures more attention than talker in both tasks. It is hard for listeners, regardless of language background, to ignore what language is being spoken when classifying talkers.

That one dimension was not the most salient dimension across all tasks underscores the importance of testing with multiple paradigms. Investigating the central question from a variety of approaches has yielded important insights. Future work could utilize yet another experimental approach to provide an even more nuanced picture of language-being-spoken's place in dimensional hierarchies across phases of processing. An additional benefit of implementing a different paradigm in future work is that it may alleviate some of the problems with interpreting results of speeded classification studies with respect to mechanisms involved, which stem from the inherent limitations of reaction times as a measure (as discussed in Section 2.8).

For example, a study using an oddball detection paradigm would provide similar types of information as the binned classification task about which dimensions are salient to participants in the presence of variability in multiple dimensions, but would do so on a faster time scale. In such a study, stimuli could be designed as in the binned classification task, containing information about three dimensions at once. Faced with a trial containing the stimuli Chinese-Female-Loud, Chinese-Female-Soft, and English-Female-Loud, would participants be more likely to choose Chinese-Female-Soft (representing amplitude) or English-Female-Loud (representing language-being-spoken) as the odd one out under the time pressure of the oddball detection task? In other words, is language-being-spoken the most salient dimension (as was the case in the binned classification task) or is amplitude (as was the case in the speeded classification tasks)? This design could provide more insight into the changing meaning of salience across tasks: If amplitude was to be found most salient in this design, that might provide more evidence that amplitude is salient early on in processing, as it would be salient in both speeded tasks. If language-being-spoken was to be found most salient in this task, that may indicate that amplitude is only most salient in the context of selective attention task, and that language-being-spoken is most salient when faced with a choice of more than two dimensions. While this is a highly simplistic depiction of the possibilities for future studies, investigating dimensional hierarchies from a variety of paradigms tapping into various phases of processing would surely provide greater insight into these questions.

6.6 Equal performance of monolinguals and bilinguals

One of the most surprising results in this series of experiments, given that the experiments were set up to determine how English monolinguals and Mandarin-English bilinguals differed their classification of language-being-spoken as related to other dimensions, was that there were no differences between these listener groups. Instead, monolinguals and bilinguals behaved similarly on all speeded classification experiments. Further, by most metrics of analysis of the multi-phase binned classification task, English monolinguals and Mandarin-English bilinguals also performed similarly; the two groups did not differ in their degree of use of dimensions of interest across phases, nor did they differ in the dimension they used in the first phase of the task.

As was described in detail in Chapter 4, a number of methodological decisions may have been responsible for this lack of difference, at least in the speeded classification tasks. In particular, the use of longer stimuli, and participants' potential lack of engagement with the linguistic level during the task, may have resulted in the equal performance of monolinguals and bilinguals on the speeded classification experiments. Future work is needed to determine whether the current findings are the result of these methodological choices.

However, as the present results stand, it is the case that English monolinguals and Mandarin-English bilinguals do not differ in their processing of language-being-spoken in relation to gender, talker, and amplitude. It does not appear to matter whether a listener is familiar with a language, or is bilingual; the language-being-spoken dimension still displays the same dependency relations with gender, talker, and amplitude. This observation makes a prediction that could be tested by future studies. The same set of speeded classification experiments in this dissertation could be conducted using the same English monolingual and Mandarin-English bilingual participants, but using stimuli spoken in two languages unknown to all participants. For these future participants classifying stimuli in two unknown languages, the patterns of dependency should be the same as they were with Mandarin and English stimuli, since language familiarity does not appear to play a role in listeners' dimensional hierarchies. The bilingual status of participants should also continue to not play a role in dependency relations between these dimensions. Such a follow-up study would confirm that the integrality of language-being-spoken with gender, talker, and amplitude truly is independent of the language background status of the listener. Even further follow-up studies could directly compare language-being-spoken with a linguistic dimension like consonant to determine whether variation in the language being spoken would interfere with listeners attempting to selectively attend to consonant. A linguistic–language-being-spoken comparison such as this would be quite interesting, and is necessary in order to understand language-being-spoken's relationship with linguistic dimensions in processing, and to further understand language-being-spoken's role in human language processing.

6.7 Conclusion

This dissertation has provided novel evidence about the patterns of integrality between language-being-spoken and gender, talker, and amplitude, and the salience of these dimensions at various points in processing, for both monolinguals and bilinguals. Moreover, the findings of these four experiments present a picture of the dimension language-being-spoken as one that is equally important to listeners whether or not they are familiar with all of the languages spoken, and more relevant to listeners later on in processing than earlier. In sum, this dissertation has contributed to the understanding of the relationships between various indexical dimensions of speech in processing, particularly with respect to bilinguals.

REFERENCES

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *115*, 3171–3183.

Andics, A., McQueen, J.M., & Turennout, M. van (2007). Phonetic context influences voice discriminability. *Proceedings of the XVIth International Congress of Phonetic Sciences [ICPhS12]* (pp. 1829–1832)*,* Saarbrücken, Germany.

Ashby, F. G., & Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology*, *38*(4), 423–466.

Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *The Journal of the Acoustical Society of America*, *71*, 975–989.

Atagi, E. & Bent, T. (2011). Perceptual dimensions of non-native speech. *Proceedings of the XVIIth International Congress of Phonetic Sciences [ICPhS 2011]* (pp. 260–263). Hong Kong: Department of Chinese, Translation and Linguistics, City University of Hong Kong.

Atagi, E., & Bent, T. (2013). Auditory free classification of nonnative speech. *Journal of Phonetics*, *41*(6), 509–519.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403,* 309–312.

Ben-Artzi, E., & Marks, L. E. (1999). Processing linguistic and perceptual dimensions of speech: Interactions in speeded classification. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(3), 579–595.

Bialystok, E. (1988). Levels of bilingualism and levels of linguistic awareness. *Developmental Psychology, 24,* 560–567.

Bialystok, E. (1992). Attentional control in children's metalinguistic performance and measures of field independence. *Developmental Psychology, 28*, 654–664.

Bialystok, E. (1999). Cognitive complexity and attentional control in the bilingual mind. *Child Development, 70*, 636–644.

Bialystok, E. (2006). Effect of bilingualism and computer video game experience on the Simon task. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale*, *60*(1), 68–79.

Bialystok, E., Craik, F. I. M., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: Evidence from the Simon task. *Psychology and Aging, 19*, 290–303.

Bialystok, E., Craik, F. I. M., & Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 859–873.

Bialystok, E., Craik, F. I., & Luk, G. (2012). Bilingualism: Consequences for mind and brain. *Trends in Cognitive Sciences*, *16*(4), 240–250.

Bialystok, E., & Martin, M. M. (2004). Attention and inhibition in bilingual children: Evidence from the developmental change card sort task. *Developmental Science, 7*, 325–339.

Bialystok, E., Martin, M. M., & Viswanathan, M. (2005). Bilingualism across the lifespan: The rise and fall of inhibitory control. *International Journal of Bilingualism*, *9*(1), 103–119.

Biederman, I., & Checkosky, S. F. (1970). Processing redundant information. *Journal of Experimental Psychology*, *83*, 486–490.

Blechner, M. J., Day, R. S., & Cutting, J. E. (1976). Processing two dimensions of nonspeech stimuli: The auditory-phonetic distinction reconsidered. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(2), 257–266.

Boenke, L. T., Ohl, F. W., Nikolaev, A. R., Lachmann, T., & Leeuwen, C. V. (2009). Different time courses of Stroop and Garner effects in perception—An Event-Related Potentials Study. *NeuroImage*, *45*(4), 1272–1288.

Boersma, P. & Weenink, D. (2011). Praat: doing phonetics by computer [Computer program].

Bőhm, T., & Shattuck-Hufnagel, S. (2009). Do Listeners Store in Memory a Speaker's Habitual Utterance-Final Phonation Type?. *Phonetica*, *66*(3), 150–168.

Bond, Z. S., & Fokes, J. (1991). Identifying foreign languages. *Proceedings of the XIIth Congress of Phonetic Sciences [ICPhS12]* (pp. 198–201). Aix-en-Provence.

Bosch, L., & Sebastián-Gallés, N. (2003). Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. *Language and Speech*, *46*(2–3), 217–243.

Bradlow, A. R. (1996). A Perceptual Comparison of the /i/–/e/ and /u/–/o/ Contrasts in English and in Spanish: Universal and Language-Specific Aspects. *Phonetica*, *53*(1–2), 55–85.

Bradlow, A. R., Ackerman, L., Burchfield, L. A., Hesterberg, L., Luque, J. S., & Mok, K. (2010). ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings. Department of Linguistics, Northwestern University. Retrieved from http://groups.linguistics.northwestern.edu/speech_comm_group/allsstar/index.html.

Bradlow, A. R., Ackerman, L., Burchfield, L., Hesterberg, L., Luque, J., and Mok, K. (2011). Language- and talker-dependent variation in global features of native and non-native speech. *Proceedings of the XVIIth International Congress of Phonetic Sciences [ICPhS 2011]* (pp. 356–359). Hong Kong: Department of Chinese, Translation and Linguistics, City University of Hong Kong.

Bradlow, A. R., Clopper, C., Smiljanic, R., & Walter, M. A. (2010). A perceptual similarity space for languages: Evidence from five native language listener groups. *Speech Communication, 52*, 930–942.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics, 61 (2)*, 206–219.

Bregman, M. R., & Creel, S. C. (2014). Gradient language dominance affects talker learning. *Cognition*, *130*(1), 85–95.

Brunelle, M. (2012). Dialect experience and perceptual integrality in phonological registers: Fundamental frequency, voice quality and the first formant in Cham. *The Journal of the Acoustical Society of America*, *131*(4), 3088–3102.

Bunge, S. A., Dudukovic, N. M., Thomason, M. E., Vaidya, C. J., & Gabrieli, J. D. (2002). Immature frontal lobe contributions to cognitive control in children: evidence from fMRI. *Neuron*, *33*(2), 301–311.

Caramazza, A., Yeni-Komshian, G. H., Zurif, E. B., and Carbone, E. (1973). The acquistion of a new phonological contrast: The case of stop consonants in French-English bilinguals. *The Journal of the Acoustical Society of America, 54*, 421–428.

Carrell, T. D., Smith, L. B., & Pisoni, D. B. (1981). Some perceptual dependencies in speeded classification of vowel color and pitch. *Perception & Psychophysics*, *29*(1), 1–10.

Childers, D. G., & Wu, K. (1991). Gender recognition from speech. Part II: Fine analysis. *The Journal of the Acoustical Society of America*, *90*(4), 1841–1856.

Clopper, C. G. (2008). Auditory free classification: Methods and analysis. *Behavior Research Methods, 40*, 575–581.

Clopper, C. G., & Pisoni, D. B. (2007). Free classification of regional dialects of American English. *Journal of Phonetics, 35*, 421–438.

Coleman, R. O. (1971). Male and female voice quality and its relationship to vowel formant frequencies. *Journal of Speech and Hearing Research, 14*, 565–577.

Colzato, L. S., Bajo, M. T., van den Wildenberg, W., Paolieri, D., Nieuwenhuis, S., La Heij, W., & Hommel, B. (2008). How does bilingualism improve executive control? A comparison of active and reactive inhibition mechanisms. *Journal of Experimental Psychology, 34*(2), 302–312.

Costa, A., Miozzo, M., & Caramazza, A. (1999). Lexical selection in bilinguals: Do words in the bilingual's two lexicons compete for selection?. *Journal of Memory and Language*, *41*(3), 365–397.

Council of Europe (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

Cox, D. R. (1970). *The Analysis of Binary Data* (2nd ed., 1989; D. R. Cox, E. J. Snell (Eds.)). London: Chapman & Hall.

Creelman, C. D. (1957). The case of unknown talker. *The Journal of the Acoustical Society of America, 29*, 655.

Cutler, A., Andics, A., & Fang, Z. (2011). Inter-dependent categorization of voices and segments. *Proceedings of the 17th International Congress of Phonetic Sciences [ICPhS 2011]* (pp. 552–555). Hong Kong: Department of Chinese, Translation and Linguistics, City University of Hong Kong.

Day, R. S., & Wood, C. C. (1972). Interactions between linguistic and nonlinguistic processing. *The Journal of the Acoustical Society of America*, *51*(79), 247–252.

Dunn, A. L., & Fox Tree, J. E. (2009). A quick, gradient bilingual dominance scale. *Bilingualism: Language and Cognition*, *12*(3), 273–289.

Eimas, P. D., Tartter, V. C, & Miller, J. L. (1981). Dependency relations during the processing of speech. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the Study of Speech* (pp. 283–309). Hillsdale, NJ: Erlbaum.

Eimas, P. D., Tartter, V. C., Miller, J. L., & Keuthen, N. J. (1978). Asymmetric dependencies in processing phonetic features. *Perception & Psychophysics*, *23*(1), 12–20.

Esposito, C. M. (2006). The effects of linguistic experience on the perception of phonation (Doctoral dissertation, University of California, Los Angeles.)

Felfoldy, G. L., & Garner, W. R. (1971). The effects on speeded classification of implicit and explicit instructions regarding redundant dimensions. *Perception & Psychophysics*, *9*(3), 289–292.

Fiszer, C. (2008). The Effect of Bilingualism on Cognition: Evidence from Early and Late Bilinguals (Master's Thesis, ENS/EHESS/P5.)

Flege, J. E., & Eefting, W. (1988). Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation. *The Journal of the Acoustical Society of America*, *83*, 729–740.

Flowers, J. H., & Garner, W. R. (1971). The effect of stimulus element redundancy on speed of discrimination as a function of state and process limitation. *Perception & Psychophysics, 9 (2)*, 158–160.

Foulkes, P. (2010) Exploring social-indexical variation: a long past but a short history. *Laboratory Phonology 1*, 5–39.

Galuske, R. A., Schlote, W., Bratzke, H., & Singer, W. (2000). Interhemispheric asymmetries of the modular structure in human temporal cortex. *Science*, *289*(5486), 1946–1949.

Ganel, T., & Goshen-Gottstein, Y. (2004). Effects of familiarity on the perceptual integrality of the identity and expression of faces: the parallel-route hypothesis revisited. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 583–597.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Lawrence Erlbaum.

Garner, W. R. (1978). Selective attention to attributes and to stimuli. *Journal of Experimental Psychology: General*.

Garner, W. R. (1983). Asymmetric interactions of stimulus dimensions in perceptual information processing. In T. J. Tighe & B. E. Shepp (Eds.), *Perception, Cognition, and Development: Interactional Analyses* (pp. 1–37). Hillsdale, NJ: Erlbaum.

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology*, *1*(3), 225–241.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition*, *19*(5), 448–458.

Goh, W. D. (2005). Talker variability and recognition memory: Instance-specific and voice-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 40–53.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological review*, *105*(2), 251–279.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 152–162.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*(02), 67–81.

Green, K. P., Tomiak, G. R., & Kuhl, P. K. (1997). The encoding of rate and talker information during phonetic perception. *Attention, Perception, & Psychophysics*, *59*(5), 675–692.

Hernández, M., Costa, A., & Humphreys, G. W. (2012). Escaping capture: Bilingualism modulates distraction from working memory. *Cognition*, *122*(1), 37–50.

Hilchey, M. D., & Klein, R. M. (2011). Are there bilingual advantages on nonlinguistic interference tasks? Implications for the plasticity of executive control processes. *Psychonomic Bulletin & Review*, *18*(4), 625–658.

James, C. T. (1975). The role of semantic information in lexical decisions. *Journal of Experimental Psychology: Human Perception and Performance, 104*, 130–136.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jerger, S., Martin, R., Pearson, D. A., & Dinh, T. (1995). Childhood hearing impairment: Auditory and linguistic interactions during multidimensional speech processing. *Journal of Speech and Hearing Research*, *38*(4), 930–948.

Jerger, S., Pearson, D. A., & Spence, M. J. (1999). Developmental course of auditory processing interactions: Garner interference and Simon interference. *Journal of Experimental Child Psychology*, *74*(1), 44–67.

Jerger, S., Pirozzolo, F., Jerger, J., Elizondo, R., Desai, S., Wright, E., & Reynosa, R. (1993). Developmental trends in the interaction between auditory and linguistic processing. *Attention, Perception, & Psychophysics*, *54*(3), 310–320.

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science, 14*, 1002–1011.

Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni and R. Remez (Eds.), *The Handbook of Speech Perception*. (pp. 363–389). Oxford: Blackwell.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, *34*(4), 485–499.

Kadlec, H., & Townsend, J. T. (1992). Signal detection analyses of dimensional interaction. In F. G. Ashby (Ed.), *Multidimensional Models of Perception and Cognition* (pp. 181–223). Hillsdale, NJ: Erlbaum.

Kaganovich, N., Francis, A. L., & Melara, R. D. (2006). Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. *Brain Research*, *1114*(1), 161–172.

Kim, M., Ackerman, L., Burchfield, L. A., Dawdy-Hesterberg, L., Luque, J., Mok, K., & Bradlow, A. R. (2013). Rate variation as a talker-specific property in bilingual talkers. In *Proceedings of Meetings on Acoustics* (Vol. 19, No. 1, p. 060235). Acoustical Society of America.

Kimchi, R., Behrmann, M., Avidan, G., & Amishav, R. (2012). Perceptual separability of featural and configural information in congenital prosopagnosia. *Cognitive Neuropsychology*, *29*(5-6), 447–463.

Kingston, J., & Macmillan, N. A. (1995). Integrality of nasalization and F1 in vowels in isolation and before oral and nasal consonants: A detection-theoretic application of the Garner paradigm. *The Journal of the Acoustical Society of America*, *97*(2), 1261–1285.

Klatt, D. H. (1989). Review of selected models of speech perception. In Marslen-Wilson, William (Ed), *Lexical Representation and Process*, (pp. 169–226). Cambridge, MA, US: The MIT Press, vii.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*(2), 820–857.

Köster, O., & Schiller, N. O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics. The International Journal of Speech, Language and the Law*, *4*, 18–28.

Köster, O., Schiller, N. O., & Künzel, H. J. (1995). The influence of native-language background on speaker recognition. In *Proceedings of the XIIIth International Congress of Phonetic Sciences [ICPhS13]* (pp. 306–309). Stockholm.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal?. *Cognitive Psychology*, *51*(2), 141–178.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15.

Krizman, J., Marian, V., Shook, A., Skoe, E., & Kraus, N. (2012). Subcortical encoding of sound is enhanced in bilinguals and relates to executive function advantages. *Proceedings of the National Academy of Sciences*, *109*(20), 7877–7881.

Kuipers, J. R., & Thierry, G. (2010). Event-related brain potentials reveal the time-course of language change detection in early bilinguals. *Neuroimage*, *50*(4), 1633–1638.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *The Journal of the Acoustical Society of America*, *59*, 675–678.

Lee, L., & Nusbaum, H. C. (1993). Processing interactions between segmental and suprasegmental information in native speakers of English and Mandarin Chinese. *Perception & Psychophysics*, *53*(2), 157–165.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*, 384–422.

Lockhead, G. R. (1972). Processing dimensional stimuli: a note. *Psychological Review*, *79*(5), 410–419.

Lockhead, G. R. (1979). Holistic versus analytic process models: A reply. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 746–755.

Logan, G. D. (1980). Attention and automaticity in Stroop and priming tasks: Theory and data. *Cognitive Psychology, 12,* 523–553.

Lorch, M. P., & Meara, P. (1995). Can people discriminate languages they don't know?. *Language Sciences*, *17*(1), 65–71.

Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition, 26,* 708 –715.

Luk, G., Green, D. W., Abutalebi, J., & Grady, C. (2011). Cognitive control for language switching in bilinguals: A quantitative metaanalysis of functional neuroimaging studies. *Language and Cognitive Processes, 27*(10), 1479–1488.

Maddox, W. T. (1992). Perceptual and decisional separability. In F. G. Ashby (Ed.), *Multidimensional Models of Perception and Cognition,* (pp. 147–180). Hillsdale, NJ: Erlbaum.

Marian, V., & Fausey, C. M. (2006). Language-dependent memory in bilingual learning. *Applied Cognitive Psychology*, *20*(8), 1025–1047.

Marian, V., & Kaushanskaya, M. (2007). Language context guides memory content. *Psychonomic Bulletin & Review*, *14*(5), 925–933.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 676–684.

Martin-Rhee, M. M., & Bialystok, E. (2008). The development of two types of inhibitory control in monolingual and bilingual children. *Bilingualism Language and Cognition*, *11*(1), 81–93.

McLennan, C. T. (2006). The time course of variability effects in the perception of spoken language: Changes across the lifespan. *Language and Speech*, *49*(1), 113–125.

Melara, R. D., & Algom, D. (2003). Driven by information: a tectonic theory of Stroop effects. *Psychological Review*, *110*(3), 422–471.

Melara, R. D., & Marks, L. E. (1990a). Dimensional interactions in language processing: investigating directions and levels of crosstalk. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 539–554.

Melara, R. D., & Marks, L. E. (1990b). Perceptual primacy of dimensions: support for a model of dimensional interaction. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(2), 398–414.

Melara, R. D., & Marks, L. E. (1990c). Processes underlying dimensional interactions: Correspondences between linguistic and nonlinguistic dimensions. *Memory & Cognition*, *18*(5), 477–495.

Miller, J. L. (1978). Interactions in processing segmental and suprasegmental features of speech. *Perception & Psychophysics*, *24*(2), 175–180.

Mitchel, A. D., & Weiss, D. J. (2010). What's in a face? Visual contributions to speech segmentation. *Language and Cognitive Processes*, *25*(4), 456–482.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology, 41,* 49–100.

Mullennix, J. W., Johnson, K. A., Topcu-Durgun, M., & Farnsworth, L. M. (1995). The perceptual representation of voice gender. *The Journal of the Acoustical Society of America*, *98*, 3080–3095.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*, 365–378.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception* & *Psychophysics, 47*, 379–390.

Muthusamy, Y. K., Jain, N., & Cole, R. A. (1994). Perceptual benchmarks for automatic language identification. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94. 1994 IEEE International Conference on* (Vol. 1, pp. I–333). IEEE.

Nakai, S., & Turk, A. E. (2011). Separability of prosodic phrase boundary and phonemic information. *The Journal of the Acoustical Society of America*, *129*(2), 966–976.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 756–766.

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by English-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, *43*(1), 1–19.

Neuhoff, J. G., Schott, S. A., Kropf, A. J., & Neuhoff, E. M. (2014). Familiarity, expertise, and change detection: Change deafness is worse in your native language. *Perception*, *43*, 219–222.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. *Speech Perception, Production and Linguistic Structure*, 113–134.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*, 42–46.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Attention, Perception, & Psychophysics*, *57*(7), 989–1001.

Nygaard L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics, 60(3),* 355–376.

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology*, *66*(2), 232–258.

Pallier, C., Cutler, A., & Sebastián-Gallés, N. (1997). Prosodic structure and phonetic processing: A cross- linguistic study. *Proceedings of Eurospeech '97, 5th European Conference on Speech Communication and Technology, 4*, 2131–2134.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 309–328.

Pastore, R. E., Ahroon, W. A., Puleo, J. S., Crimmins, D. B., Golowner, L., & Berger, R. S. (1976). Processing interaction between two dimensions of nonphonetic auditory signals. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(2), 267–276.

Pearson Education, Inc. (2008). Versant^{TM} English Test: Test Description & Validation Summary. Retrieved from http://www.versanttest.co.uk/pdf/ValidationReport.pdf.

Peirce, C.S. (1940). *The Philosophy of Peirce: Selected Writings*. Ed. J. Buchler. London.

Perani, D., Paulesu, E., Sebastián-Gallés, N.., Dupoux, E., Dehaene, S., Bettinardi, V., Cappa, S. F., Fazio, F., Mehler, J. (1998). The bilingual brain. Proficiency and age of acquisition of the second language. *Brain*, *121*(10), 1841–1852.

Perrachione, T. K., Chiao, J. Y., & Wong, P. C. M. (2010). Asymmetric cultural effects on perceptual expertise underlie an own-race bias for voices. *Cognition, 114(1)*, 42–55.

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. (2011). Human voice recognition depends on language ability. *Science*, *333*(6042), 595−595.

Perrachione, T. K., Pierrehumbert, J. B. & Wong, P. C. M. (2009). Differential neural contributions to native- and foreign-language talker identification. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1950−1960.

Perrachione, T. K., & Wong, P. C. M. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia*, *45*(8), 1899−1910.

Peters, R. W. (1955). The relative intelligibility of single-voice and multiple-voice messages under various conditions of noise. Joint Project Report, 56. U.S. Naval School of Aviation Medicine, Pensacola, Florida 1−9.

Pierrehumbert, J. B. (2002). Word-specific phonetics. In Carlos Gussenhoven, Tanya Rietvelt and Natasha Warner (eds.), *Laboratory Phonology VII*. (pp. 101−139). Berlin/New York: Mouton de Gruyter.

Pisoni, D. B. (1992). Some comments on invariance, variability and perceptual normalization in speech perception. *Proceedings of ICSLP*, 92, (pp. 587−590).

Pollack, I., Pickett, J. M., & Sumby, W. H. (1974). On the identification of speakers by voice. *Experimental Phonetics*, *26*(3), 251−258.

Ramus, F., Hauser, M. D., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science*, *288*(5464), 349−351.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265−292.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 651−666.

Repp, B. H., & Lin, H.-B. (1990). Integration of segmental and tonal information in speech perception: A cross-linguistic study. *Journal of Phonetics,* 18, 481−495.

Robinson, K., & Patterson, R. D. (1995). The stimulus duration required to identify vowels, their octave, and their pitch chroma. *The Journal of the Acoustical Society of America*, *98*, 1858−1865.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 487–494.

Salvatierra, J. L., & Rosselli, M. (2010). The effect of bilingualism and age on inhibitory control. *International Journal of Bilingualism*, *15*(1), 26–37.

Schiller, N. O., Köster, O., & Duckworth, M. (2013). The effect of removing linguistic information upon identifying speakers of a foreign language. *International Journal of Speech Language and the Law*, *4*(1), 1–17.

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(6), 1447–1469.

Singh, L., Lee, Q., & Goh, W. D. (2011). Processing Dependencies between Suprasegmental and Segmental Information: Effects of Emotion and Lexical Tone on Interference. *Proceedings of the 17th International Congress of Phonetic Sciences [ICPhS 2011]* (pp. 1858–1861). Hong Kong: Department of Chinese, Translation and Linguistics, City University of Hong Kong.

Smith, L. B., & Kemler, D. G. (1978). Levels of experienced dimensionality in children and adults. *Cognitive Psychology*, *10*(4), 502–532.

Soli, S. D. (1980). Some effects of acoustic attributes of speech on the processing of phonetic feature information. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(4), 622–638.

Soli, S. D., Wong, L. L. N. (2008). Assessment of speech intelligibility in noise with the hearing in noise test. *Intl. J. Audiology* 47, 356–361.

Sommers, M. S., & Barcroft, J. (2006). Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *The Journal of the Acoustical Society of America*, *119*(4), 2406–2416.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *The Journal of the Acoustical Society of America, 96*, 1314–1324.

Stockmal, V. (1995). Discrimination of unknown foreign languages in spoken utterances: A developmental study (Doctoral dissertation, Ohio University).

Stockmal, V., Moates, D. R., & Bond, Z. S. (2000). Same talker, different language. *Applied Psycholinguistics*, *21*(3), 383–393.

Stockmal, V., Moates, D., & Bond, Z. S. (2011). On talker voice in language identification. *Proceedings of the 17th International Congress of Phonetic Sciences [ICPhS 2011]* (pp. 1906−1909). Hong Kong: Department of Chinese, Translation and Linguistics, City University of Hong Kong.

Stockmal, V., Muljani, D., & Bond, Z. (1996). Perceptual features of unknown foreign languages as revealed by multi-dimensional scaling. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on* (Vol. 3, pp. 1748−1751). IEEE.

Stone, G. O., & Van Orden, G. C. (1993). Strategic control of processing in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(4), 744−774.

Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology*, *18*(1), 86−100.

Sullivan, K. P. H., & Schlichting, F. (2000). Speaker discrimination in a foreign language: First language environment, second language learners. *Forensic Linguistics, 7*, 95−111.

Sundara, M., Polka, L., & Molnar, M. (2008). Development of coronal stop perception: Bilingual infants keep pace with their monolingual peers. *Cognition*, *108*(1), 232−242.

Sundara, M., & Scutellaro, A. (2011). Rhythmic distance between languages affects the development of speech perception in bilingual infants. *Journal of Phonetics*, *39*(4), 505−513.

Swartz, B. L. (1992). Gender difference in voice onset time. *Perceptual and Motor Skills 75*, 983−992.

Tomiak, G. R., Mullennix, J. W., & Sawusch, J. R. (1987). Integral processing of phonemes: Evidence for a phonetic mode of perception. *The Journal of the Acoustical Society of America*, *81*(3), 755−764.

Tong, Y., Francis, A. L., & Gandour, J. T. (2008). Processing dependencies between segmental and suprasegmental features in Mandarin Chinese. *Language and Cognitive Processes*, *23*(5), 689−708.

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology*, *1*(2), 121−131.

Van Lancker, D. R., Cummings, J. L., Kreiman, J., & Dobkin, B. H. (1988). Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex, 24(2),* 195−209.

Van Lancker, D., Kreiman, J., & Wickens, T. D. (1985). Familiar voice recognition: Patterns and parameters. Part II: Recognition of rate-altered voices. *Journal of Phonetics*, *13*, 39–52.

Vicenik, C., & Sundara, M. (2009). Role of rhythmic and intonational cues in language discrimination. *The Journal of the Acoustical Society of America*, *125*(4), 2764–2764.

Vitevitch, M. S. (2003). Change deafness: The inability to detect changes in a talker's voice. *Journal of Experimental Psychology: Human Perception and Performance*, 29, 333–342.

Walden, B. E., Montgomery, A. A., Gibeily, G. J., Prosek, R. A., & Schwartz, D. M. (1978). Correlates of psychological dimensions in talker similarity. *Journal of Speech, Language, and Hearing Research*, *21*(2), 265–275.

Weiss, D. J., Gerfen, C., & Mitchel, A. D. (2009). Speech segmentation in a simulated bilingual environment: A challenge for statistical learning?. *Language Learning and Development*, *5*(1), 30–49.

Wester, M. (2012). Talker discrimination across languages. *Speech Communication, 54*, 781–790.

Winters, S. J., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *The Journal of the Acoustical Society of America*, 123, 4524–4538.

Wood, C. C. (1974). Parallel processing of auditory and phonetic information in speech discrimination. *Perception & Psychophysics*, *15*(3), 501–508.

Wood, C. C. (1975). Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analyses. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(1), 3–20.

Wood, C. C., & Day, R. S. (1975). Failure of selective attention to phonetic segments in consonant-vowel syllables. *Perception & Psychophysics*, *17*(4), 346–350.

Wong, L. L. N., Liu, S., Han, N., Huang, V. M., & Soli, S. D. (2007). Development of two versions of the Mandarin Hearing In Noise Test (MHINT). *Ear & Hearing, 28*(2 Suppl.), 70–74.