NORTHWESTERN UNIVERSITY

INTONATION THROUGH EMOTION:
EVIDENCE OF FORM AND FUNCTION IN AMERICAN ENGLISH

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Linguistics

By

Daniel Robert Turner

EVANSTON, ILLINOIS

June 2025

Abstract

Phonologically defined intonational forms are notoriously challenging to identify in context because the surface acoustic form can be highly variable depending on the context, although the sources of variation are not fully understood (Pierrehumbert 1980, Ladd 2008, among others). This problem obstructs a deeper understanding of intonation from the listener's perspective as well, for the perception of intonational contrasts and their interpretations. *Strategy.* Empirical intonation research typically strives to elicit productions of intonational categories in emotionally sterile contexts to sidestep nonlinguistic variation, but the current study takes the opposite strategy by considering speaker emotion as a case of structured variation. The unique approach tested here was to constrain acoustic variation that may blur tune distinctions by jointly considering sources of variation in phonological specification and emotional portrayal. Importantly, emotion was also formalized through adoption of a recognized analytical framework from psychology (Fontaine et al. 2007). *Experiments.* The scope of this project included four total experiments, one for production (imitating tunes while portraying emotions; Phase I), one for perception (distinguishing tunes produced in combination with emotional portrayal; Phase II), and two for interpretation (judging the meanings of tunes produced with emotion; Phase III). The research objective of considering multiple methodological perspectives was to thoroughly test the hypothesis that intonational distinctions will be enhanced by eliciting, then controlling for speaker emotion. With a clearer picture of intonational form as it relates to distinctions predicted from the phonological model, a secondary objective was to use emotional variation to better understand how listeners perceive and interpret the linguistic meaning encoded by intonation in conditions of emotional variation. *Findings.* This project found that effects of

emotion on intonation, measured in the F0 trajectories that implement phonologically specified tunes, was observed mainly in production. Listeners seemed to easily account for emotion conditioned variation, both in terms of perceptual distinctiveness and the tune-meaning associations listeners tended to endorse. The phonological specification of intonational tunes was found to be the primary driver of acoustic variation in F0 trajectories, and a strong predictor of whether listeners perceived tunes to convey contrastive meanings. This bolsters a common assumption within linguistics that the expression of linguistic content can be independent of speaker's emotion, to a large extent. Speakers tend to preserve contrasts predicted to be critical for conveying linguistic distinctions per the prevailing Autosegmental-Metrical (AM) model, within specific emotional contexts. This suggests that, despite emotion-conditioned variation in the phonetic realization of intonational contrasts, the linguistic system remains robust. Evidence from the perception and interpretation of tunes builds on these findings to further demonstrate that listeners can cope well with emotional variation in their linguistic evaluations of tunes. That said, fully half of phonologically distinct tunes were often treated as functionally interchangeable, based on findings from the interpretation phase of this study. Comparisons across *perception, production,* and *interpretation* show that perceptual discriminability and distinctiveness of tune-meaning associations tend to correlate with F0 trajectory similarity. This research fills a gap between a formal model of Mainstream American English (MAE) intonation, and by extension the AM theory on which it is based, and empirical evidence from a series of four experiments testing perception, production, and interpretation of intonational tunes. The findings broadly support the AM model. By considering how tune-emotion combinations were treated by speakers and listeners, it was concluded that despite acoustic confounds, phonological and emotional factors function separately in the production and perception of intonation.

Acknowledgments

Special thanks to the chair and defense committee. This thesis would have been impossible without the expertise, guidance, and collaboration of the chair, Prof. Jennifer S. Cole, and defense committee, Prof. Ann R. Bradlow and Prof. Matt Goldrick. Bringing together these thinkers around this research topic was an immense privilege.

Thanks to the funders behind this research, particularly the review committees:

- NSF DDRI Grant #2215338 "Intonational Cues in Emotional Contexts"
- Graduate Research Grant, Northwestern University TGS
- Additional research funding provided by Northwestern Linguistics Department

Thanks to the following people who helped enable this research:

- Chun Chan (Northwestern University Research Technology Software Developer) assisted in software development for the Auditory Free Classification experiment, and provided valuable advice about how to configure the project's data collection server.
- Fariz Firdausi, Vincci Chan, Caitlin Hubert, this project's team of Research Assistants who worked dutifully to audit the production data.
- Prof. Jeremy Steffman for assistance specifying and interpreting the GAMMs.
- Chris Serani for voice actor recruitment planning assistance.
- To the hundreds of research participants whose data contributed to this project, particularly the voice actors.

Thanks to my research participants, since data collection for this thesis was intensive and involved hundreds of participants working through experimental tasks that were all challenging in some way. The voice actors who contributed their talents were vital to laying a solid foundation for the present work and deserve appreciation.

Thanks to the following groups that provided valuable feedback:

- Audience of the 2023 Linguistic Society of America (LSA) Annual Meeting in Denver, who provided a platform and feedback for the first version of the production analysis.
- Audience of the 2024 DGfS (Deutsche Gesellschaft für Sprachwissenschaft) Annual Meeting in Bochum, who provided feedback on production and perception experiments.
- Fellow members of the Northwestern University's ProSD (Prosody and Speech Dynamics) Lab, who provided valuable stimulus and experiment piloting throughout the design process across experiments.
- Owners and moderators of CastingClubCall.com for supporting and enabling the recruitment of voice actors on their talent platform free of charge.

Thanks to my teachers… at Northwestern University, especially my Ph.D advisor and defense committee, as well as Prof. Masaya Yoshida. At University of Minnesota Duluth, especially Prof. Chongwon Park, Prof. Krista Sue-Lo Twu, and Prof. Michael Linn.

Thanks to my fellow data science consultants, data science collaborators and my assistantship supervisors Dr. Christina Maimone and Dr. Colby Wood.

Thanks to my family... especially to my wife Nicole for supporting me through this journey, my parents Michael and Ann who encouraged me to pursue truth, and my grandfathers Maurice Turner and Robert Fehner who taught me the value of science and applying a technical approach.

Table of Contents

List of Tables, Illustrations, Figures, and Graphs

Tables

Chapter 2. Production

Chapter 3. Perception

Chapter 4. Interpretation

Chapter 5. Discussion

Figures

Chapter 2. Production

Chapter 3. Perception

Chapter 4. Interpretation

Chapter 5. Discussion

Chapter 1: Motivation

1A. Defining the problem

Intonation conveys a broad spectrum of linguistic meaning (pragmatic, syntactic, etc.) primarily through pitch dynamics, which corresponds to F0 trajectories in the speech signal. A fundamental challenge for work in this area is F0 variability in production: the F0 trajectories speakers ultimately produce depend on many factors. For example, F0 dynamics can be shaped by factors that are linguistic in nature (e.g., intonational features, dialect, or segmental influence) or nonlinguistic (e.g., disfluencies, speaker emotion, or speaker's response to environmental noise), in addition to random noise arising from degrees of freedom in phonetic implementation. In order to more effectively investigate linguistic intonation, the purpose of the current project is first to isolate the information in the F0 signal that cues phonological contrasts responsible for intonational form-meaning mapping.

According to the Autosegmental-Metrical (AM) phonological model for Mainstream American English (MAE), a higher-level prosodic phrase (the Intonational Phrase, or IP) licenses the specification of an intonational tune defined by a sequence of tones consisting of (i) a pitch accent, (ii) a phrase accent, and (iii) a boundary tone (Pierrehumbert, 1980; Ladd, 2008). These phonological building blocks define the pitch pattern of the "nuclear tune" that spans from the rightmost word with phrasal prominence (the nuclear pitch-accented word) to the end of the phrase. The linguistic interpretation of nuclear tunes conveys pragmatic meaning related to information structure (focus, givenness) and the speaker's communicative intentions, such asking, telling, or floor-holding (Büring, 2016). In research on form and meaning distinctions among nuclear tunes (simply 'tunes' hereafter), researchers typically use experimental paradigms

that avoid or minimize nonlinguistic confounds such as speaker emotion, presumably to avoid introducing linguistically irrelevant acoustic variation. The current study takes a different approach by jointly modeling the effects of tune and emotion on F0 dynamics, in order to statistically model how the phonetic implementation of intonation varies if the effect of non-linguistic variation on F0 can be accounted for. A central motivation for the current study is to enrich the empirical evidence for phonetic distinctions among the phonological contrasts hypothesized in the AM model – the prevalent formalism for research on intonation in English and many other languages (Jun, 2005, 2014). Typically, studies investigating the production of these intonation patterns in MAE rely upon analyses of tunes elicited in the absence of a specified communicative context (Braun et al., 2006; Dilley & Heffner, 2021; Pierrehumbert & Steele, 1989; Tilsen et al., 2013, Cole et al., 2023; Steffman et al., 2024, among others). This leaves open the possibility that an inability to validate the phonological model using empirical data is partly a consequence of eliciting intonation outside the broader contexts that accompany natural speech. In order to illustrate the implications of this problem, next we will review the details of how the phonological model specifies phonetic targets, and apparent gaps in the empirical evidence supporting the AM model.

1B. Status of intonational phonological categories

While the primary acoustic correlate of intonation is F0, which can be continuously sampled from any voiced intervals in the speech signal, according to the AM model its phonological representation can be analyzed as a sequence of discrete High or Low (H or L) tones (Ladd, 2008). Importantly, the surface acoustic form is not a direct representation of the underlying phonological form; rather, tones are thought to encode abstract relative pitch targets. As noted, the phrase-final melody that defines the nuclear tune consists of three components: a pitch accent, phrase accent and boundary tone. Considering only the monotonal pitch accents (H*, L*) in combination with phrase accents (H-, L-) and boundary tones (H%, L%), this model generates eight tonally distinct tunes: HHH, HHL, HLH, HLL, LHH, LHL, LLH, LLL (here suppressing the diacritics marking the tone position, e.g., H*H-H%). Continuous F0 trajectories are generated via interpolation between the pitch targets of successive tones, with monotonic and non-monotonic trajectories[1]. If the AM model is correct, then the phonological contrasts represented in this 'basic' eight-tune inventory should reflect the implicit knowledge of native speakers. Yet, as discussed below, evidence from a recent study (Cole et al., 2023) fails to show the full set of predicted contrasts.

[1] A particular sequence of pitch targets might be implemented in different ways. Pierrehumbert (1980) proposed different interpolation rules based on tone (Low vs. High) and position (phrase-medial vs. final). Subsequent work suggests that nonlinearities (which appear as 'cupped' or 'domed' pitch excursions) are both perceptually salient and phonologically important. Converging evidence for the importance of these differences, which are sometimes quantified in terms of Tonal Center of Gravity (TCoG) comes from Italian (D'Imperio, 2000), German (Niebuhr, 2007), and English (Barnes *et al.,* 2012; Cole et al., 2023; Steffman et al., 2024).

Cole et al. (2023) lays the foundation for the current study's methodology in its use of pitch-resynthesized model tunes within an imitative production paradigm. In this method, participants imitate model tunes that implement AM's predicted contrasts, presented in stimuli that were created using pitch resynthesis. Participants in that study imitated the model tunes over simple target sentences presented without a specified pragmatic or emotional context. Participants heard pre-recorded model sentences like "Her name is Marilyn" where the F0 trajectory over the final word was controlled using pitch-resynthesis to represent one of the tunes of the AM model. Then, participants were asked to read aloud a new sentence (same syllable and stress structure, but different lexical items) using the intonational cues they heard. The lexical items were phonetically controlled to promote voicing, in order to obtain continuous F0 trajectories that facilitate accurate pitch tracking. This was particularly important for tune-bearing target words, which were always proper names in sentence-final position.

Cole et al. (2023) used various quantitative models to analyze differences in the time-normalized F0 trajectories produced as imitations of the model tunes, including a generalized additive mixed model (GAMM) and k-means clustering. The purpose of the GAMM is to predict variation in the F0 trajectories from the phonological tone labels of imitated tunes, while accounting for speaker-level variation. k-means clustering is a data-driven classification method that identifies distinct patterns that emerge from the F0 trajectories of the imitated tunes irrespective of the tune labels associated with individual trajectories. Overall, the results challenge the claim that the AM model adequately predicts the linguistic behavior of native speakers in producing or perceiving nuclear tunes. Based on their clustering analysis, evidence emerged supporting no more than five tunes, meaning that several of the predicted phonological contrasts were not well represented in

the speech signal. One feature that was particularly robust was the distinction between tunes that terminated with a 'high-rising' F0 excursion and tunes with other terminal trajectories, e.g., falling, fall-rise, or mid-plateau. Other types of distinctions were weak or absent, including distinctions among tunes whose imitations were grouped together in each of the two primary classes. On one hand, these results cast doubt on the ability of the AM model to capture phonological distinctions using the High/Low tone sequence model. On the other hand, the results raise the question of whether contextual factors that are present in normal communication contexts but absent in the experimental methods used by Cole et al. (2023) might be critical for the elicitation and/or recognition of some tunes in the inventory generated by the AM model. The latter possibility suggests that more tune distinctions may emerge if tunes are considered meaningful in only certain contexts, which is directly tested in the present study.

The notion that context is important for guiding the linguistic interpretation of intonation is in line with previous claims about the many-to-many mapping between tunes and pragmatic meaning (Roettger *et al.,* 2019). In such a system, context may constrain the meaning space available for the interpretation of a particular intonational melody, potentially shifting the listener's expectations about the weighting of intonational cues. For example, marking questions is a relatively well-established function of intonation, but no particular tune is uniquely assigned to this function. Rather, the polar (yes/no) question-marking function is more strongly associated with certain tunes, such as L*H-H% (Pierrehumbert & Hirschberg, 1990; Goodhue *et al.,* 2016; Gussenhoven, 2002) though the same tune might carry different linguistic meanings depending

on the context.[2] For example, while rising tunes such as L*H-H% are usually associated with polar questions, in certain instances a declarative sentence can also be produced with a rising tune in English. On a surface level, rising declaratives bear similarities to interrogative sentences with question syntax (subject-auxiliary inversion) and listeners use the context, specifically the perceived epistemic certainty of the speaker, to help interpret the intonational meaning (Jeong, 2018). According to the literature, H*L-L% is the default tune for declaratives that express assertion (Bartels & Kingston, 1994; Bartels, 1999; Pierrehumbert & Hirschberg, 1990; Goodhue *et al.,* 2016; Gussenhoven, 2002). Many other tunes are relatively unexamined regarding their meaning, such as L*H-L%, which is weakly linked with a "prompting" meaning (Bartels, 2014; Pierrehumbert & Hirschberg, 1990). Without knowing how the context of tunes affects how they are linguistically treated, in this study's view, stable tune meanings (and by extension, the categorical status of tunes) are likely to remain elusive.

[2] 'Context' in this case refers to extra-sentential information relevant to the linguistic evaluation of the speech, rather than a phonological or syntactic arrangement.

1C. Prior work on effect of speaker emotion on intonation

Peripheral to linguistics, there has been a significant effort to characterize the vocal markers of emotional speech, although this is usually done without respect to formal linguistic factors, such as pragmatics and information structure (Scherer 2013). Central to this line of research has been a methodological debate about whether empirical studies examining portrayed emotions are less valid than studies using authentic emotional experiences, for the purposes of understanding the acoustic phonetic encoding of emotion. In this literature, there are two types of emotion elicitation methods: either the researcher can "push" participants to induce an authentic emotional experience, or they can "pull" the emotion from participants by asking them to perform a version of it based on parameters that can be experimentally manipulated. For the purposes of the current study, pulled emotions seem to be appropriate because it minimizes the risk of interference between linguistic and emotional manipulations. The present analysis depends partly on the ecological validity of acoustically modeling pulled (portrayed/acted) emotions, which has been well established by emotion researchers (Scherer 2013). The core hypothesis is that speakers consciously encode emotional cues into their speech to accomplish social functions, and since they are aware of the conventionalized features of emotion, one should expect the portrayal to acoustically approximate a real emotional experience. To test this, Scherer (2013) compared a set of acoustic dimensions (F0, energy, spectrum, and speech rate) between pushed and pulled emotions.

Participants in Scherer's study were 83 adult male speakers who completed the experiment in their native language, which was either English, French, or German. The emotional possibility

space was two-dimensional, from positive/happy versus negative/sad. In the task that pushed participants into an induced emotional state, participants read target sentences that reinforced the target emotion such as, “For the rest of the day, things will go really well” for positive, versus “Life seems boring and uninteresting”, for negative, while hearing congruent music. In the task that pulled emotion from participants by asking them to act it out, the elicitation procedure for “satisfied” was used as a proxy for positive as there is no ‘happy’ within their emotion formalism, while ‘sad’ did, so it could be directly used. In other words, the elicitation method relied on systematic differences between the same sentence portrayed with satisfaction versus sadness. When participants portrayed the emotions, they were guided by short scenarios (for example, “During the holidays I built a modern desk for my apartment. Several of my guests have asked me already where I bought this beautiful piece of furniture” for positive, versus “When I moved from my old home, I had to give away my pet dog, whom I had raised and lived with for many years. My new apartment now seems very empty”, for negative). After reading the context, participants produced a target sentence, such as “At the moment I feel <satisfied/sad>” or “This is task number <digit sequence>”. Throughout the experiment, participants were also asked about their present emotional state, in order to track the effectiveness of the critical manipulation. To summarize, both the push/induction task and pull/portrayal task involve participants reading controlled materials (setting aside important linguistic differences by task) but the prior aims to shift the speaker’s mood positively or negatively (e.g. valence) before collecting the speech sample while the latter is comparable to a stage direction.

Scherer (2013) analyzed the productions in terms of F0, energy, and speech rate which were averaged across sentences and submitted to multiple measures ANOVA. They found a large and

significant effect of emotion (positive versus negative) for all acoustic dimensions, and a weak effect of task (induced versus portrayed) emerged in terms of energy. Specifically, they found that speakers in the pull/portrayal task produced more energy modulations, which was also the case for the induced positive condition, but the effect mainly depended on task. For the current study, the focus is on how F0 varies depending on elicitation method, and Scherer (2013) found a significant but small effect of task, indicating that F0, the key acoustic correlate of linguistic intonation, is minimally different depending on how emotion is elicited. This bodes well for the prospect of using emotional portrayals in conjunction with conventional experimental methods for intonation research to cross emotionally conditioned variation in F0 with controlled intonation production. In fact, a multimodal corpus based on Scherer (2013) has already been constructed, the Geneva Multimodal Emotion Portrayals (GEMEP) corpus (Bänziger et al. 2012). GEMEP is composed of 10 actors portraying 18 emotions over three types of verbal material: a declarative sentence or exclamation, a question sentence, or a sustained vowel. Trials were recorded in audio and video, and though only the audio data are discussed here, it should be noted that audio-only emotion recognition was always worse than video-only or audio and video together. The audio data was validated by 18 participants who had to (i) choose which of the tested emotions was most strongly invoked by the production ('none' was also an option) and (ii) indicate the intensity of the emotion and (iii) how believable and plausible the emotional qualities of the enactment sounded. This study is mainly interested in whether the enacted emotions were accurately classified into the intended category. Accuracy ranged from 16% (Pride) to 66% (Amusement) with a mean of 34% which shows significant perceptual confusion between emotions, although the details of the results show little confusion between emotion classes (e.g. types of Fear). Considering the size of the possibility space, this level of accuracy is

reasonable and suggests that raters are effectively perceiving and using vocal markers of emotion. In conclusion, GEMEP shows that portrayed emotions yield speech with informative acoustic cues to emotion, and as such, the elicitation method used for emotional portrayal may be suitable for investigating effects of speaker emotion on linguistic intonation with enough data to support acoustic and statistical analysis.

1D. Implications of speaker emotion for intonation

Contemporary intonation research assumes that the phonological form of an utterance is determined independently from factors like the speaker's psychological state, but this was not always the case. One early study testing the separation of intonation from the vocal cues of emotion in German is reported in Ladd *et al.* (1985). That study used synthesized speech to capture intonation patterns recognized as distinct both in terms of their phonological form and meaning function, and then asked whether F0 in these utterances could also be interpreted as a cue to the speaker's emotion. Participants in the study rated manipulated sentences based on whether they conveyed certain emotions using an eight-point scale. The results only showed a minor interaction of emotion with intonation tune, but a strong effect of pitch range, whereby tunes that were produced with a larger range were perceived to be more emotionally charged. Based on these findings, the conclusion was that listeners have independent perceptual experiences of linguistic intonation and emotion, which was taken as evidence that separate mental processes are involved, seemingly ruling out a deep connection.

There was wide disagreement, however, as around the same time the prominent intonation researcher Dwight Bolinger published a book on the meaning of intonation that emphasized the link with emotion, drawing a critical review from Ladd (1990). Ultimately Bolinger's 'dependent' hypothesis (that emotion can be integral to understanding linguistic intonation in form and meaning; they interact) was set aside in favor of Ladd's 'independent' hypothesis (that emotion is a nuisance variable that listeners handle separately from phonology; no interaction). The structure of academic departments reflects the outcome of this debate: the structure of

emotional meaning and the vocal cues of emotion are objects of interest for psychologists, while the structure of intonational meaning and the salience of phonological contrasts are studied by linguists (Ladd, 2008). Unfortunately, there has been minimal empirical validation of the 'independent' hypothesis from a linguistic perspective, which the current project aims to address.

A project similar to the present work is Mozzioconacci (1998), which tested the effect of emotions on the phonetic implementation of Dutch speech, specifically in F0. This work focused on the possibility of integrating linguistic and emotional representations in order to improve the naturalness of perceived emotions in synthesized speech without compromising the linguistic message. The core of Mozziconacci (1998) is a production study, but there is also a follow-up perception experiment reported that investigates how F0 trajectories contribute to the perception of emotion in speech. For present purposes, only the production study is directly relevant, which involved crossing a set of eight target sentences with 13 emotions (later refined to seven), as produced by three speakers. The sentences were simple statements which were presented without being situated in a context, such as "His girlfriend came by plane" and "She phoned yesterday" (as noted, the sentences were presented in Dutch). The analytical methods involved visually sorting F0 trajectories into one of 11 shape categories, then comparing the emotion-F0 trajectory correspondences.

The main finding from this analysis was that "there is no one-to-one coupling" of particular emotions and F0 trajectories. Rather, emotions and F0 trajectories appear to be in a many-to-many relationship whereby it is possible to invoke different emotions with the same F0 trajectory and vice versa. In fact, one F0 shape ("pointed hat") was found to be associated with every

emotion except ‘indignation’. From a linguistic point of view, this result is expected. It suggests that the same emotion can be realized over F0 trajectories that encode different elements of linguistic (pragmatic) meaning in a coordinated expression of linguistic meaning and emotion, such that phonological and emotional features can be simultaneously and harmoniously expressed.

Motivated by Mozziconacci’s findings, the present study seeks a deeper understanding of how phonological features are phonetically implemented by jointly modeling the emotional and linguistic contributions to F0 dynamics. That said, extending empirical intonation research by adding emotional dimensions poses a practical challenge. The present study’s approach to this issue is to formalize emotion using a psychometric model, which defines individual emotions in terms of perceived meaning using a limited number of dimensions. The use of a psychometric model facilitates cross-linguistic research because every language’s emotion words map to a universal possibility space. Ideally, emotion words would be unified in a psychological system that functions independently from the lexical system. As long as a language has emotion words that map to the same psychological possibility space, an equivalent emotion inventory could hypothetically be constructed. In other words, the key is to ground the specific emotion words in a formal system that transcends any particular language. Additionally, with emotions formalized in a psychological model, the findings from this study will be informative for emotion researchers, since considering linguistic and emotional variation together is likely to shed light on the latter, even if the focus in designing this study is the former.

1E. Psychometric models of emotion

The present study adopts the emotion model in Fontaine *et al.* (2007), which is based on psychometric dimensions that have been empirically validated on English emotion words, but which could be extendable to other words and languages. This well-cited work brings together dimensions that have been identified as important across languages, and which are shown to have acoustic correlates. One of the defining features of emotion theories is the number of dimensions used to capture relationships between emotions (Fontaine *et al.*, 2007). While other models use a greater number of dimensions—for example, Cochrane (2009) proposes eight—the one proposed by Fontaine *et al.* (2007) uses fewer dimensions which are more directly associated with F0. For the purposes of the present work, this means it is possible to systematically manipulate the most important non-linguistic dimensions that contribute (along with the linguistic context) to determining the F0 trajectories speakers produce. Fontaine *et al.*'s model uses three psychometric dimensions to represent the emotion possibility space, and the present study focuses on the two that are claimed to correlate with F0 modulations, Valence and Potency, discussed in more detail in Chapter 2. Critical for the present work, according to Fontaine et al., these dimensions have stable acoustic correlates: Valence tends to be encoded in F0, while Potency correlates with speech rate and sound pressure. This study will construct the emotion manipulation by covarying Valence and Potency using a counterbalanced set of emotions. Conveniently, the model by Fontaine et al. was normed using English emotion words, so their results can directly inform the design of our experimental materials. That said, the formalism undergirding the emotion model is not specific for English and this study's methods could be straightforwardly extended to other languages given a normed set of emotions.

Eliciting emotional speech brings methodological challenges worth considering, especially in the case of unpleasant emotions (e.g. Anger, Shame) that, if genuinely experienced by the participant, could lead to negative experimental outcomes. Rather than eliciting 'real' emotions from participants, this study relies on enacted, intentionally portrayed emotions. Enacted emotion, as conveyed through speech, has the advantage of sounding authentic while allowing experimenter control over the text. There is also evidence from Scherer (2013) that portrayed emotions are an effective proxy for naturally produced emotion in prior work. Moreover, listeners perceive portrayed emotion as authentic and identify the portrayed emotion with above-chance accuracy on the basis of audio cues alone (Bänziger et al. 2012). Based on this prior work, for the present study professional voice actors were recruited to provide an initial set of recordings for production analysis and for use as stimuli in follow-up experiments testing the perceptual distinctiveness and interpretation of tunes as a function of speaker emotion.

To preview, the core experiment tests the distinctions among intonational tunes, as defined by the prevalent phonological model (AM), using emotions that are complementary according to a psychological model of emotion. The rationale of the study is that if phonological and emotional factors interact in determining the F0 trajectory tunes, then by factoring out the emotion effect a better understanding of tune contrasts can be found. In other words, by generalizing over the phonetic realization of intonation across systematically different emotional contexts, the linguistically informative parts of the signal should emerge.

1F. Research goals and preview of methods

In order to shed light on the intersection of intonation and emotion with empirical data, a three-phase study was designed and conducted to explore the production (Phase I, the core experiment) along with perception (Phase II), and interpretation (Phase III) of tunes.

Phase I: Production

The core experiment involves intonation production across specified emotional contexts. The goal was to shed light on how the phonetic implementation of intonation varies as a function of speaker emotion. The methods are based on Cole et al. (2023), extended here with the additional condition of emotion portrayal. In line with emotion research, such as Scherer (2013), the production study was conducted with trained actors as participants, who read short text passages to establish a broader context (discourse, situational, emotional) for their imitation of an intonational tune. As a follow-up, the same experiment was also conducted with undergraduate students as participants, who were not trained actors. Participants in both cohorts listened to a recorded short utterance with a particular nuclear tune and were asked to reproduce the melody they heard on a new sentence they would read aloud from text. Participants were further instructed to imitate the heard tune while portraying a specified emotion, which was introduced in the form of a Situation that defines the discourse context, followed by a discourse Prompt (John's line in the example below), which was then followed by the participant's Response. The Response consists of a very short phrase, e.g., *"My Melanie"*, which ends in a trisyllabic proper name that will carry the target tune (HHH and HLL shown; see the example in (1) below). Each unique Response is paired with a unique combination of Situation and Prompt, chosen to

represent the distinct pragmatic meanings associated with each tune as proposed in the linguistic literature on intonational meaning (Bartels & Kingston, 1994; Burdin & Tyler, 2018; Büring, 2016; Goodhue et al., 2016; Pierrehumbert & Hirschberg, 1990; Westera et al., 2020, among others). Participants also produced the text presented in parentheses following the target phrase, which however is not analyzed here. The full set of materials consist of 32 sentences (see Appendix A) which represent every unique combination of eight tunes and four emotions.

(1) [Proud] *Situation: Jessie has been bragging about their daughter **Melanie** who is a prominent local doctor, and now an important local figure. Jessie feels proud about Melanie.*	
HHH Prompt/John:	I was reading about your daughter's award-winning medical unit.
Response/Jessie:	My **Melanie**? (She's so talented.)
HLL Prompt/John:	Who do you think will win the civic service award this year?
Response/Jessie:	My **Melanie**. (I'm almost sure.)

Acoustic measurements from the speech data are analyzed for variation that relates to the portrayed emotion and the phonologically specified tune. Here the project focuses on the analysis of F0, setting aside other acoustic correlates like intensity and tempo modulation, because the phonological model gives F0-based predictions[3]. If results indicating that the phonetic realization of an intonational tune varies according to the speaker's emotional state, it would raise questions about the factors that are important in the perception of tunes and how listeners interpret them in relation to pragmatic meaning.

[3] A preliminary analysis was conducted using F0 and non-F0 acoustic correlates, including intensity, spectral tilt, signal to noise ratio, and tempo modulation. The motivation for this limited exploration was the possibility that intonation in emotional contexts may be acoustically encoded differently than in neutral ones. However, there was no evidence that motivated a deeper analysis of F0 in conjunction with other dimensions because they tended to correlate with F0, and therefore seemed to provide little or no additional information. Consequently, the question of how intonation affects acoustics beyond F0 is a question left for future work.

Phase II: Perception

The study includes a perceptual discrimination experiment using an AX paradigm wherein participants judge pairs of stimuli (drawn from the production phase) as conveying the same or different intonation. The same method was used by Cole et al. (2023), but without considering possible variation related to speaker emotion. Stimuli for Experiment 2 were selected from the utterances elicited from voice actor participants in Experiment 1. Participants heard a pair of recordings in short succession, each containing a target phrase from Experiment 1 (e.g., "My Melanie", "For Oliver"), where each name carries one of the nuclear tunes. On a given trial, each target phrase presents a certain tune-emotion combination, and participants perform a two-alternative forced choice, responding "same" or "different". Critically, participants were instructed to base their decision on whether the speakers are "attempting to say the phrase in the same way" focusing on "how the speaker is saying the phrase, not how speakers naturally sound differently" to direct participant attention to *how* the utterance is said, and not *who* said it.

Modeling the perceptual response data will indicate to what extent listeners are able to account for the perceptual confound of emotion, when comprehending the linguistic meaning conveyed through intonation. The analysis will also compare the results with predictions from the acoustic model (Phase I), which will help to illustrate the similarities and differences between the empirical measurements and the subjective perceptual experience of listeners. In other words, Experiment 2 will gauge the magnitude of the challenge that emotional expression poses for the perception of linguistic intonation. The findings may show intonational tunes to be perceptually robust regardless of the speaker's emotional speech cues, which would support Ladd's 'independent' hypothesis (Ladd *et al.,* 1985 and Ladd, 1990). Alternatively, they may present evidence for Bolinger's 'dependent' hypothesis (Bolinger, 1986), which predicts significant

interactions between intonation and emotion. Since perception is integral to the imitation-based production paradigm, the results will also provide context for prior findings, and the following research phase.

Phase III: Interpretation

The final phase includes two experiments that ask participants about the meaning of intonational contrasts as implemented in different emotional contexts. For participants to successfully interpret the linguistic function encoded by intonational contrasts, they must be perceptually disentangled from linguistically uninformative emotional variation. In Experiment 3, the goal is to assess how an enriched linguistic context in the elicitation of tunes affects the interpretation that listeners assign to the same tunes by again making use of the speech corpus from Experiment 1. Whereas Experiment 2 tests listeners' perceptual discrimination among intonational forms, Experiment 3 has listeners sort tunes based on their linguistic function, hence its nickname for our purposes, 'Sorting'. This experiment uses the paradigm of Auditory Free Classification (AFC) and a novel implementation of the task includes design features special to this experiment in order to improve the informativity of the groups participants create. Participants were allowed up to an hour to sort 32 tune-emotion combinations into as many groups as they think are appropriate to represent the distinct meanings conveyed by the recordings. If participants are able to disentangle the linguistic from the emotional use of intonation in these stimuli, then the intonational tune should drive grouping behavior and the emergent groups should mirror the tune inventory.

In contrast to the free classification approach of Experiment 3, Experiment 4 ("Rating" task) tests how tune-emotion combinations are interpreted in relation to a small set of linguistic functions that prior research has linked to the tune inventory (Pierrehumbert & Hirschberg, 1990; Büring, 2016; Westera, 2020, among others). Participants in Experiment 4 rated how strongly an audio stimulus (implementing a tune-emotion combination drawn from Experiment. 1) invoked a particular linguistic function specified in the form of a simple text prompt (or 'probe'). The stimuli consisted of 16 tune-emotion combinations (8 tunes x 2 emotions) from one talker, the same as Experiments 2 and 3. Two complementary probes were used for each meaning dimension of interest, six total.

Probe	**Dimension**	**Direction**
1. It sounds like the speaker believes what they're saying.	*Belief*	*Positive*
2. It sounds like the speaker is asking a question.	*Question*	*Positive*
3. It sounds like the speaker wants to keep talking.	*Continue*	*Positive*
4. It sounds like the speaker doubts what they're saying.	*Belief*	*Negative*
5. It sounds like the speaker is giving a statement.	*Question*	*Negative*
6. It sounds like the speaker is finished talking.	*Continue*	*Negative*

The results were analyzed in terms of differences between complementary probes. Several possible outcomes were anticipated: (i) no difference between the Belief, Question and Continue probes, indicating a weak tune-meaning link in these dimensions; (ii) many positive and few negative ratings for a particular tune-probe mapping, indicating a strong link for the positive probe, or (iii) many negative and few positive ratings, also indicating a strong link for the negative probe. Strong alignment is expected between the results of Experiment. 3 (Sorting/AFC) and Experiment. 4 (Rating) since they both use the same stimulus files and ask

participants to judge the speaker's intended meaning while ignoring emotion. If this holds true, then it sheds light on how linguistic meaning remains robust even when acoustic distinctiveness is complicated by nonlinguistic sources of variation.

1G. Scope and strategy

The scope of the study is deliberately broad (production, perception, interpretation) in order to advance our scientific understanding of this under-researched area to enable future research that jointly considers linguistic intonation and emotional prosody in the speech signal. Numerous pilot and follow-up experiments were conducted to arrive at the final design of each part of the study, but due to the novelty of some of the tasks (in terms of task design generally and as tasks were applied to intonation) the nature of the present work is exploratory. The work is grounded in formal models of both intonation and emotion, which is expected to lead to a finer-grained analysis of each, although the effect of emotion is only methodologically relevant to the linguistically-motivated research goals. If the strategy of manipulating and modeling intonational and emotional variables together is successful, it could help pave the way for the consideration of other linguistic and nonlinguistic contrasts that may also inform the speech signal in terms of production, perception, and interpretation.

Chapter 2: Production

2A. Introduction

Intonation refers to patterns of variation in pitch, tempo and other acoustic parameters in structures that unfold over phrases and utterances, often multiple words, which convey pragmatic meaning and discourse information. Much of the recent phonological research into English intonation adopts the Autosegmental Metrical (AM) model, which has fruitfully guided the field in many ways (Pierrehumbert 1980, Ladd 2008). Despite the popularity and successes of the AM model, few attempts have been made to directly test whether it generates valid predictions for the phonetic implementation of basic phonological contrasts. As discussed in Chapter 1, Cole *et al*. (2023) recently showed that participants imitating stimuli encoding predicted intonational differences from the AM model did not consistently reproduce all the critical distinctions represented in the tonal specification of the model. That study tested a set of eight AM-generated forms (called 'tunes') that varied in their underlying tonal specification (High or Low), which AM posits as the building blocks of intonation. Converging results from quantitative analysis of F0 trajectories and perceptual discrimination failed to confirm all the predicted differences. From a set of stimuli representing eight intonational tunes, they found evidence for only a five-way tune distinction, which suggests either a problem with the AM model for English (it allows too many tonally distinct tunes), or a problem in how the predictions from that model were empirically evaluated. The current study applies novel experimental methods to tease apart these possibilities.

While the design of the present study is based on Cole *et al*. (2023), it is unique in the use of speaker enacted emotion as an additional experimental variable, which was constructed using a well-cited model adopted from psychology (Fontaine *et al*. 2007). The objective is to reduce variation in the speech signal by invoking, then factoring out, emotion-conditioned variation. Next, the theoretical underpinnings for intonational phonology and emotion will be considered in further detail.

2B. Background

Phonological model

As discussed in Chapter 1, currently the dominant phonological theory for analyzing Mainstream American English (MAE) is the Autosegmental-Metrical (AM) model (Pierrehumbert, 1980; Ladd, 2008). At issue in the present study is the lack of empirical evidence to support fundamental contrasts that AM predicts, which was most recently demonstrated by Cole *et al*. (2023). Their key finding is that speakers failed to consistently produce distinct F0 trajectories for each of the unique phonological tunes of the model. The present study asks whether the missing distinctions predicted by the AM model emerge when considered in relation to speaker emotion, which is not encoded in the phonological specification. Of interest is whether all the predicted tune distinctions are observed within an individual emotion context, or whether the full set of predicted distinctions emerges when emotion contexts are aggregated. The analytical methods are consequently designed to evaluate whether speakers produce acoustically distinct forms for phonologically distinct F0 trajectories.

Acoustic distinctiveness of AM model-generated tunes

In order to test the relative distinctiveness of intonational tunes, Cole *et al*. (2023) generated an inventory of F0 trajectories that instantiate a subset of the phonologically contrastive tunes in the AM model. Specifically, they varied the tonal specification of each phonological position, consisting of a pitch accent, phrase accent and boundary tone, each of which may be a single High (H) or Low (L) tone, yielding eight phonologically distinct sequences: HHH, HHL, HLH,

HLL, LHH, LHL, LLH, and LLL. Cole *et al*. (2023) tested whether these phonologically contrastive tunes were reproduced as phonetically distinct (which would constitute evidence for the AM model) or not (providing only partial support for the model). A production study provided the data for their analysis, involving participants imitating F0-resynthesized versions of each tune, which were designed to exemplify the critical differences between tunes, based on descriptions and examples in Pierrehumbert (1980) and Veilleux *et al*. (2006). Each tune was presented over a three-syllable name (offering distinct syllables as anchors for the pitch accent, phrase accent and boundary tone) which was the last word in a short sentence ("Her name is Marilyn").

The experimental task involved participants hearing the stimuli then reading aloud a new sentence with different lexical items (but same the same syllable and stress structure) while imitating the heard "melody". Stimuli were presented out-of-the-blue without a specified discourse or situational context, nor were they designed to invoke specific linguistic functions associated with intonation. Participants repeated the same sentences multiple times for each tune, and from those recordings Cole *et al*. (2023) extracted time-normalized F0 trajectories. Several methods were used to analyze the F0 trajectories for distinctions among imitations of the eight tunes, including k-means clustering and a generalized additive mixed model (GAMM). The purpose of the GAMM was to evaluate whether and to what extent variation in the imitated F0 trajectories could be predicted from the tune label of the stimulus that was imitated. The k-means clustering analysis served to identify the optimal partition over the set of imitated F0 trajectories, where individual trajectories were grouped together based on their mutual similarity over the entire F0 trajectory, irrespective of the tone label of the tune that was imitated on each trial.

Overall, Cole *et al*. (2023) found incomplete evidence for the hypothesized set of tune contrasts that were tested, such that the phonological predictions were not fully supported by the empirical results. The k-means clustering analysis showed five emergent clusters of F0 trajectories that differed in shape, suggesting a reduction in contrast from the eight AM model-generated tunes. In other words, over the course of the experiment, although participants heard stimuli with an eight-way distinction in F0 trajectories, corresponding to the eight tonally specified tunes, the F0 trajectories produced in imitation of those eight tunes support only a five-way distinction. In the clustering analysis, three of the emergent clusters were composed of imitated productions of two input tunes, with >90% of the imitations of each tune assigned to the merged cluster: HLL-HLH, LHL-LLH, and HHH-HHL. Two other clusters emerged, each composed of nearly all the imitations of just a single tune: LHH and LLL. Based on converging evidence across analyses of the production data (clustering, GAMM, machine classification) and perception data (AX discrimination), Cole *et al*. (2023) claims that tunes are hierarchically related, with a high-level, robust distinction between "high-rising" tunes that rise monotonically and terminate with a high final F0 value (mainly imitations of HHH and HHL) vs. "non-high-rising" tunes (mainly imitations of all other tunes), and smaller, less reliable distinctions among tunes within each of those classes. Their GAMM analysis showed fine grained differences between tunes that clustered together, usually in the final region of the F0 trajectory, and offered further support for a major distinction between high-rising and non-high-rising tunes.

These results are problematic for the AM model, as they fail to support the predicted eight-way distinction, and because the AM model does not predict a larger, more robust and more reliable distinction for high-rising tunes than for other, non-high-rising tunes. That said, it is possible that the predicted phonological contrasts fail to emerge in Cole *et al*.'s data due to the

lack of a supporting context, which when present, might critically license additional tune distinctions and provide better evidence for the predicted eight-way tune distinction. The current project extends their general design by specifying a discourse context for each tune presented as stimuli, and by using speaker emotion to further constrain acoustic variation in F0.

Emotion model

Intonation research conventionally avoids invoking emotion probably under the valid concern that doing so would introduce additional uninformative variation ("noise"). However, the aim of the present study is to identify emotion-related variation from the observations of the phonetic implementation of tunes in order to separate F0 variation due to emotion from variation related to phonological tune distinctions. This project adopts a well-cited emotion model from psychology that conceptualizes all emotions as a combination of a small number of psychometric dimensions which are theorized to be shared across languages and cultures (Fontaine *et al*., 2007). Hence, one of the key advantages of adopting this emotion model is the ability to adapt the procedure beyond English in the future. The current study focuses on the two dimensions that have been shown to share acoustic correlates with linguistic intonation, particularly F0 (Fontaine *et al*., 2007). The first dimension of interest is evaluation-pleasantness which is characterized by "appraisals of intrinsic pleasantness and goal conduciveness, as well as action tendencies of approach versus avoidance or moving against" (Fontaine *et al*., 2007: 1051), hereafter called 'Valence', in line with conventions in the emotion research field (Shuman *et al*., 2013). The other dimension is potency-control (hereafter 'Potency') which is characterized by "appraisals of control, leading to feelings of power or weakness; interpersonal dominance or submission,

including impulses to act or refrain from action" (Fontaine *et al*., 2007: 1051). Important for the present work, according to Fontaine *et al*.'s model, these dimensions have stable acoustic correlates: Valence tends to be encoded in F0, while Potency mainly correlates with speech rate and energy. The present study restricts its consideration of the many acoustic correlates of intonation to F0, so it is possible that the effect of emotion will be mainly driven by Valence. That said, differences in energy and speech rate also tend to manifest in F0, so Potency may correlate with F0 vis-à-vis its correlations with other acoustic dimensions. Specifically, in terms of energy, a greater volume of airflow is required across the glottis to produce High versus Low tones, and in terms of speech rate, a longer duration is often observed for High versus low tones (Arvaniti, 2020). Therefore, the F0-based approach taken by this study is poised[4] to capture related modulations in other acoustic dimensions that may also vary in relation to emotion. The emotion manipulation was constructed by covarying Valence and Potency using a counterbalanced set of emotions, the details of which are discussed in the Methods section.

[4] The question of whether F0 trajectories alone are sufficient to characterize the phonetic implementation of tunes is directly addressed by follow up experiments in perception and pragmatic interpretation, which should (and do) provide results which correlate with trends in F0 trajectories.

2C. Hypotheses & predictions

The interplay between phonological structure and phonetic realization in speech is a complex and ongoing area of investigation. Determining the degree to which phonetic implementation of tunes aligns with phonological expectations, and how speaker emotion influences this alignment, presents a significant challenge for current research. This study aims to clarify the relationship between phonologically specified tunes and F0 trajectory shape through direct investigation of the joint production of tunes and vocal expression of speaker emotion. To formally examine this relationship, the study introduces Hypothesis 1, which posits that the joint production of phonological specifications and emotional expression yields patterns of variation that ultimately preserve phonological distinctions. This predicts that speakers may adapt their imitation of tunes to maintain tune distinctiveness in terms of F0 shape within a given emotion. A finding that distinctions among phonologically specified tunes in their F0 trajectories are diminished (e.g. masked) by emotional variation, such that tunes lack distinctive F0 shapes, would lead to the rejection of Hypothesis 1[5].

Hypothesis 1: As jointly produced with emotional expression, phonological distinctions among tunes will be reflected in distinctions among the corresponding F0 trajectories.

Building on Hypothesis 1, Hypothesis 2 further posits that speakers encode similar degrees of tune distinctiveness regardless of the specific emotion, including Neutral.

[5] This finding seems improbable, since it implies that expression of speaker emotion imperils linguistic meaning conveyed by intonational contrasts. But since speakers are presumably always conveying an emotional state, it would be surprising counterevidence for Hypothesis 1 if tune distinctions in F0 were found to be masked by emotional variation.

Hypothesis 2: Distinctions among F0 trajectories for phonologically different tunes will be of similar magnitude across speaker emotions.

Hypothesis 2 predicts that the F0 shapes corresponding to tunes will not be systematically more or less distinct in any given speaker emotion. For example, the same number of tune distinctions are expected to emerge under Neutral as under Love, Pride, Shame or Anger. A finding that the F0 shapes of tunes are more distinct under Neutral than under a given emotion (e.g., if emotional contexts constrain F0 shape variation) or a finding that F0 trajectories or less distinct under Neutral (e.g., if unconstrained context leads to more variation in tune F0 shape) would constitute evidence against Hypothesis 2.

Interim summary

This study seeks to answer basic questions about how intonational tunes and vocal cues of emotion are expressed together in F0 trajectories, based on the idea that the general properties of tunes can be better understood if the concurrent emotional variation can be accounted for. By jointly considering how F0 varies in relation to tunes and emotions, this study tests whether it is possible to separate the contributions from emotion, thereby sharpening the view of the F0-based properties that encode phonological distinctions. If the findings from this study show that F0 variation due to emotion can be isolated from variation due to linguistic factors, it opens the door to exploring how listeners cope with variation and the form-meaning mapping of intonation. Alternatively, if the findings fail to disentangle F0 variation related to emotion vs. tune, it hints at a deeper connection between linguistic intonation and emotional expression in F0 and possibly challenges fundamental assumptions about the distinction between linguistic and “paralinguistic”

prosody (Gussenhoven 2004; Ladd 2008). Is emotion effectively a surface-level phenomenon that can be set aside for the purpose of understanding intonation, or will evidence of intonational dependencies with emotion emerge?

2D. Experimental methods

The experimental and statistical procedures for Experiment 1 were adapted from the imitative production task, materials, and analysis detailed by Cole et al. (2023), with three key modifications introduced to address the need to effectively elicit an emotion portrayal:

i. Tunes are elicited in the context of a rich dialog with a virtual conversation partner, which provides a pragmatic context that is congruent with the meaning associated with the tune, based on existing accounts of intonational meaning in English (Búring 2016, Westera 2020, Pierrehumbert & Hirschberg, 1990).

ii. The inclusion of visual and dialogue cues reinforced the emotion being elicited while imitating a particular tune, as detailed below.

iii. Voice actors with professional training and experience are recruited as participants, along with untrained speakers. The anticipated benefits of specialized training and experience includes exemplary vocal control and the ability to produce natural-sounding portrayals of emotions with minimal prompting.

The remainder of this section details how, and with what materials, the experiment was conducted. All of the materials, including audio (model tunes, raw recordings) and experiment files (executables, stimulus lists), are freely available through the study's Open Science Framework repository (https://osf.io/gbk8z/, see Experiment #1).

Tune inventory

The tune inventory parallels Cole *et al*. (2023) with eight phonologically distinct tone sequences: HHH, HHL, HLH, HLL, LHH, LHL, LLH, LLL, each being composed of a sequence of pitch accent, phrase accent and boundary tone (ToBI diacritics suppressed here, as noted in Chapter 1). This set represents the phonological tune contrasts predicted by the AM model (Pierrehumbert 1980, Ladd 2008), setting aside bitonal pitch accents and downstepped High tones. For these tunes to convey distinct linguistic functions, the critical features that encode the phonological tone sequence (an F0-based contrast) should be salient in the speech signal. As discussed in Chapter 1, there is a rich literature on the pragmatic meaning associated with some tunes, e.g., HLL encodes assertion and LHH encodes question, while other tunes have received less (or no) attention for their meaning function. Consequently, the contrastive status of tunes based on pragmatic function remains untested and therefore uncertain.

Consistent with Cole *et al*. (2023), the present study considers each phonologically specified tune as a *potential* category that is available for association with linguistic (pragmatic) meaning, and evidence that the predicted categories are phonetically distinct is examined in F0 trajectories. If the AM model accurately predicts the tune inventory distinctions for MAE, then speakers of that variety should perceive each phonologically specified tune as distinct from others, based on its characteristic F0 trajectory. Further, the same tune will be faithfully reproduced with an F0 trajectory of the same shape, in the production of a novel sentence. Conversely, and as reported in Cole *et al*. (2023), if certain tunes across speakers are not distinct from one or more other tunes in production, it would raise questions about the underlying tune inventory, with possible implications for the phonological characterization of the nuclear tune.

Resynthesized model tunes

As an imitation-based experiment, it is critical that the audio stimuli adequately guide participants in the phonetic implementation of phonologically conditioned F0 contrasts. Because the study design crosses intonational tunes with emotions without a specific hypothesis about how particular combinations might be produced, tunes were always represented by the same audio stimuli regardless of the emotion being portrayed. Specifically, the present study adopts a subset of the model tune audio files developed and used by Cole *et al*. (2023) for the same purposes, giving the present study the chance to replicate their results with (some of) the same materials. These model tunes were based on schematic and real examples from the AM literature (Pierrehumbert 1980; Veilleux *et al*. 2006*)* and were implemented as model tunes on short sentences through a process of pitch resynthesis using Praat (Boersma & Weenink, 2001). Each model tune was resynthesized with the same sentence, “Her name is Marilyn”, using source recordings produced by a pair of native English speakers (female and male). Critically, the F0 of the final tune-bearing word (‘Marilyn’) was resynthesized to capture the intended F0 contour shape for each tune, adjusted relative to the speaker’s mean (see Figure 1: Audio stimuli F0 trajectories). The set of F0 targets used for resynthesis are given in Appendix 2A.

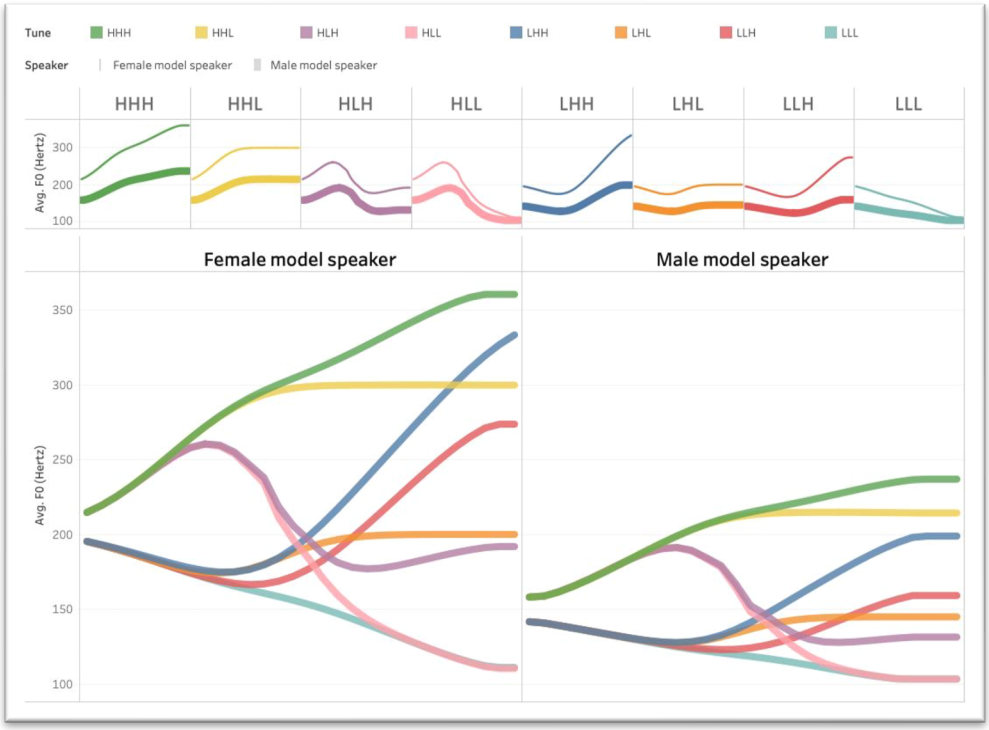


Figure 1: Audio stimuli F0 trajectories organized by tune (top) and speaker (bottom), as estimated from VoiceSauce (Shue et al., 2009).

Emotion inventory

As mentioned in Chapter 1, the organization of the adopted emotion model is based on emotions being defined in terms of multiple psychometric dimensions, two of which (Valence and Potency) are known to impact the acoustic correlates of intonation, including F0 (Fontaine *et al.* 2007). To serve the design of the experimental manipulation, a set of emotion words which systematically vary Valence and Potency was identified from the normed English emotion terms from the model (see Figure 2: Emotion possibility space). Specifically, the goal was to identify four emotion words to represent emotions representing extreme positive and negative values along both axes, which was done by inspecting their relative values. Selection of the final set of emotions was also driven by how well they worked in combination with the attested pragmatic

meanings for the tunes according to the literature. The process was iterative and involved soliciting and integrating feedback from members of the Prosody and Speech Dynamics Lab at Northwestern University. This step helped to identify and exclude emotions that were judged to be less compatible with the pragmatic function associated with certain tunes, such that their combination could be irreconcilable for speakers, leading to unpredictable results. Because the goal was not to maximally explore the emotion possibility space, the project focused on emotions that were judged by the author to be compatible with the greatest variety of linguistic meanings. For instance, combining the canonical meaning of the tune LHH, 'seeking information', with the meaning of Contempt seemed more challenging than pairing it with Anger, which led to selecting Anger over Contempt as the emotion with maximum Valence and minimum Potency. The emotion inventory consists of: Pride [+Valence, +Potency], Love [+Valence, -Potency], Anger [-Valence, -Potency], and Shame [-Valence, +Potency][6]. In principle, a different set of emotion terms could convey the same dimensional relationships with similar acoustic effects, (e.g. Joy, Surprise, Disappointment, Irritation) but this was not tested.

[6] Emotions in the model are given continuous values, but in the context of the current study they represent extreme values along both dimensions, hence the simplified polar coding.

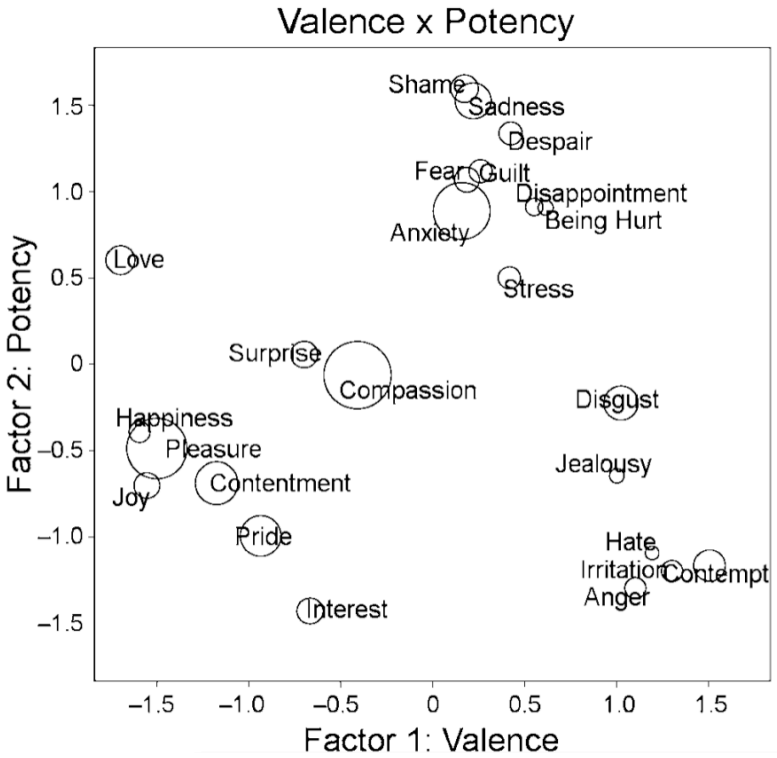


Figure 2: Emotion model dimensions of interest (Fontaine et al. 2007)

Emoticons/Emoji

When participants were asked to portray a given emotion, it was displayed as text flanked by two identical corresponding emoticons. Emoticons, also popularly known as emoji, are ideogrammatic icons that depict a certain emotion with a simple cartoon face, which served as a non-linguistic cue that did not have to be read, and therefore should minimally interfere with cognitive resources needed for linguistic perception and production, such as verbal working memory. Participants first see the emoticons in the experiment instructions, so they are familiar with the emotion-emoticon correspondences before trials are presented. The set of emoticons was primarily developed through consultation with other members of our intonation research

group: Pride [🤩], Love [🥰], Anger [😤], Shame [😰][7]. Since the emoticons are always accompanied by the target emotion spelled out, ambiguity about their intended meaning is unlikely to arise, therefore alternative sets were not independently vetted. In addition to the four emoticons representing the emotion inventory, the experiment uses an emoticon to represent the participant's virtual interlocutor, 'Bill', represented by [👨]. The purpose for representing Bill via emoticon was to promote an implementation of tune-emotion combinations that maintains the critical intonational features a real interlocutor would need in order to understand the tune. Additionally, it was thought that if the speaker had an interlocutor in mind, that their vocal cues to emotion might be produced in a more personal and thus naturalistic way.

Written materials

With the goal of encouraging participants to draw upon their implicit intonational knowledge each tune-emotion combination was elicited in a unique semantic-pragmatic frame that was congruent with the tune's attested (or for some tunes, claimed) linguistic function. For example, the canonical questioning tune (LHH) was elicited using an interrogative sentence ('My Melanie?'). This means that, in addition to attending to the compatibility of tunes and emotions, the design considered how the semantic-pragmatic frame of the tune-emotion combination implies illocutionary force (i.e., speech acts). The study did not evaluate alternative written materials for tune-emotion combinations. Text materials fall into three categories: the preceding dialog (read silently), the target sentence (read aloud with a target tune and emotion), and a

[7] The set of emoticons used is published by Microsoft and is widely used on mobile devices and online and are thus going to be familiar to many participants.

continuation sentence. The continuation was meant to put distance between the end of the target word and the end of the trial, since participants might encode words at the very end of trials differently, e.g., with lower amplitude or creaky voice. Participants were instructed to maintain their emotional portrayal through the continuation, though the intonation of the continuation was neither specified nor analyzed. See Appendix 2B for the full set of written materials.

Pragmatic constraints

This project was guided by the literature on intonational meaning in the construction of materials, to elicit tunes in a congruent dialog context, in scenarios where speakers could leverage their implicit linguistic knowledge (see Table 1: Attested tune meanings). Two core challenges arose in the creation of the text materials. First, the intonational meaning literature provides uneven (and sometimes contradictory) information about tune meaning for the tunes being investigated; certain tunes have a canonical meaning established over decades of research, while the meanings of other tunes are under-researched. Second, tunes in the pragmatic meaning literature are often characterized in simple terms (e.g., “rising” or “falling”), or in terms of phonological tone sequences, leaving open questions about the F0 trajectory associated with a given tune. To help identify potential issues, the materials were developed in close consultation with members of the Northwestern Prosody & Speech Dynamics Lab, such that each tune was assigned a unique function congruent with the group intuition and the literature. The following table synthesizes the attested meanings of each tune based on prior research with specified tunes. Note that participants were not instructed or queried about particular tune interpretations and

they were principally elicited using an imitation paradigm which (as mentioned) with model tunes drawn from Cole *et al*. (2023).

Table 1: Attested tune meanings

Tune	Linguistic meaning/function	Source*
HHH	Questioning, possibly when the answer is believed to be positive	PH, Je
HHL	Elaborating on something that's been previously mentioned	PH, Ba
HLH	Non-finality, uncertainty, selecting the addressee	PH, Gu, WH
HLL	Declarative, asserting information, possibly incomplete	PH, Go, Gu, Ba
LHH	Questioning, typical for polar questions, or incredulity	PH, Ho, Je, Gu
LLH	Speaker believes listener should already know this information	PH, BT
LHL	Prompting for the speaker to respond, possibly as a reminder	PH, Ba
LLL	Finality, non-predication	PH

*Key to citations: Ba = Bartels (1994); BT = Burdin and Tyler (2018); Go = Goodhue *et al*. (2016); Gu = Gussenhoven (2002); He = Heim (2019); Je = Jeong (2018); PH = Pierrehumbert & Hirschberg (1990); WH = Ward & Hirschberg (1985)

Rather than eliciting the tunes through meaning-based prompts, participants were presented with two auditory models of a particular tune on each trial and were tasked with reproducing the perceived melody over a new sentence, presented in a dialog context. The point of crafting compatible semantic-pragmatic contexts was to encourage participants to recruit their implicit intonational knowledge of how tunes in a particular pragmatic context should be produced, with the aim of maintaining linguistic informativity despite the presence of other salient sources of variation in the speech signal.

Phonetic constraints. The tune-bearing words (Melanie, Marilyn, Lavender, or Gallagher) were chosen because of their phonetic properties and metrical structure (syllable count, stress pattern, etc.) which facilitates F0 extraction and analysis. To this end, tune-bearing words were selected to minimize unvoiced segments, favoring phones like [d]/[b]/[g] over [t]/[p]/[k], and were uniform in terms of syllable count and stress pattern. The tune-bearing (target) word was always the final word in a short two-word sentence, preceded by an unstressed function word that was not analyzed ('about', 'also', 'and', 'for', 'from', 'it', 'just', 'my', 'not', 'our', 'that', 'other', and 'with').

Participants

Participants were recruited in two groups: 13 voice actors ('VA'; 7 female, 6 male; mean age 26.5 years) from an industry crowdsourcing platform (details below) and 19 untrained speakers from the Linguistics subject pool ('SP'; 17 female, 2 male; mean age 19.5 years). Most of the data (4,764 recordings) were collected from VAs; although fewer in number than the SP participants, the VAs repeated trials five times (200 total trials), versus three repetitions for SP participants (120 total trials). On average, VAs took 61 minutes to finish the study (Min 33, Max 80), compared to 30 minutes (Min 25, Max 35) for SP participants. VAs received monetary compensation for their participation while SP participants, undergraduate students in Northwestern University introductory level linguistics classes who were enrolled through the Linguistics Subject Pool, earned course credit for their participation.

Inclusion criteria: All participants were native speakers of MAE who spoke English at least 90% of the time (based on self-report), were between the ages of 18-65, and reported no speech or hearing deficits. Speakers also had to have native proficiency in the Northern Cities or Midlands dialects of American English based on self-report of regional residential history, to reduce dialectical variation. Specifically, participants were recruited who spent the majority of their early life in the region defined by the following states: Michigan, Minnesota, Missouri, New Jersey, Illinois, Indiana, Ohio, Pennsylvania, and Wisconsin. Individuals with speaking proficiency in another language (e.g., bilinguals) were included if they currently spoke their other language less than 10% of the time. This was done to recruit participants with ample experience with American English as spoken in the dialect region of interest, so they are drawing from a common well of linguistic experience, which should foster consistent results.

Voice actors (VAs) These participants were recruited through a popular industry job posting platform specifically for freelance voice actors, CastingCallClub.com (CCC). The projects on CCC typically relate to video games, audio books, film, and TV voiceovers; ours was the first research study to use the platform according to CCC staff, who were consulted prior to launching the project[8]. After the experiment session, VAs who completed all parts of the experiment were compensated for their time via $40 electronic transfer.

[8] The project page, titled "Linguistics Study", included a description of the experiment and inclusion criteria, and interested VAs were invited to audition through a built-in feature of the website. The norm on CCC is for members to submit an audition in the form of sample recording, which helped to exclude participants with inadequate equipment and recording practices. Unlike other crowdsourcing websites, CCC's primary focus is professional-quality remote VA work, so members typically own studio-quality equipment. That said, one piece of equipment not available to VAs is a sound-attenuated recording space, and background noise was a problem in some cases. Fortunately, in most cases background noise, when present, did not affect F0 measurement, but to facilitate the

Subject pool. As mentioned, the methodological basis for the present study is Cole *et al*. (2023), which was conducted using the Northwestern University subject pool; therefore, additional data was collected from the same student population. With the addition of subject pool data, the analysis of emotional variation does not wholly rest on the conventionalized emotional cues that voice actors recruit. Another difference between participant cohorts is that SP participants were not financially compensated, unlike VAs, but this is expected to be congruent with each group's expectations, and SPs were not aware that another group was compensated differently. Drawing contrasts between the participant cohorts was not a research goal of the current project, since the current project focuses on probing the common linguistic knowledge of native speakers, which is expected to be similar between these groups.

Procedure

The experiment was implemented using OpenSesame (Mathôt *et al*. 2012), which is open-source software that enabled the implementation of the same procedure online for remote participants (VAs)—using a virtual private server running JATOS software (Lange *et al*. 2015)—and locally in the Northwestern University Phonetics Laboratory (SPs)[9]. The laboratory was equipped with a Shure WH-20 dynamic head-mounted microphone, ensuring a constant distance even if the

future analysis of acoustic dimensions like amplitude and spectral tilt, only recordings that were suitable for a broader acoustic analysis were included in the data presented here.

[9] SP data was collected while we were strictly observing university-wide COVID-19 precautions which limited the time participants could spend in the recording spaces, required the sanitizing of all surfaces in the recording booth, as well as the use of compliant face masks. SPs removed their face mask for the duration of the experiment, as it would have interfered with recording, while the experimenter's face (the first author) remained masked at all times.

participant moves, which was anticipated might occur with the production of strong emotional cues (‘lively speech’). Audio routing was managed through a dedicated digital interface (MOTU M4) set to the sample rate of 44,100 Hertz, which was captured in WAV format. For VAs, who provided their own recording equipment, WAV files were captured through the experimental software at the same sample rate. Unfortunately, some participants lacked the necessary internet connection bandwidth to support uncompressed audio streaming, which resulted in the accelerated introduction of audio artifacts. The F0 extraction and preprocessing script includes provisions to detect and repair audio artifacts that would affect the quantitative analysis, which is detailed in the statistical methods below.

Instructions. The instructions opened with a brief explanation of the goals of the experiment, which was given in terms of crossing a set of intonational tunes with a set of emotions. The concept of an intonational tune was explained using canonical tune-meaning associations that are easily represented in text and punctuation: questions versus statements. The instructions then explained that tunes and emotions are fundamentally different, yet both influence what is eventually produced. Participants were told that the most important part of the experiment was faithfully recreating the perceived F0 trajectories presented in the model tunes (the critical audio stimuli). Recall that the model tunes were created with the intention of not invoking any specific emotion context, so in the context of the present study they are considered emotionally neutral. During the instructions, participants hear all model tunes played in a random order, which serves to familiarize them with the intonational forms they will be imitating. There was no equivalent

training or guidance on how emotions were expected to be produced as participants were assumed to be familiar with them. That said, the design of the text materials was meant to unambiguously cue the target emotion (in addition to other cues) using a cast of characters that evoke specific emotions. The characters were introduced to the participant at the start of the experiment session, as show below (emotion labels were not given):

i. Pride: Madelyn is your character's adult daughter, and a prominent local physician.
ii. Shame: Lavender is your troublesome pet dog who embarrasses you in public.
iii. Love: Melanie is your teenage granddaughter and you two often cook and bake together.
iv. Anger: Gallagher is your penny-pinching landlord, who is dodging your calls (and *their* bills).

Using a specific character for each emotion allowed the creation of an internally consistent world to motivate the linguistic and nonlinguistic cues bring elicited, and this way each name itself (Madelyn, Lavender, Melanie, Gallagher) is associated with a unique emotion, reinforcing the cues from the emoticons and the orthographically presented emotion word. An additional character is the virtual interlocutor, Bill, to whom the preceding context is attributed, and to whom the participant is responding with the target sentence.

Neutral elicitations. After the instructions, participants completed a trial block without a specified emotion, called Neutral for the purposes of the present study although it may appropriately be considered an "unspecified" emotion condition. Since this is the type of elicitation used in Cole *et al*. (2023) it is expected that Neutral elicitations in the present study will most closely resemble results from that study. Neutral trials proceed in a self-paced manner

that begins with the participant silently reading the target sentence, which only for the Neutral trials was the same sentence as the model stimulus, presented with no supporting discourse context. After reading the text of the target/model sentence, the participant clicks the mouse to initiate playback of the model tune as presented in two recordings, always in the order: female, male. Immediately after the last model tune, a beep plays to indicate the microphone is recording, then three seconds later another beep plays signaling the end of recording. Trials were separated by a short (~five second) interstimulus interval.

Emotion elicitations. After giving their tune imitations in a Neutral context, participants began the main part of the study, which involved combining vocal cues of emotion with the tune productions. These trials followed a four-step process (see Figure 3) of tune-emotion elicitation, as follows. (1) Trials began with the participant silently reading the virtual interlocutor's prompt ("I received a phone call from your favorite granddaughter on my birthday."). When ready to continue, the participant clicked the mouse, then (2) two audio models of the specified tune played, always with the sentence "Her name is Marilyn". As the last model played, the specified emotion word and target sentence ("From Madeline?", with Love) appeared together in the center of the screen. Then, (3) a beep played to indicate microphone activation and the participant gives their emotion-imbued imitation of the tune over the target sentence, before a second beep marked the end of recording. After the target sentence, (4) the continuation appears, which the participant produces with the same emotion but with an intonational expression of their choosing.

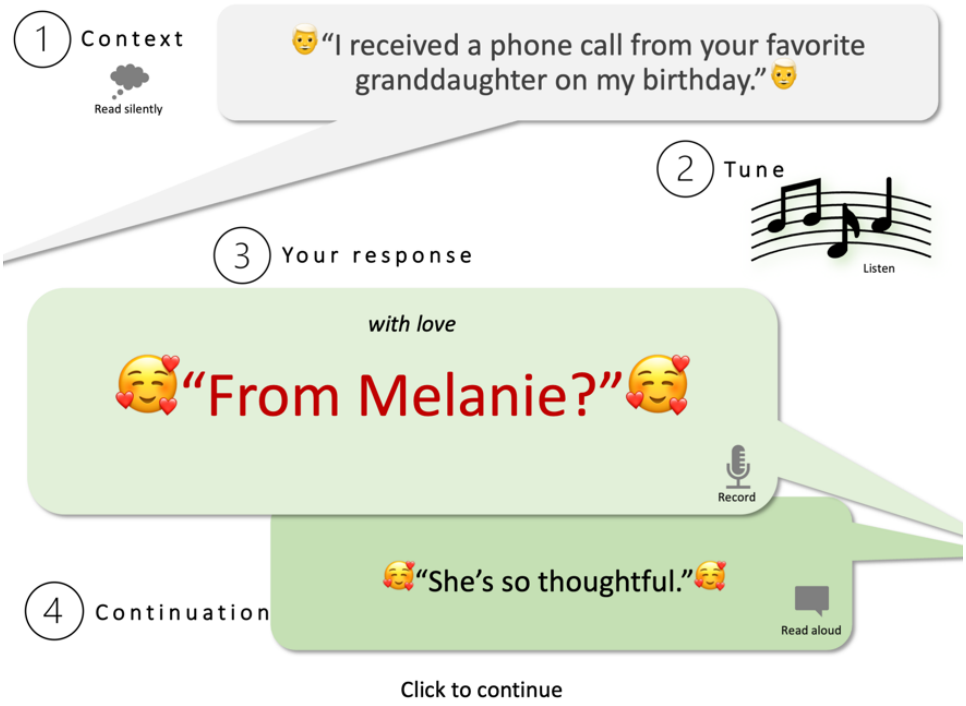


Figure 3: Tune-emotion elicitation procedure. (1) Participant silently reads context from "Bill". (2) Participant hears two models of the tune and emotion cues appear. (3—critical) Participant reads aloud the target sentence with the specified tune and emotion.

After the last trial block, participants were forwarded to a brief questionnaire to double-check their eligibility and, in the case of VAs, to also collect information about the microphone they used, to ensure it was appropriate. In all cases, equipment used by VAs was of equal or greater quality than the professional equipment used for SPs. Data from all eligible participants were submitted to the same analysis pipeline, which involved extracting F0 trajectories and submitting them to quantitative analyses, as described in the next section.

2E. Quantitative methods

The goal of quantitative analysis was to understand the acoustic characteristics of phonologically distinct linguistic forms, as represented in F0, the primary cue for intonation. This study focuses on F0 dynamics, which reveals patterns across time that are obscured by summary statistics (e.g. F0 mean, maximum, minimum) but are integral to the concept of F0 shape. The data analysis pipeline, described next, was designed to run in a streamlined unsupervised manner, so it can be easily iterated or extended.

TextGrid preparation

A key part of the acoustic analysis process is the creation of Praat TextGrid files which contain the annotations denoting the temporal boundaries of target words in the continuous audio recordings from trials. The first step involved aligning the TextGrids using the Montreal Forced Aligner (MFA; McAuliffe *et al*. 2017) using its pretrained English acoustic model, which generally returned usable but sometimes coarse word boundaries. Additionally, members of the research team hand-corrected the data collected from voice actors, mainly as a check of the adequacy of MFA for the data, since the emotional qualities of the data make it a potentially challenging case. To help gauge the extent of the problem of emotional variation for the TextGrid alignment, the same statistical model (a GAMM described below) is conducted on data with and without hand-correction. The model with hand-corrected TextGrids was a better fit and explained ~6% more of the deviance in F0, so in the present study the hand-corrected data are used where possible. Due to resource limitations, the subject pool data was not manually audited.

F0 data preparation

F0 is usually represented on the Hertz scale, which is linear, however human perception of pitch differences is fundamentally nonlinear, making Hertz inappropriate for the present study's research goals. This issue was compensated for by converting from Hertz to ERB, which helps to interpret differences in trajectories with respect to perception (Glasberg & Moore, 1990). Conversion was accomplished in R using the `soundgen` package (Anikin, 2019) using the 'linear' method, which applies the formula: $ERB = 21.4 \times \log_{10}(1 + 0.00437 \times Hz)$. The F0 trajectory extraction and preparation process proceeded as follows:

i. From each target word, F0 was sampled in 10ms intervals using the STRAIGHT algorithm for F0 estimation (Kawahara *et al*., 2005) implemented in VoiceSauce (Shue *et al*., 2009) in Hertz.

ii. Samples with possible F0 tracking errors were identified and removed in two stages. First, from a total of 6,002,175 continuous F0 samples from VoiceSauce, 137 samples where an F0 value could not be estimated, resulting in an NA, were removed (.002%). Failure to be able to estimate F0 at a given timepoint might be due to phonetics (unvoiced segments) or random audio artefacts. Second, the remaining data was analyzed for F0 jumps using a method proposed by Steffman & Cole (2022), which identified 408 additional samples with likely errors (.007%) which were removed.

iii. Using the clean F0 tracking data, 30 equidistant samples were selected for each trial in order to capture its holistic shape. A consequence of this step is a considerable down-sampling from ~6M to 4,764 total measurements.

iv. F0 values were speaker-normalized through centering, which meant calculating each participant's mean F0 across all Neutral trials, which is then subtracted from the samples. This was to reduce individual variation due to factors including speaker sex and gender.

v. Samples, which to this point were treated in Hertz, are converted to ERB to more faithfully reflect perceptual differences across the range of possible F0 values (see above).

The prepared F0 trajectories were submitted to two types of modeling, k-means clustering and GAMMs, described below.

K-means clustering

The patterns of F0 variation among the aggregated data were analyzed using k-means clustering to discover emergent groups of similar F0 trajectories based on the Euclidean distance between trajectories at each (time-normalized) time point. In other words, the associated labels that encode tune, emotion, and speaker are not included in the analysis, rather just the time-normalized F0 trajectories. This was implemented with the R script from Cole *et al*. (2023), which uses the `kml` package for clustering longitudinal data (Genolini & Falissard, 2011). The input to the model was each speaker's mean F0 trajectory across identical trials (same tune-emotion combination) not individual trials, which helps constrain measurement noise. The clustering algorithm maximizes between-cluster distance while minimizing within-cluster distance.

The result of the k-means cluster analysis gives the optimal number of clusters for the data (k), given a certain range of solutions to consider, which for this analysis was 2-20. This range is appropriate to evaluate the three possible outcomes given the design. i) a solution with fewer clusters than stimulus tunes—e.g., in line with Cole *et al*. (2023)'s finding of five emergent F0 patterns; (ii) a solution with eight clusters that map neatly onto the stimulus tunes, an ideal outcome for the AM model; or (iii) a solution with more than eight clusters, which could arise if a particular tune-emotion combination yields F0 trajectories that are unlike other trajectories for the same tune, and unlike trajectories of any other tune. In the case of this data, if the imitations faithfully reproduce all of the distinctions present in the stimuli, the phonologically-optimal output will have eight clusters, with all (and only) the imitations of a single stimulus tune grouped together in the same cluster.

GAMMs

The imitated F0 trajectories were also analyzed using generalized additive mixed models (GAMMs), adapted from Cole *et al* (2023) and implemented using the `mgcv` R package (Wood, 2017). Whereas Cole *et al*. (2023) models Tune alone, the primary GAMM in the present study considers tune-emotion combinations as terms (e.g. LLL-Neutral[10], HLH-Love, etc.) which will enable a comparison of how tunes are implemented differently as a function of emotion, independently and jointly with random effects by speaker (following Sóskuthy (2021).

[10] The reference level was LLL-Neutral. The choice of LLL was due to the fact its tune shape is very distinctive in this particular inventory; there are few falling F0 trajectories and LLL's monotonic fall seemed like a suitable point of comparison. The choice of Neutral is due to the lower expected levels of emotional variation.

Given that this design differentiated between VAs and SPs, the effect of cohort (and Age[11]) was added to the model, but ultimately did not improve the model fit and was subsequently removed.

Recall that Cole *et al*. (2023), testing the same set of eight 'basic' tunes with the same resynthesized pitch stimuli, found evidence for maximally five distinct tunes produced as imitations of those stimuli. In addition, Cole *et al*. presents results from an AX discrimination test showing four pairs of tunes that were not accurately discriminated above chance levels: LHL-LLH, LHH-LLH, HLH-HLL, and HHH-HHL. The present study is premised on the idea that the tunes in these confusable pairs in particular may be produced more distinctly when elicited in the context of a compatible discourse and specified emotion. To statistically evaluate such a claim, predictions for tunes have to be compared within the same GAMM, which is accomplished using a 'difference GAMM'. Difference GAMMs are method for comparing two levels of an independent variable (e.g. tune) over time, and it is possible for some time spans to be statistically indistinguishable. For the purposes of the present study, difference GAMMs are mainly used to compare the same tune in different emotions. Additionally, different tunes in the same emotion can be compared in the same way, which should reflect AM-conditioned phonological contrasts. In line with Cole *et al*. (2023), a visual inspection of the difference GAMMs was conducted to evaluate statistical significance (the numeric coefficients for the parametric or smooth terms are not particularly informative for the questions asked here). Significance is determined by visualizing the mean trajectories predicted by the smooth terms and the 95% confidence interval (CI) around the mean trajectories; if the difference between

[11] Because SPs were college age and VAs tended to be older, speaker age was also tested as an effect. However, it was dropped as it did not systematically affect F0 variation related to tunes, although differences did emerge that merit further analysis of age-based differences on tune production in the future with a more age-balanced participant pool.

model predictions for two tunes is greater than the mean +/- 95% CI, based on Sóskuthy (2021), it is considered significant.

The purpose of this secondary GAMM is to test the extent to which phonologically predicted distinctions emerge from data lacking tune labels. Unlike the clustering algorithm, the GAMM uses different levels for each speaker (as a random effect), which is expected to decrease the amount of apparent noise overall (see Table 2). This makes it possible for the GAMM representation of the clustering solution to be more in line with the phonological predictions compared to the distributional qualities of the clustering solution or means of clusters, which are also examined.

Table 2: GAMM formulae

Name	Code

Tune-Emotion	`F0 ~ Combination +` `# Parametric term`
	`s( sample, by = Combination, bs = "tp", k = 10 ) +` `# Smooth term`
	`s( sample, spkrID, bs = "fs", m = 1, k = 10 ) # Random effects`

Cluster	`F0 ~ Cluster +` `# Parametric term`
	`s( sample, by = Cluster, bs = "tp", k = 10 ) +` `# Smooth term`
	`s( sample, spkrID, bs = "fs", m = 1, k = 10 ) # Random effects`

2F. Results

The time-normalized F0 trajectories (by-tune averages within speaker) were submitted to GAMM analysis and k-means clustering. Each model's goals, results, interpretation, and findings are presented in this section. Raw data, R analysis scripts, and models are freely publicly available through the study's Open Science Framework repository (https://osf.io/gbk8z/, see Experiment #1).

Empirical results

To gain insight about the model's findings, the aggregated empirical data is presented in this section in the form of mean F0 trajectories plotted as a function of tune and emotion, both across and within participants. Comparing the average F0 trajectories of each tune across the five emotion conditions, in the data aggregated across participants (Figure 4), it appears that the emotion context has little impact on the overall shape of the tune. Likewise, it appears that positive Valence emotions (Love, Pride) are generally implemented with a higher average F0 versus negative ones (Anger, Shame). As mentioned, the Neutral emotion condition most closely parallels the elicitation procedure of Cole *et al*. (2023). The F0 trajectories in the Neutral condition do not appear identical to any particular emotion, for any of the tunes, nor do the Neutral trajectories resemble an average of the different emotions for any tune. Rather, the effect of Neutral depends on the tune—for HHL it has the highest F0 maximum while for LHH it has the lowest F0 minimum. Therefore, a preliminary takeaway from this analysis is that Neutral/decontextualized tune elicitations are not reliably representative of the tune as it would be produced in a richer, emotionally specified context.

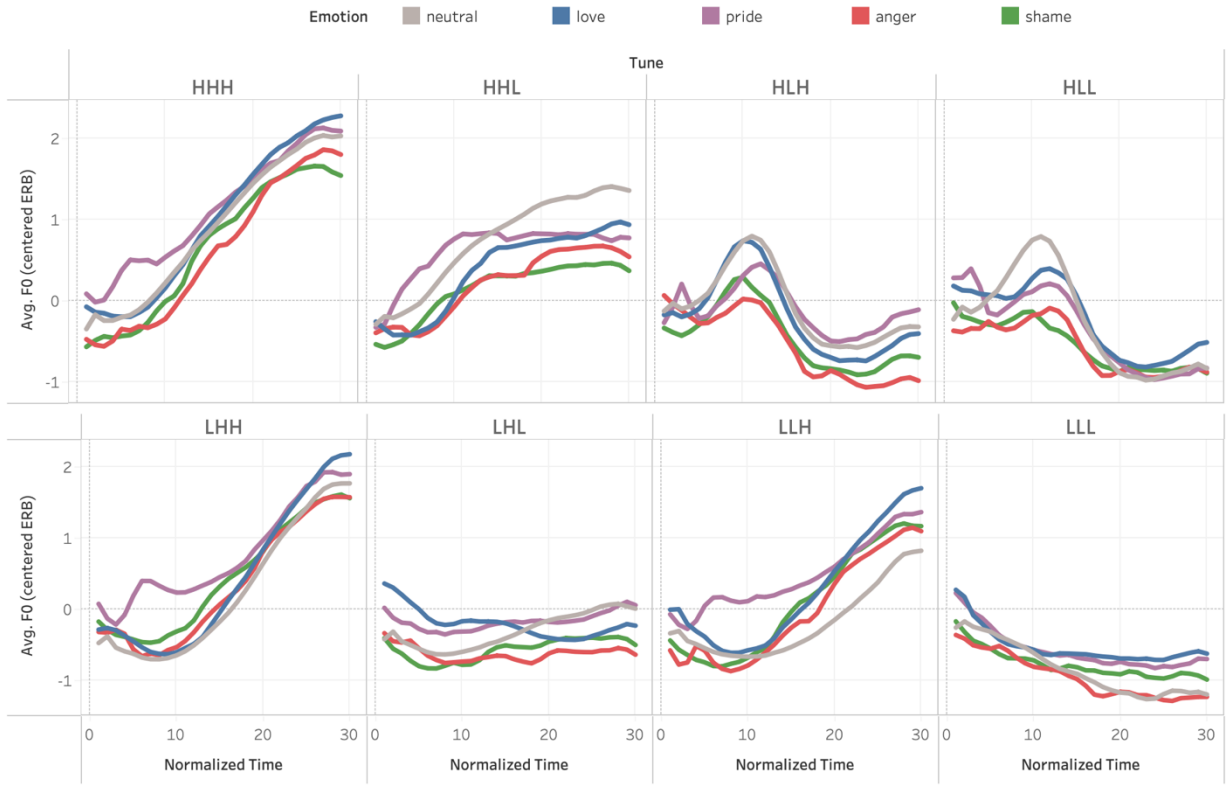


Figure 4: Across-participant empirical average F0 trajectories by tune and emotion. Each line represents the average F0 trajectory for one tune produced in the context of one emotion.

Expanding the visual assessment to the within-participant average F0 trajectories (Figure 5), there appears to be substantial variation by participant within each tune-emotion combination. Despite the degree of variation within each tune, the expected distinctions among tunes, based on their tonal specification in the AM model appear, if only in coarse-grain, with tune shapes that appear less distinct than in Figure 4. Based on the trajectories in Figure 4 (repeated in the top panel of Fig. 5), participants seem to imitate the general shape of the tunes while disregarding variation by emotion (comparing different colored lines within each panel) certain pairs of tunes appear to share markedly similar F0 trajectories, e.g., HHH-HHL, HLH-HLL, and LHH-LLH. A core question for this study—which cannot be answered from visual inspection of these average

F0 trajectories —is whether phonologically distinct tunes were produced with acoustically distinct F0 trajectories; modeling is required

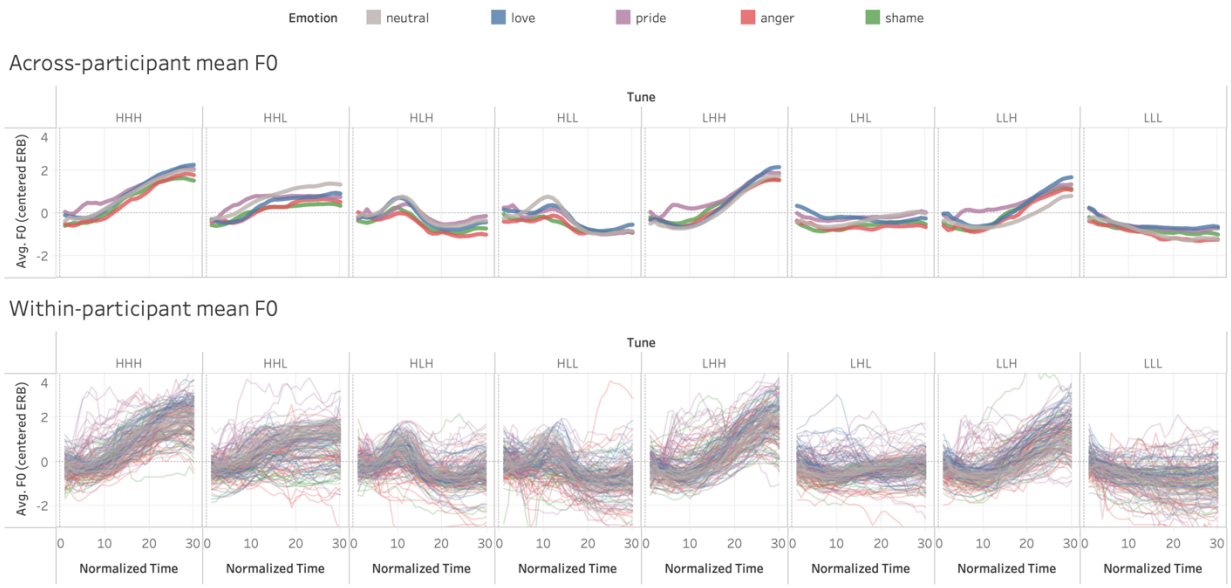


Figure 5: Empirical F0 trajectories by tune and emotion, aggregated across participants (top) and within participants (bottom) . Bottom: Each line represents one participant's average F0 trajectory for a tune-emotion combination (1,280 total)

Clustering results

Overall, k-means clustering reveals six distinct F0 trajectory shapes emerged from the imitated tune productions (k=6), the mean trajectories of which are in Figure 6. The shape differences between clusters appear to show most of the contrasts that the model tunes encoded, but the underlying data (plotted behind the means in Figure 6) show huge variation. The cluster distinctions show a distinction between high rising clusters (E) from shallower rises (C and B), as well as a rise-fall cluster (D), a 'flat' or monotone cluster (A) and a falling cluster (F).

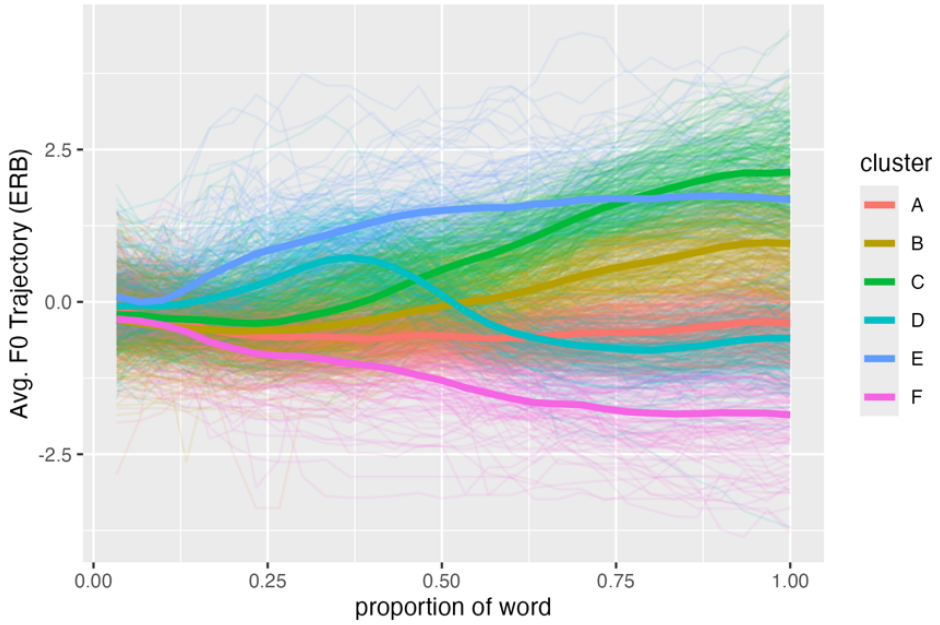


Figure 6: Mean trajectory of each of the six clusters. Comparable to Figure 5 in Cole et al. (2023).

The clustering solution is analyzed in terms of tune-cluster correspondence (Figure 7a, like Cole *et al*.'s Figure 5B) and tune-emotion-cluster correspondence (Figure 7b). Note that while Figure 7b shows the breakdown of the clustering solution by emotion, the model was not run on subsets of the data for this analysis[12]. For example, Figure 7A shows that HHH tokens were split between clusters B, C, and E, whereas Figure 7B further shows that many HHH tokens are Neutral in cluster C. No clusters in 7A contained the majority of any particular tune, but LLL was closest as about 50% were assigned to cluster F. Based on Figure 7B, the distribution of LLL within cluster F does not vary as a factor of emotional variation. Certain tunes also appear

[12] A version of this analysis wherein clustering was also performed over subsets of the data by emotion in order to better understand how emotional variation impacting the clustering results. Doing so constrains the emotional variation in the data, since all tunes are expected to be encoded similarly across conditions, but emotion is expected to have a general effect that clustering might partly account for. Since the results of this modeling exploration mostly speak to the relative effectiveness of k-means clustering for highly variable data, and not the research question at hand, the current analysis focuses on a more unified analysis of tunes across emotions.

to pattern together in Figure 7a, such as the "high-rising" tunes HHH and HHL, which have similar distributions across clusters, a trend which holds across emotions in Figure 7b.

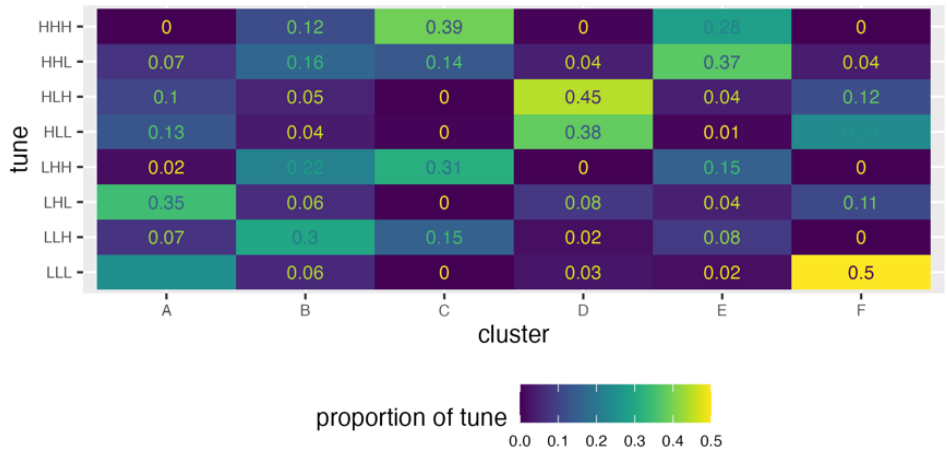


Figure 7a: Clustering results: tune-cluster correspondence by TUNE across clustering result (k=6). Numbers indicate proportion of tune tokens per cluster; rows sum to 1.

While participants were imitating eight tunes, only six clusters emerged from the analysis, suggesting some clusters contain trajectories from multiple tunes. As shown in Figure 6b, while there is variation in tune-cluster correspondence across emotions, no obvious patterns emerge, which is in line with the empirical means (Figure 5), tunes seem to be the primary driver of F0 trajectory distinctions in this data. For example, in the Neutral condition, Cluster D is composed of similar proportion of imitations of HLH and HLL (84% each), but under Love HLH-HLL share the corresponding cluster with LHL tokens (34%) and under Pride with HHL tokens (22%). Under Anger and Shame, Cluster D only includes trace proportions of tokens from other tunes (>10%) so these emotion conditions closely approximated Neutral. To take Cluster F as another example, under Neutral it is dominated by LLL (45%) and is similar for Love, but under Pride Cluster F has LLL joined by HLL (22%) while under Anger there are even more instances of HLL imitations in Cluster F (31%).

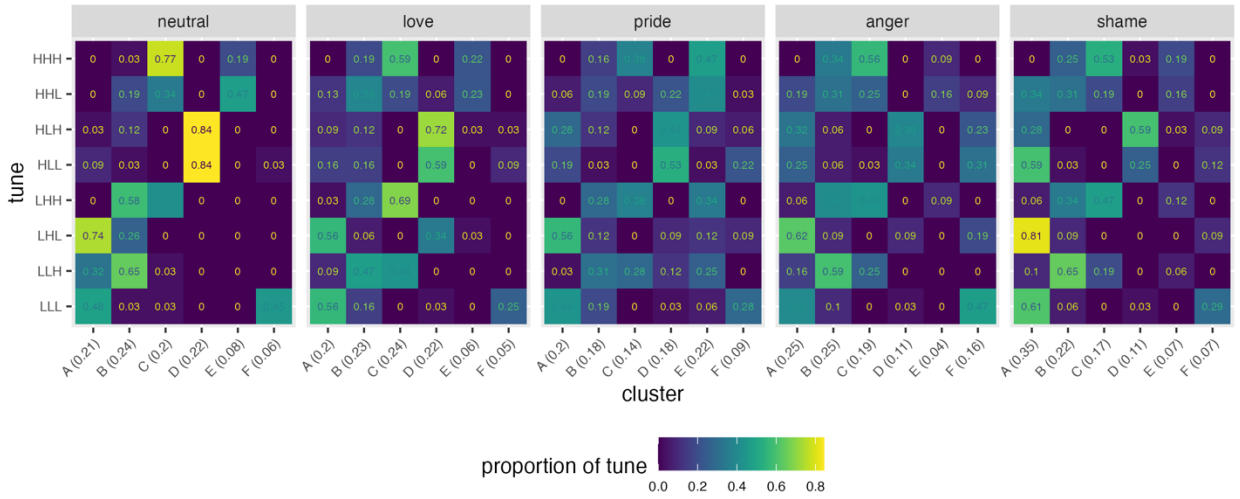


Figure 7b: Tune-cluster correspondence across clustering result. Cells indicate proportion of tune tokens per cluster; each row (a tune-emotion combination) sums to 1 across emotions. Cluster labels shown with proportion of emotions per cluster and sum to 1 within emotion.

While emotion did not drive the clustering results, there were some notable effects. The least-structured ('noisiest') emotion appears to be Pride, where no Cluster contains more than the maximum of 56% of tokens for any given tune, followed by Anger where the maximum is 62%. The other emotions show a more consistent tune-cluster mapping, with greater proportions of tune imitations in any one cluster and with more clusters dominated by one or two tunes, including Love with a maximum tune proportion of 72% and Shame with 81%, which is comparable to Neutral's 84%. The Neutral condition closely resembles the data reported in Cole *et al*. (2023), and results from the current study show a partial replication of their clustering solution (see Table 3). The solution for the present study includes an additional cluster compared to Cole *et al*., which is partly explained by the emergence of an HHL-dominated cluster (E). Also, half of the clusters in the present study's Neutral condition include tokens from at least three tunes (A, B, C), and in one case five tunes (B), while Cole *et al*. found a maximum (and median) of two tunes composing each cluster. Relatedly, just one cluster in the present study's

Neutral condition was composed of tokens from only a single tune (F), while for Cole *et al*. two clusters did. Together, this suggests that the mapping from tune to cluster in the present study is more variable than prior work not invoking emotion.

Table 3: Comparing clustering results to Cole *et al*. (2023)

Cluster	Current study: Neutral condition only	Cole *et al.*
A	LHL (74%), LLL (48%), LLH (32%)	LHL (100%), LLH (97%) {A}
B	LLH (65%), LHH (58%), LHL (26%), HHL (19%), HLH (12%)	LHH (93%) {D}
C	HHH (77%), LHH (40%), HHL (34%)	HHH (83%), HHL (83%) {C}
D	HLL (84%), HLH (84%)	HLL (97%), HLH (97%) {B}
E	HHL (47%), HHH (19%)	
F	LLL (45%)	LLL (97%) {E}

- Threshold for inclusion in this table was 10%.
- Cole *et al*. (2023)'s cluster labels given in curly brackets beside similar cluster from current study.

To better understand the effect of emotion on the clustering findings, the distribution of emotions across the solution was also visualized, see Figure 7c. If emotion is a primary determinant of clusters, it would predict an asymmetric distribution whereby each cluster would be dominated by one and only one emotion. Instead, no single emotion represented more than 50% of any cluster, although one case came close. The strongest cluster-tune correlation was found in Cluster E for Pride (47%), followed by Cluster F for Shame (37%). On the other hand, the weakest tune-cluster correlation was also in Cluster E but for Anger (9%), followed by Cluster F for Love (11%), which is interesting because Love and Shame are both high Potency, based on our psychometric model, contra Pride and Anger which are low Potency. If this is a Potency

effect, however, it does visibly manifest in other clusters, as outside of these edge cases the distribution of emotions across clusters is mostly uniform (with a mean ~20%). The distribution of Neutral is similar to Love (within 5%) across clusters.

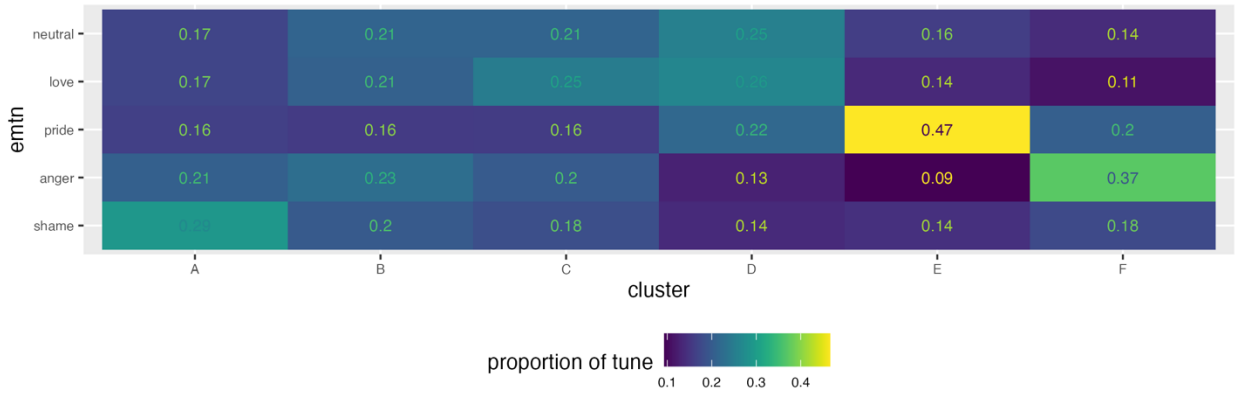


Figure 6c: Emotion-cluster correspondence (each column sums to 1)

GAMM results

The primary GAMM predicted variation in F0 trajectories as a function of tune and emotion and their interaction, with speaker-level random effects. The full output of this model is given in Appendix 2C. A secondary GAMM was fit predicting F0 variation by cluster. Following the recommendations by Sóskuthy (2021), the focus of analysis is on the smooth terms, which capture significant differences in the shape of the F0 trajectories between conditions. Figure 8 shows the predicted F0 trajectory for tune-emotion interactions from the GAMM, paneled by tune. Inspection of the empirical means (Figure 4) revealed significant variation compared to the GAMM. The emergent tune shapes appear to reflect the expected phonological contrasts defined by the AM model, but variation by emotion makes it difficult to tell whether tune shapes are distinct within the same emotion.

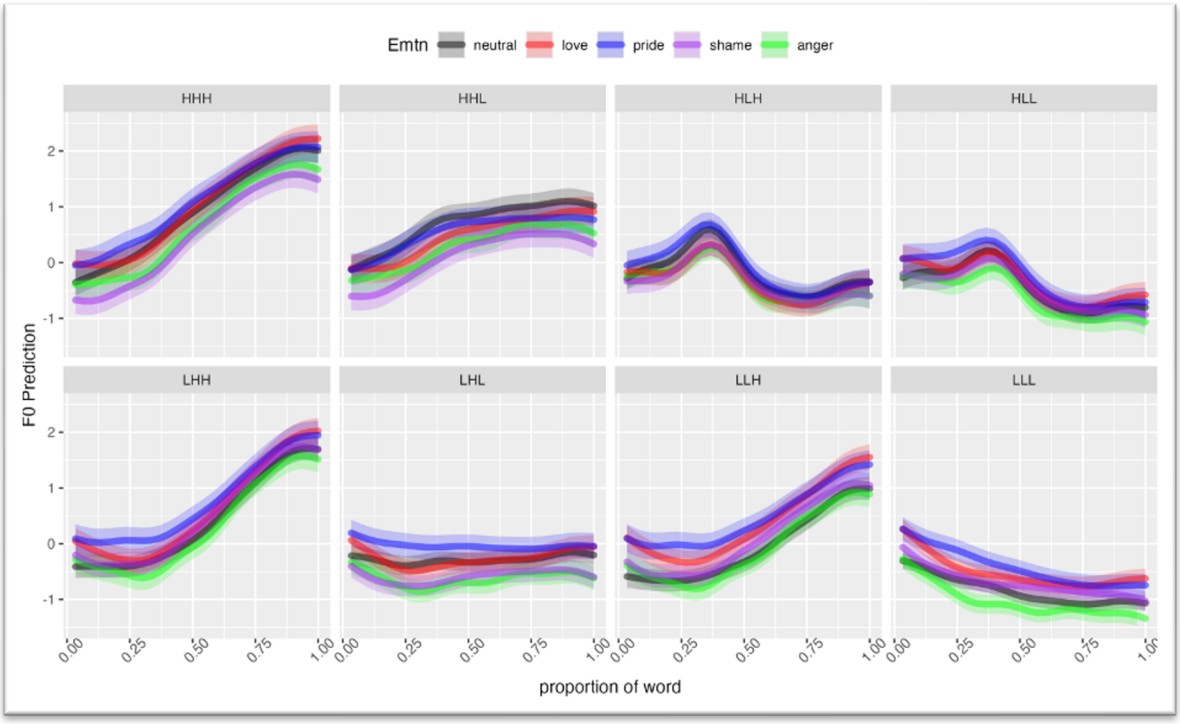


Figure 7a: GAMM predictions for tunes by emotion, relative to Neutral.

One trend in the GAMM predictions is that positive Valence emotions (Love, Pride) tended to be produced with overall higher F0, which is present in Figure 8a but easier to see when the respective emotions are plotted together, as in Figures 8b-c below. Almost all tunes (not HHL) had a higher F0 prediction in one of the positive Valence emotions versus Neutral (Figure 8b). For negative Valence emotions, the F0 predictions overlap with Neutral more often, but when they diverge, F0 is typically lower, although the F0 scaling effect is less clear than for positive Valence. The tune that most clearly exhibits a Valence effect is HLL, which has a predicted F0 trajectory for Neutral that lies between the positive and negative Valence emotions.

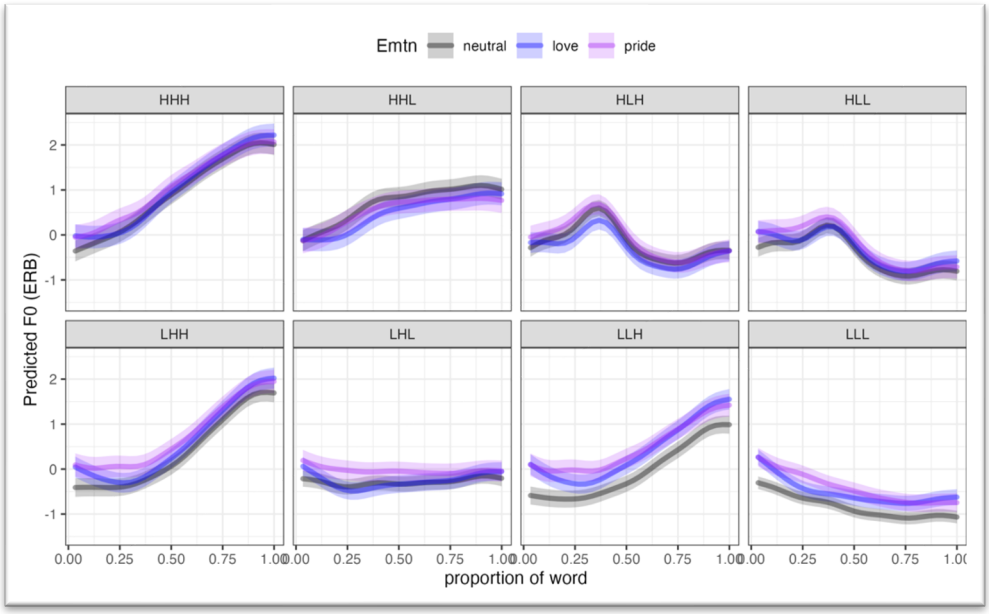


Figure 8b: GAMM predictions for Positive valence tunes by emotion, relative to Neutral.

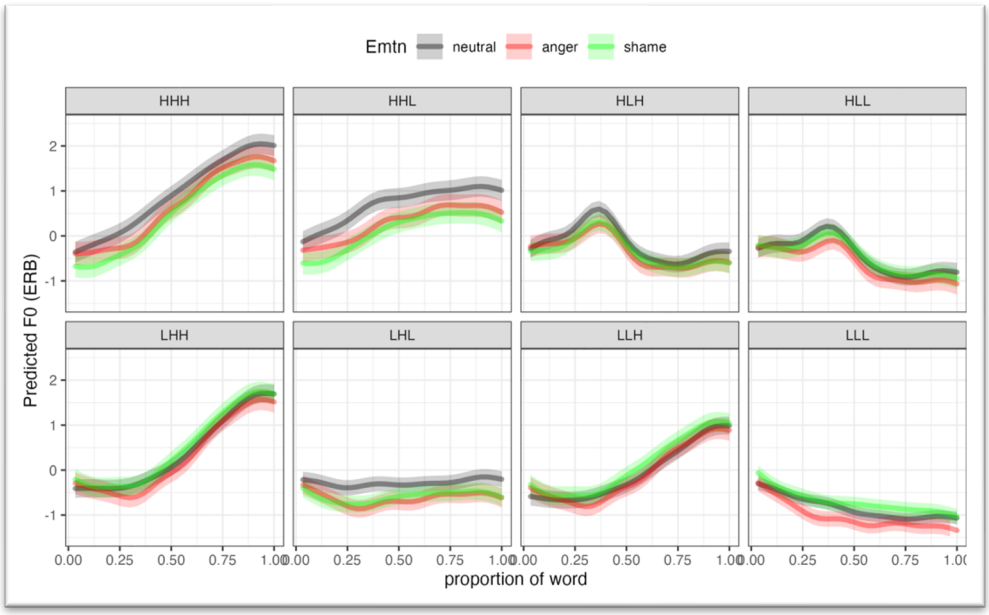


Figure 8c: GAMM predictions for Negative valence tunes by emotion, relative to Neutral.

Based on the GAMM, the phonetic realization of phonological contrasts is remarkably consistent once other sources of variation are considered, which partly validates the present study's strategy and thus bodes well for answering the research questions. However, the plots of Figure 8 do not lend themselves to fine-grained comparison between different tunes, in order to evaluate their distinctiveness in a given emotion. To address this gap and thereby assess the statistical significance of the observations made thus far, an exhaustive set of difference GAMMs was generated, discussed next.

Difference GAMM Analysis

Between-tune plots. In the following plots (Figure 9a-d) difference between the GAMM's F0 trajectory predictions are plotted for different tunes under the same emotion (Tune 1 – Tune 2), with the y-axis being F0 (in ERB) and the x-axis being time (in 30 equidistant samples) The dark line shows the fit smooths and the gray band around the line shows the 95% confidence intervals (CI). The difference GAMM is positive when Tune 1 has a higher F0 than Tune 2 at a given sample (in normalized time). The difference is negative when Tune 2 has a higher F0 than Tune 1 in the corresponding interval. For two trajectories to be considered statistically significantly different (following Sóskuthy, 2021), the 95% confidence interval (CI) for the difference curve must exclude zero for some part of the curve. If the difference curve approximates zero across the predicted trajectory, then it indicates that the trajectories being compared are statistically indistinguishable. In some cases, the regions of significant differences consisted of a single (or very few) samples, which do not easily lend themselves to phonological interpretation. Therefore, a criterion was developed whereby significant regions had to include at least five

contiguous samples (~16% of total duration) in order to be considered relevant to the difference GAMM analyses. First the difference GAMMs for the pairs of tunes that clustered together (in this study and for Cole *et al*. 2023) are presented: LHL-LLH, HHH-HHL, and HLL-HLH.

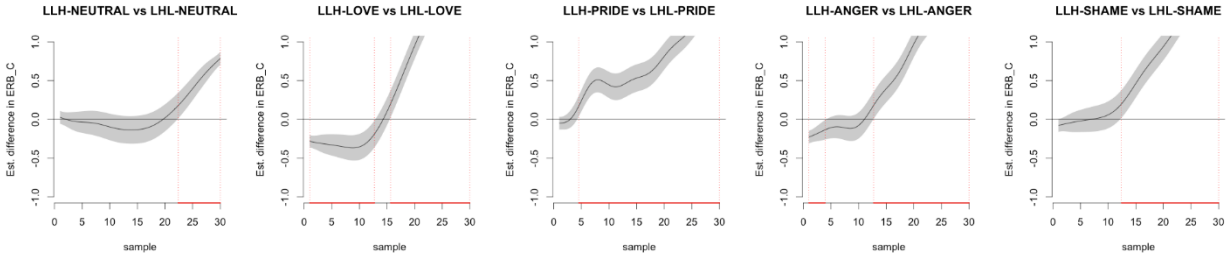


Figure 8a: LHL-LLH difference GAMMs

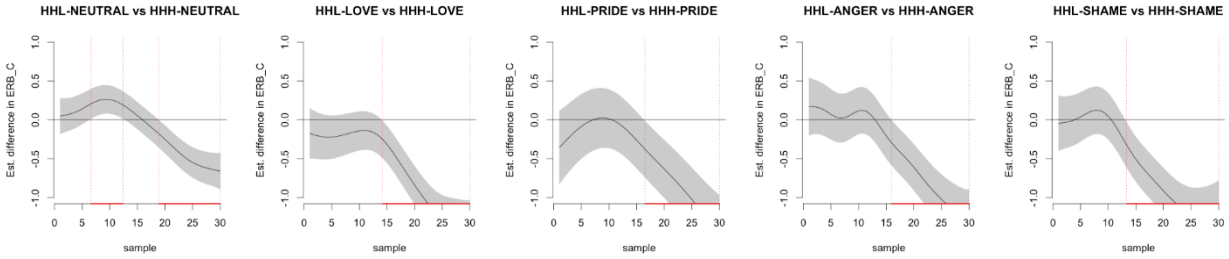


Figure 9b: HHH-HHL difference GAMMs

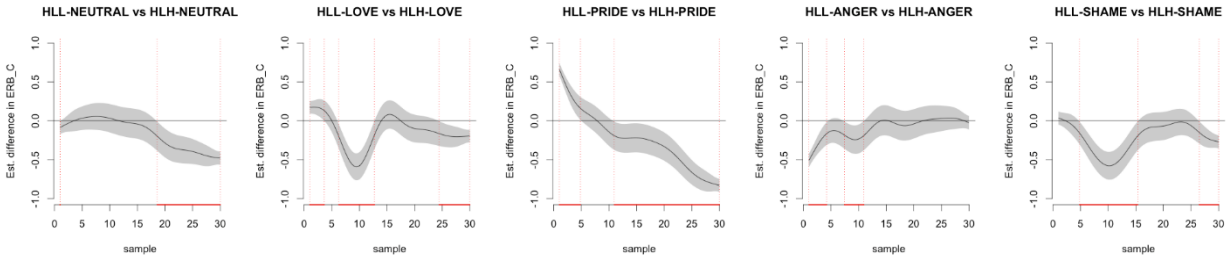


Figure 9c: HLL-HLH difference GAMMs

All difference GAMM in Figure 9 shows regions of statistically significant differences between tunes, although one tune pair is not distinct according to the analysis criterion, HLL-HLH under Anger. The other pairs, LHL-LLH and HHH-HHL had significant differences over an interval longer than five contiguous samples. What stands out in these figures is that while the GAMM finds significant differences between tunes for the pairs {HLL, HLH} and {LHL, LLH} in all emotion conditions (Figure 9c, top and middle panels), the difference is of roughly the same magnitude across emotion conditions. The result is more varied for the pair {HLL, HLH} (Figure 9c, bottom panel), where the difference between these tunes appears in different regions of the interval depending on the emotion, but again with no clear evidence that the difference is enhanced by emotion. Appendix D contains all difference GAMMs for pairs of mismatching tunes under the same emotion; no additional tune pairs failed to emerge as distinct. While the effect of emotion on tunes is not the focus of the present study, this illustrates the effects of the emotion manipulation, which typically manifested in the raising or lowering of F0 across the trajectory.

To help visualize the many (156 total) difference GAMMs, the cumulative significant difference in F0 (‘delta’ hereafter) between every combination of experimental variables was calculated. This involved extracting every timepoint that the difference GAMM identified as significant, meaning the CI of the difference curve excluded zero, and taking the cumulative sum of the absolute value of all the differences in ERB. This treats F0 space like Euclidean space, so the deltas are always positive, which arguably makes this a measure of acoustic distance. Deltas are first shown on a scatterplot where tune pairs are in ascending order, so the distribution of values and the relative rank of pairs can be seen within and between each emotion (Figure 10a). Then

the values are shown in a heatmap (Figure 10b) and finally a bar plot which normalizes the deltas to Neutral by subtraction (Figure 10c).

Starting with Figure 10a, the deltas for when both pairs in Neutral are shown in the top panel, which is why no same-tune pairs appear, since comparisons between the same tune in the same emotion are identical in the GAMM (delta = 0). The Neutral delta cline starts ~10 (LLH-LHL, HLL-HLH, and HHL-HHH) and gradually increases to ~270 (HHH-LLL). The four labeled emotions also showed gradual differences across tune comparisons, but three of the emotions (Love, Anger, Pride) show a positive deflection around 50 that breaks otherwise linear trend. For example, see LLH-HHL (purple square) which is always >50 versus LLH-LHL (purple square) which is usually near LHH-HHL in rank order, but has a delta <60, on the other side of the deflection. Most but not all of the most similar tune pairs (lowest deltas) had the same tune, which means the tune was produced with a similar F0 trajectory in that emotion and Neutral. There are exceptions, though, such as in Shame where LLL-HLL was the most similar pair and HLL-HLH was the third most similar.

Figure 10a: Scatterplot of cumulative significant differences (CSD) paneled by emotion. Within each panel, tune pairs (represented by a unique color-shape combination) are in ascending order. Low values indicate little difference between predictions for that tune pair/emotion combination; high values indicate larger differences. See Figure 10b-c for alternative visualizations of this data.

Moving onto the heatmap presentation of this data (Figure 10b), if tune is the primary factor determining F0 shape, and emotion has no effect on F0 shape, then the delta values are expected to be comparable for the same tune pairs across emotions. Generally, tune distinctions were of larger magnitude in Neutral productions, but many exceptions emerged. One exception is LLH-LHL, which was one of the least distinct tunes under Neutral but not so with the named

emotions. Compared to LLH-LHL, HLL-HLH is not well differentiated under Neutral nor the named emotions, meaning there are few cases where HLL and HLH are as (relatively) distinct.

Figure 10b: Heatmap of deltas from difference GAMMs. Low values indicate little difference between predictions for that tune pair/emotion combination; high values indicate larger differences.

To more easily compare tune pairs in terms of robustness across emotions, Figure 10c subtracts the Neutral value from each of the named emotions, which should remove the differences that tend to emerge because of the phonological specification. In other words, if the four plots where the differences are normalized to Neutral are identical, it means emotions drove consistent effects across tunes, rather than there being evidence of particularized tune-emotion effects.

While not identical, the four plots show tune pairs have roughly the same acoustic distinctiveness regardless of emotion, relative to Neutral. There are a few exceptions where a tune pair is higher than Neutral in one emotion but lower in Neutral in another, such as LLL-LHH which is higher under Pride but lower under Love, but the differences are numerically small.

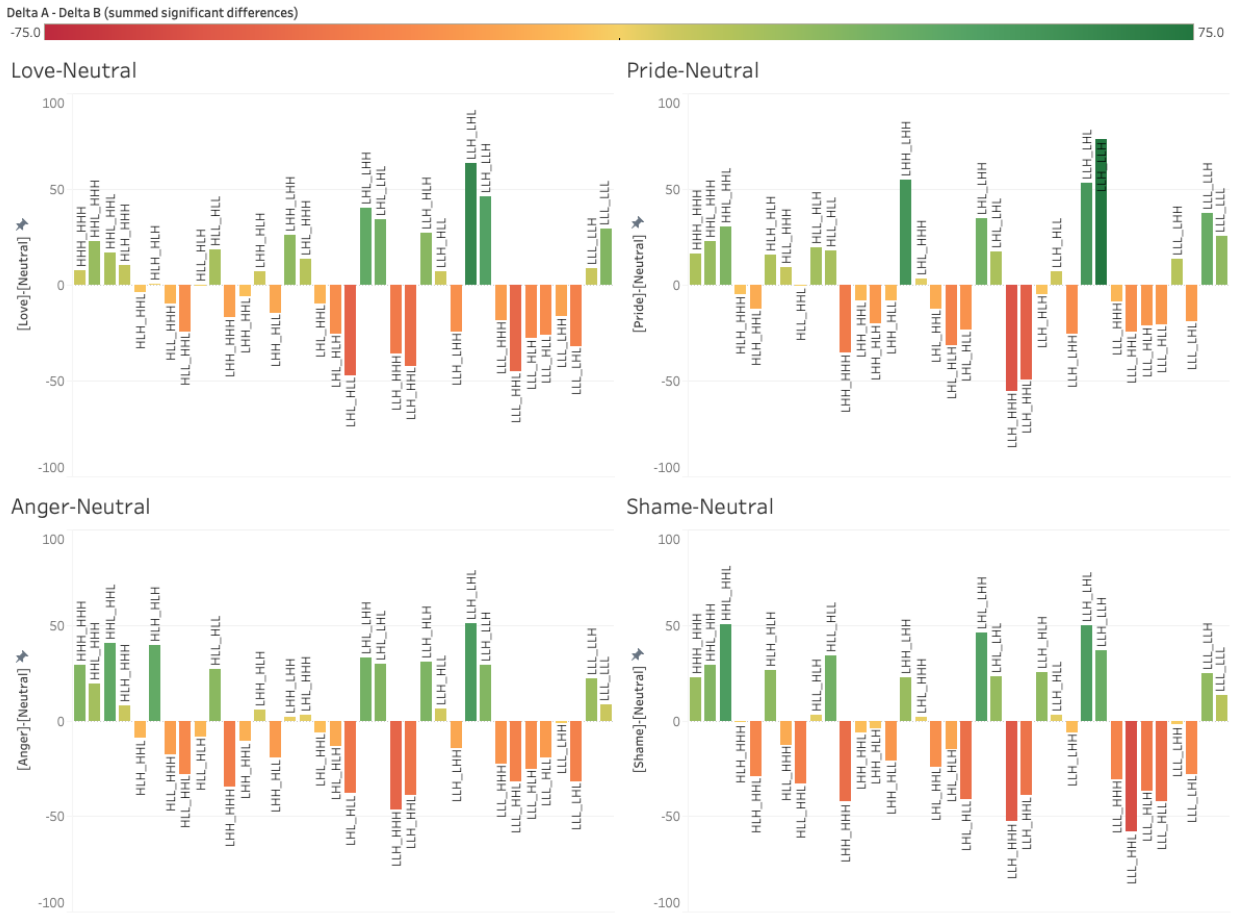


Figure 10c: Bar plot of deltas in identical order, difference between tune pairs in labeled emotions and Neutral (e.g. Love minus Neutral). Positive values are green, negative are red. Since same-tune comparisons were always zero in Neutral, they were omitted here, as in previous figures. If emotions affected tune pair distinctiveness the same way, the four panels should match.

Whereas Figure 10b shows the deltas within a particular emotion, Figure 11 below displays differences between each tune in Neutral versus a specified emotion. The magnitude of within-

tune differences between emotions was smaller than between-tune differences within emotions, and there was relatively little variation in the effect of emotion across tunes and across emotions. All within-tune difference GAMMs are given in Appendix 2E. Overall, this evidence helps support the idea that the effect of the emotion on tune shape was minimal, which is a pattern of results that looks quite different from between-tune comparisons.

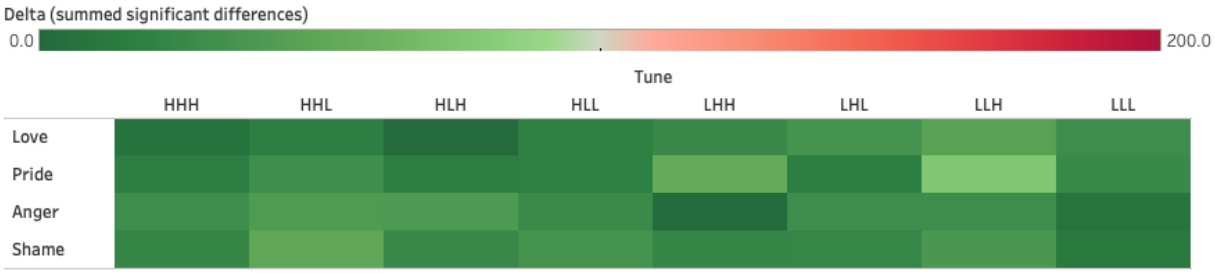


Figure 9: Summed difference GAMM regions (deltas). Low values indicate little difference between predictions for that tune in the given emotion compared to Neutral; high values indicate larger differences. Note same color scale as Figure 10b.

Based on the GAMM predictions by tune (Figure 8a) and the difference GAMMs (as in Figure 9), F0 trajectory shape appears to be a robust feature of tunes across emotion contexts. Moreover, similarity relationships between tunes tend to hold across emotions, such that the same tune pair tends to exhibit F0 differences of the same relative magnitude, i.e., larger or smaller on the color scale of Figure 10b, same patterns observed across tune pairs in emotions relative to Neutral in Figure 10c. Overall, these results suggest highly systematic effects for emotion across tunes, which bodes well for the testing of the experimental hypotheses. The next section summarizes the goals of the study, considers the experimental evidence relative to the predictions, and offers interpretations of the findings.

2G. Discussion

Reviewing the objectives

The present study considered the effect of speaker emotion on the production of phonologically specified tunes, which was analyzed in terms of F0 trajectories over the nuclear region of the intonational phrase. A speech production experiment was conducted where speakers imitated eight tonally distinct tunes presented auditorily in the form of pitch-resynthesized speech, while portraying one of four specified emotions and in an unspecified (Neutral) emotion condition. Tunes were produced in target sentences embedded in the context of a rich dialogue that provided a pragmatic context appropriate for each target tune. The F0 trajectories were extracted, prepared, and modeled, first using k-means clustering to test for distinctions among imitated productions that reflect the tonal specifications of the eight target tunes and/or the effects of emotion portrayal on the production of those tunes. In addition, GAMMs were used to model variation in the shape of F0 trajectories as predicted by the phonological tune labels plus the portrayed emotion. In a separate model, a GAMM modeled F0 variation predicted by the cluster to which the imitated production was assigned in the clustering analysis. The k-means clustering analysis showed an optimal partition of the imitated F0 productions into six distinct clusters, each with a distinct F0 shape. Identifying each imitated production with the tune label of the stimulus it was meant to imitate revealed a tune-to-cluster mapping showing that while many of the eight tunes were distinguished in the imitated productions, certain tune pairs were imitated with very similar F0 patterns and were grouped together in the clustering solution. The clustering solution for the Neutral emotion condition in the present work is similar to the solution in Cole *et al*. (2023), though with more variability in the tune-to-cluster mapping and with one additional

cluster. The clustering solutions for data in the four specified emotion conditions show even more variability in tune-to-cluster mapping and are thus less similar to the solution in the Cole *et al*. study. These findings show that tunes produced in rich dialog contexts are overall more variable, blurring tune distinctions that are present in tune produced with no dialog context (the Neutral emotion condition of the present study). As a way to better understand how emotion influenced F0 in conjunction with tune in these data, GAMMs were used. Smooths from the primary GAMM showed that F0 trajectories for a given tune had a highly similar shape across the emotion conditions, and also showed consistent effects of emotion on F0 variation for emotions grouped by psychometric dimension, particularly Valence. Specifically, tunes in the context of emotions with negative valence tunes (Anger, Shame) were produced with F0 trajectories that were overall lower than those with positive valence (Love, Pride), and tunes produced in the Neutral (unspecified) emotion context typically patterned between the two (see Figure 8b-c). The relative differences in F0 height for tunes by Valence accords with the psychometric model which laid the foundation for the emotion manipulation, Fontaine *et al* (2007), so the results also provide some validation for predictions from that theory.

Hypothesis evaluation

This study evaluated two experimental hypotheses, first that the joint production of phonological specifications and emotional expression yields patterns of variation that ultimately preserves phonological distinctions, and second (only if the first hypothesis is supported) that speakers encode similar degrees of tune distinctiveness regardless of the specific emotion, including Neutral.

To evaluate how the evidence supports or challenges the first hypothesis, recall that it was predicted that speakers will adapt their imitation of tunes to maintain their mutual distinctiveness in terms of F0 shape within a given emotion. Broad evidence emerged for tune-emotion interactions in the clustering and GAMM analyses, supporting the idea that speakers are adapting their realization of phonological contrasts while also producing emotion cues in F0. Put simply, the properties of F0 trajectories that encode contrastive tune shapes appear "warped" systematically by the co-expression of emotion, without losing consistency in the shape of F0 trajectories by tune across the emotion conditions (Figure 8a). According to the difference GAMMs, nearly all tune pairs were statistically distinct within each emotion condition, though for a given tune pair, the degree of difference varies across the emotion conditions (Figure 11).[13] In the clustering analysis, maximally six distinctive F0 shapes emerged rather than the full eight tune inventory, which suggests that most tune distinctions might be robust even without the need to account for emotional variation. Several of the predicted pairwise tune distinctions fail to emerge in the clustering analysis (Figure 6), and the same tune pairs are among those with the weakest differences by GAMM analysis (Figure 11). The findings from GAMM analysis point to tune as the primary determinant of the F0 trajectories in this study, with significant differences for all tune pairs in all emotion contexts along at least some portion of their F0 trajectory. Given the strong results in the GAMM, Hypothesis 1 seems well supported despite certain tune pairs being produced similarly in certain emotional contexts.

[13] While it was not the primary focus of this thesis to characterize how emotions are portrayed through F0 patterning, it's notable that there were only two tune-emotion combinations which failed to show a significant effect of emotion (compared to Neutral), HLH-Love and LHH-Anger, raising questions about how the emotion is be encoded in those cases.

The positive finding for Hypothesis 1 opens the possibility to evaluate Hypothesis 2, which predicted that the F0 shapes corresponding to tunes will not be systematically more or less distinct in any given speaker emotion, including Neutral. Looking at the GAMM predictions by tune (Figure 4) the degree of variation between emotions within each tune appears to be similar, without clear visual evidence of emotion enhancing or diminishing distinctions in F0 trajectories between tunes. The heatmap summarizing the difference GAMM analysis (Figure 10b) also suggests that tunes in F0 space may be slightly more distinctive under Neutral on the whole. That said, it is difficult to argue from this evidence that Neutral provides a general distinctiveness advantage considering the many cases where tunes were produced under emotion exhibited greater degrees of difference. If it is the case that Neutral and particular emotions can have general effects on tune distinctiveness, then it seems the emotions selected for the present study did little to muddle tune productions. Failing to find greater magnitude of distinctions across tunes in Neutral, i.e., a distinctiveness advantage under Neutral, the experimental evidence supports Hypothesis 2. Overall, the results show that (i) across the emotion contexts, tunes are phonetically implemented with F0 trajectories that generally conform to predictions of the AM model, and (ii) variation in the phonetic implementation of tunes is structured in relation to the emotion context with similar degrees of distinctiveness.

Evaluating tune contrasts

Clustering identified six clusters of F0 trajectories, which means not all eight phonologically distinct tunes were realized with distinct trajectories (see Figure 6: Tune-cluster correspondence). This is similar to the level of contrast found by Cole *et al*. (2023), which identified five clusters

using the same methodology but without an emotion manipulation or pragmatically motivated dialog. The clustering solution here includes one additional cluster than Cole *et al*., and a weaker correspondence between tunes and clusters, but the same tunes tended to cluster together across studies and (important for the research question) across emotions. Several of the predicted pairwise tune distinctions fail to emerge in the clustering analysis (Fig. 6), and the same tune pairs are among those with the weakest differences by GAMM analysis (Fig. 11). Most notable among these poorly distinguished pairs are (HLL, HLH), (HHH, HHL), and (LHL, LLH). There are the same tune pairs that failed to be distinguished in the Cole *et al*. (2023) findings.

Tunes in specified emotion contexts were more widely distributed across clusters, meaning that a smaller proportion of a tune was assigned to a single cluster, which is taken as an indication of weakened contrasts. This means that many clusters had traces of multiple tunes (<10%), whereas in Cole *et al* (2023) the smallest proportion of a tune in a cluster was 83% and some clusters contained 100% of a tune, which never occurred in the present study. Small but significant differences between tunes were observed for nearly all tune pairs (within-emotion) based on difference GAMMs, which supports the predictions of the AM model, at least at a fine-grain level of analysis. Only one tune pair in one emotion condition showed no significant differences, HLL-HLH in Love (see Figure 9c). With this sole exception in mind, the finding that tunes are distinguished from one another to some degree in every emotion condition suggests that if the speaker emotion is identifiable, most tunes are potentially identifiable based on F0.

The clustering results for trials in the Neutral condition in the present study most closely follows that of Cole *et al*. (2023)—see Table 3: Comparing clustering results, as predicted, since in both cases tunes were imitated in the absence of a dialog providing emotional and pragmatic context. This similarity between the two studies further strengthens the case that to some degree,

variation in the F0 implementation of tunes is due to emotional and pragmatic context. A further observation is that no particular emotion uniformly helped the phonetic realization of all AM model-predicted F0 contrasts, and between-tune differences tend to be comparable between emotions relative to Neutral based on delta-based exploration of the difference GAMM analysis, (see Figure 10c: Bar plot of GAMM deltas relative to Neutral). Overall, the GAMM was able to account for much of the emotion-driven variation observed earlier in the clustering solution, with model predictions showing all eight tunes as distinct, despite the fact that for some tune pairs, contrasts were diminished within emotion conditions.

2H. Conclusion

This study examined the joint phonetic implementation of intonational and emotion contrasts through the lens of eight phonologically distinct tunes, defined by the AM model (Pierrehumbert 1980, Ladd 2008). Due to the broad scope of the study and constrained research resources, there were necessary limitations in the design and analysis of the study which are discussed in detail in Chapter 5 (Discussion). This study found, based on clustering analysis over imitated tunes produced with emotion, that while most of the AM-predicted distinctions among tunes are maintained, certain distinctions are blurred when tunes are produced in the context of a specified emotion. Further, three pairs of tunes that clustered together in the Cole *et al*. (2023) study also tend to cluster together here: (HLL, HLH), (HHH, HHL), and (LHL, LLH). In a finer grain analysis where F0 trajectories are labeled for the tune they were intended to imitate, GAMM results show evidence for all the predicted contrasts across the emotions, with the sole exception being HLL-HLH in the Love condition. More generally, a speaker's emotion, especially in terms of Valence, was observed to influence F0 trajectories, such that specific emotions had general effects that held across tunes.

Additional work is needed to fully understand how tune-emotion combinations are perceived and interpreted, which could shed light on why some tune-emotion interactions were not significant. Knowing more about how listeners evaluate tune-emotion combinations for linguistic meaning would also shed light on the perceived function of tunes. With a fuller understanding of what tunes mean across contexts, it may be possible to construct contexts that are more relevant for the purpose of eliciting tunes that are faithful to the AM model's predictions.

Despite its limitations, which are more fully explored in Ch. 5, the novel methodology of eliciting intonation in rich dialog contexts showed that accounting for variation related to a speaker's emotional portrayal reveals clearer distinctions in the phonetic implementation of intonational contrasts. There was partial, though compelling evidence for the AM model of MAE tunes, as many of the distinctions predicted by the proposed inventory of contrastive tunes were maintained across emotions, though more work is needed to understand how their distinctiveness (acoustic and perceptual) and interpretations might be linked. This question is explored in perception and tune-meaning associations in the following chapters.

Chapter 3: Perception

3A. Introduction

The goal here is to examine whether the presence of emotional variation in the signal impacts the perceptual salience of phonologically predicted differences in the phonetic form (an F0 trajectory). Chapter 1 introduced the phonological formalism, the Autosegmental-Metrical (AM) model (Pierrehumbert 1980, Ladd 2008), which predicts discrete intonational forms, called tunes. Chapter 2 described a production study examining the acoustic distinctiveness of tunes as a function of a simultaneous emotional portrayal, within an F0 trajectory-based analysis of tune-emotion combinations. This raises the question of whether the observed acoustic differences between tunes predict perceptual salience and whether or how tune perception is affected by variation in emotional portrayal.

One possibility is that listeners will be less able to distinguish between tunes when there is more acoustic variation in the signal due to emotional portrayal. Alternatively, emotional variation might not be a problem for perceiving differences between tunes, and could even help, if identifying a speaker's emotion helps listeners to decode the tune. Consider that listeners may discriminate tunes better than even the F0 trajectory-based analysis suggests given emotion, since they are able make use of secondary cues, like intensity (sound pressure level) and timing cues, which were not considered in Chapter 2. If additional or stronger tune distinctions are evident in the perception results (e.g. HLL-HLH) compared to production, it would motivate a broader consideration of how intonation is phonetically encoded beyond F0, which is the core of the AM model. On the other hand, if listeners' perception of tune distinctions correlates with

speakers' production of F0-based distinctions, it would be a demonstration of the AM model's predictive power, and evidence for a continued reliance on F0 by researchers in this area.

The present study joins a recent push to find empirical support for the Autosegmental-Metrical (AM) model's predictions in the speech signal, which has examined nuclear tunes (Cole *et al*., 2023) and bitonal pitch accents (Steffman *et al.,* 2024) in perception and production, and metrical enhancement in production (Steffman & Cole, 2024). Specifically, the present study extends the methodologies of Cole *et al*. (2023)—experimental and analytical—to test the same set of nuclear tunes under the effects of emotional variation. Cole *et al*. (2023)'s experiment employed an AX discrimination task involving trials with a single 2-alternative forced choice between pairs of auditory stimuli on the basis of perceived similarity (e.g. a same/different judgement). The stimuli being judged by participants in Cole *et al.* (2023) were the model tunes that a different set of participants imitated in their production study, a subset of which were used here, as model tunes for the imitation production experiment in Chapter 2. The F0 trajectories for the model tunes illustrate the critical phonological differences which correspond to a distinct underlying tone sequence, as shown in Figure 1 (repeated from Figure 1 in Chapter 2).

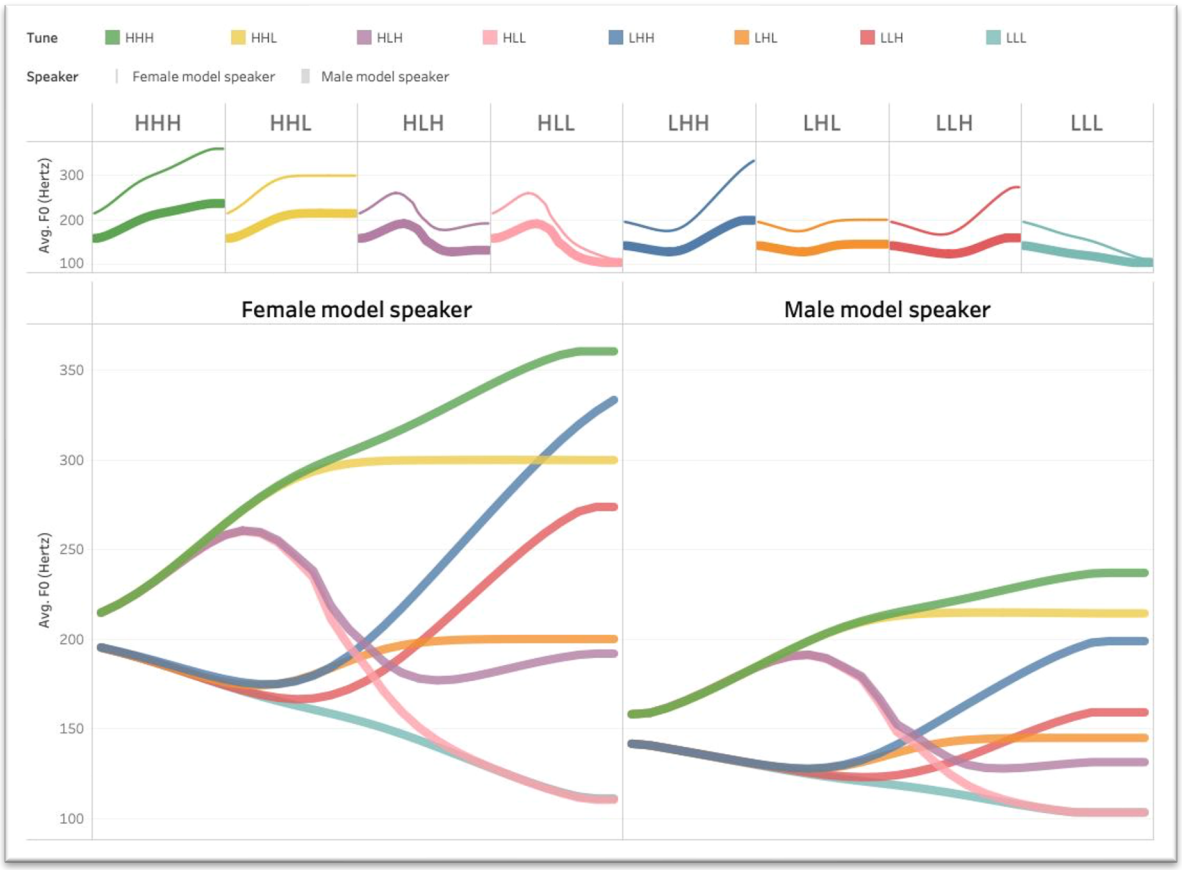


Figure 10: F0 trajectories for model tune stimuli in Chapter 2, which are the basis of the imitative productions that serve as stimuli in the present study.

Because the model tunes were created through F0 resynthesis of the same source recordings, they did not vary in speaker[14], lexical content, or other acoustic factors. This meant that participants did not necessarily have to make a linguistic judgement to succeed at the task; they merely had to detect a difference. Establishing the perceptual distinctiveness of phonetic features is important to understanding the results of imitative production experiments, because it is generally expected that weakly perceived differences will result in weakly produced differences.

[14] Whereas Cole et al. (2023) uses audio stimuli produced by two model speakers (male and female) for their perception experiment, the present study uses one. Importantly, the speaker of both items in a given trial match for both experiments.

In order to understand the degree to which the production results are predictable based on perception results, an overview of the relevant findings from Chapter 2 follows.

Production findings. The tune imitations collected in Chapter 2 were analyzed in terms of their time normalized F0 trajectories, which were submitted to GAMMs and k-means clustering. Both analytical methods showed that tunes tended to be produced with distinct F0 trajectories across emotions, but not all tune differences were found to be robust, as Figure 2 (repeated from Chapter 2: Fig. 8a) suggests. Difference GAMMs provided a fine-grained analysis of F0 differences for all tune emotion combinations and showed that only one tune pair, HLH -HLL, showed no significant differences in their F0 trajectories across emotion conditions.

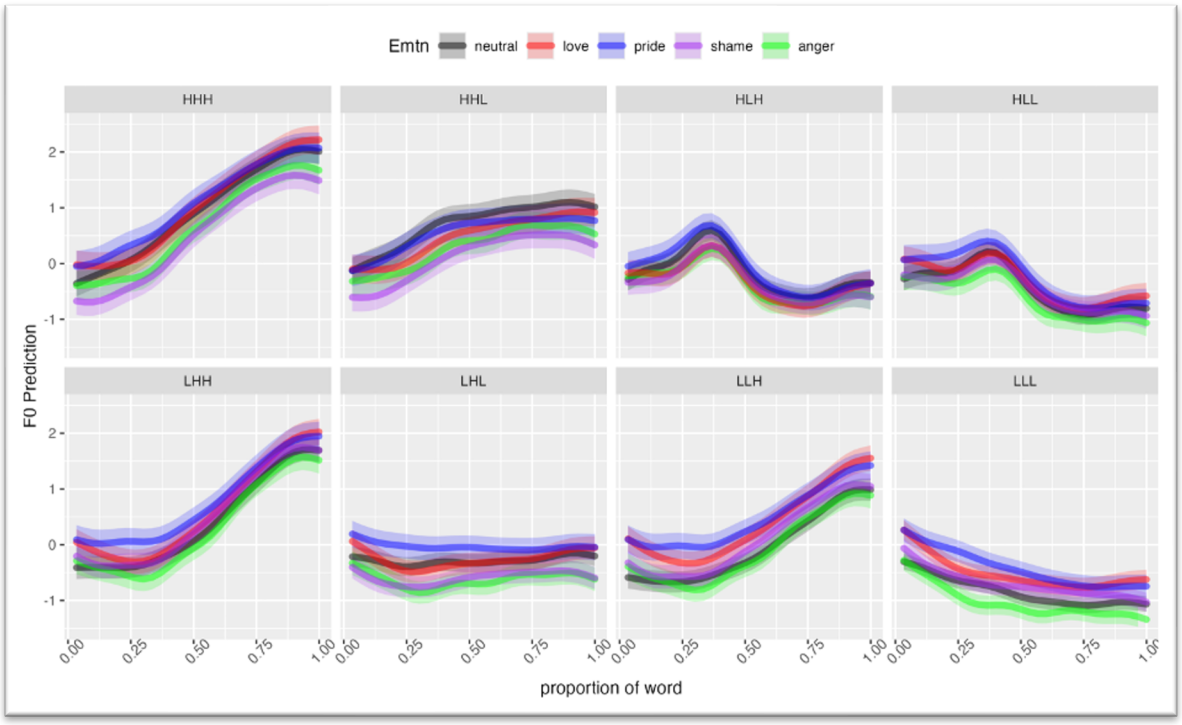


Figure 2: GAMM predictions for all tune-emotion combinations. Repeated from Ch. 2: Fig. 8A.

The clustering solution, visualized by tune in Figure 3 below (repeated from Ch. 2), shows the F0 trajectories in these data are optimally partitioned in five clusters, suggesting a smaller emergent set of tune shapes than predicted by the phonological tune labels of the auditory stimuli being imitated. As in the GAMM analysis, imitations of HLL and HLH tended to be clustered together (mainly in Clusters 'D' and 'A'), but other tune pairs tended to merge as well, such as HHH-HHL (Clusters 'E' and 'B'), and LHH-LLH (Clusters 'B' and 'C'). Interestingly, these are also tune pairs that were merged based on the clustering solution by Cole *et al*. (2023), although LHL-LLH also merges for Cole *et al*. but not in Ch. 2's findings. In general, tunes were also more broadly distributed across clusters compared to Cole *et al*.'s results; their tunes composed 83 to 100% of clusters while the Ch. 2's tunes composed >10 to 84%. One cluster (B) even had >10% of five tunes—indicating that more than half of the tune inventory partially merged under the influence of emotional variation (see Table 3 in Chapter 2 for full breakdown).

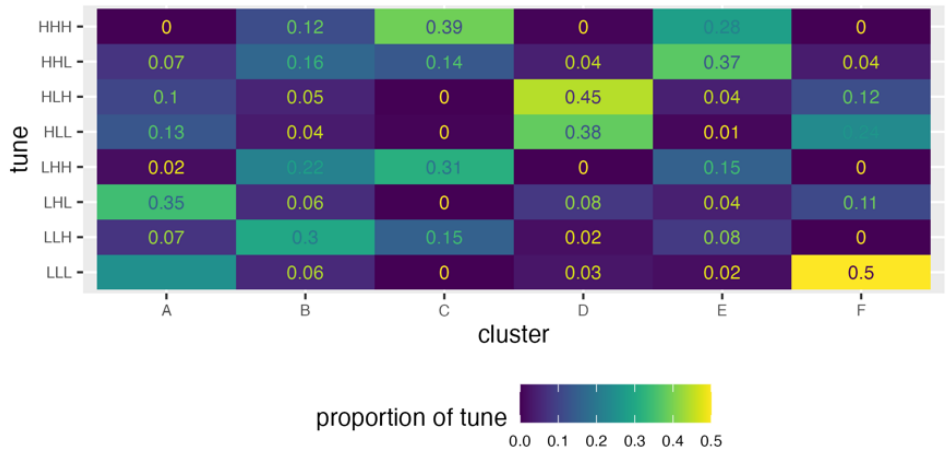


Figure 3: Clustering solution for tune-emotion combinations (repeated from Chapter 2: Fig. 7a)

Given that Cole *et al*. (2023) found strong parallels between their production and perception studies, similar results might be obtained in comparing the results of the present study to those from Chapter 2. If so, then the tune pairs that are produced with similar F0 trajectories will also be perceptually proximal, based on AX discrimination judgements. That said, it is possible that judging tune distinctions in stimuli that also vary by emotion will pose a considerable challenge, resulting in a weaker correlation with the production results. The possible outcomes of the present study and these comparisons is considered next.

3B. Hypotheses & predictions

This study tests the hypothesis that listeners can perceptually discriminate tune contrasts produced within any type of emotional context equally well, including Neutral. Under this view, listeners can make use of information in the speech signal about speaker emotion to adjust their perceptual expectations about F0, facilitating the accurate interpretation of intonation. For example, it would benefit a listener debating whether a tune is HHH or HHL (which differ mainly in the F0 maximum at the end of the trajectory) to know if the speaker is expressing Love (boosts F0 based on Chapter 1 GAMMS) or Anger (lowers F0). Taking a strong version of this hypothesis, it is predicted that emotion has no impact on tune discrimination; tunes will be discriminated with similar accuracy across speaker emotions, including Neutral. Under a weaker version of this hypothesis, the specific type of emotional context may play a role, such that listeners' ability to discriminate between tunes will partly depend on the speaker's emotion. Under the weak hypothesis, it is predicted that the emotional context plays a role in determining tune distinctiveness, for example, such that tune discrimination accuracy would be higher under Neutral than Love (or vice versa). If the strong hypothesis holds, then results should resemble Cole *et al*. (2023)'s perceptual discrimination results, as their materials were designed to be emotionally invariable.

3C. Experimental methods

The methodology of Cole *et al*. (2023)'s perceptual study was adapted for the present study in order to test the effects of emotional variation on the perceptual salience of distinct intonational tunes, using an AX discrimination experimental paradigm. The experiment was conducted using OpenSesame (Mathôt *et al*., 2012) deployed on a private server running JATOS (Lange *et al*., 2015). This combination of software allowed for multiple remote participants to simultaneously complete the study. The code and files to run the experiment along with raw results are freely available through the study's Open Science Framework repository (https://osf.io/gbk8z/, see Experiment #2).

Task design

AX discrimination involves a 2-alternative forced choice for a pair of stimuli (A and X) given some criteria, such as a same/different judgement. In the present study, stimuli vary along two dimensions, tune and emotion (unlike in the Cole *et al*. study, where stimuli varied by tune only), so participants were instructed to focus their attention and judgment on the tune. This is accomplished by reminding participants to consider what the speaker is trying to say, which is presented in different ways, starting before the experiment with the consent form, which reads: "Your job will be to decide whether pairs of words are being said in such a way as to convey the same meaning, while ignoring factors like the speaker's mood." After starting the experiment, they receive the following instructions which are intended to reinforce this concept:

- This experiment is about the meaning that words convey when said a particular way.

- You will hear pairs of the same 4 names said in many ways by the same speaker and judge them as having the same or different intended meaning.
- Note that the names (Melanie, Madelyn, Lavender, and Gallagher) do not convey an inherent meaning, other than to identify a person.
- However, if said in a specific way, words can convey a variety of different meanings.
- In written language, some (but not all) of these meanings can be signaled by punctuation:
    - That's Mary.
    - That's Mary?
    - That's Mary…
- Your job is to decide whether the speaker is trying to say the words in order to convey the same meaning, while ignoring other factors such as the name and the speaker's expression of emotion in their voice.
- You will answer the question "Is the speaker trying to say the words in the same way, or a different way?" using the keyboard.

The objective using these instructions was to define an intonational difference as a speaker intentionally modulating their voice to change the received meaning, specifically the kind of meanings associated with statements, questions, and holding the floor (e.g. pragmatic). Moreover, participants are repeatedly instructed to ignore the speaker's "mood" and "expression of emotion", which is expected to further focus their attention on phonologically relevant features.

Allowed responses were "Same" (Left Shift key) or "Different" (Right Shift key). The trials timed out after 6 seconds if no response was recorded, which was deemed appropriate based on early piloting. Participants (N=3) with >10% timeouts were excluded, as this indicated a lack of attention to the task. Due to the need to test all possible combinations of tune and emotion in all orders, the number of unique trials exceeded the reasonable length of an experiment (8^2 tunes x 5^2 emotions = 1,600 possible combinations of tune and emotion pairs). Therefore, the experiment

adopted a between-participants design, whereby each participant was assigned 80 trials. Each experimental session was about 20 minutes long.

Participants (N=153)

From the online crowdsourcing platform Prolific, 160 eligible participants were recruited through selective filtering. Specifically, they had to meet the following criteria based on self-report:

(i) L1 speakers of Mainstream American English (MAE) who used it as their primary language but did not have to be monolingual. For multilingual speakers, MAE was acquired first or simultaneous with other languages.

(ii) Between 18 and 65 years of age.

(iii) No reported hearing, speech, or language processing related deficits.

Exclusions consisted of seven participants: three had an overly high proportion of trial timeouts (>10%), two encountered technical problems and could not complete the experiment, and two did not me*et al*l eligibility criteria based on our post-experiment questionnaire. Individuals were only recruited into a single study; no participants from the imitative production study in Chapter 2 also contributed to the present study.

Materials

From the production study with tune and emotion (Chapter 2), 40 recordings were selected to represent all tune-emotion combinations from a single speaker, a female voice actor. The main

criteria for including a particular recording in the inventory was faithfulness to the F0-resynthesized model tunes that were imitated for the tune-emotion production study (and used by Cole *et al*. (2023) for production and perception). Comparisons were done by visually inspecting F0 trajectories of stimuli (see Figure 4) to ensure AM-predicted differences (reflective of an underlying tone sequence) were present in the speech signal. The emotional qualities of the stimuli were not independently assessed but sounded natural and conversational according to members in our research group based on early piloting.

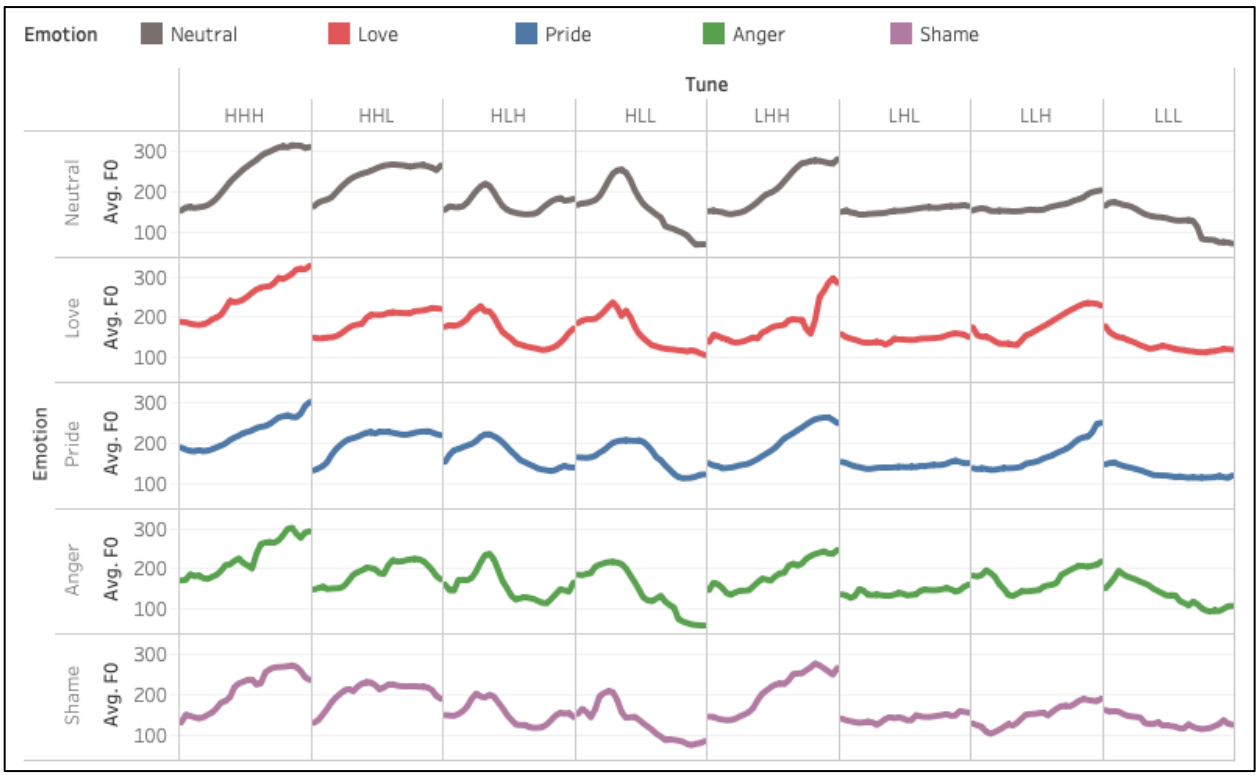


Figure 4: F0 trajectories for experimental stimuli (target word only) by tune and emotion. Recordings were naturally produced (not F0-resynthesized) by a single speaker (a female voice actor) in the preceding production experiment. Colors for emotions match Figure 3 to aid comparison between GAMM generalizations and F0 trajectories for the present study's stimuli.

To quantify the differences between the F0 trajectories of every stimulus pair, the root mean squared deviated (RMSD) was calculated, shown in Figure 5 where each dot represents a tune pair under a particular emotion. It appears that Pride is the emotion condition in which RMSD measures for tune pairs is frequently the lowest (indicating weaker distinctions), while Neutral and Love are the emotion conditions in which RMSD for tune pairs were usually the highest (indicating stronger distinctions). To fully understand how these differences predict the perception results, an upcoming plot compares these values to model estimates of perceptual discriminability.

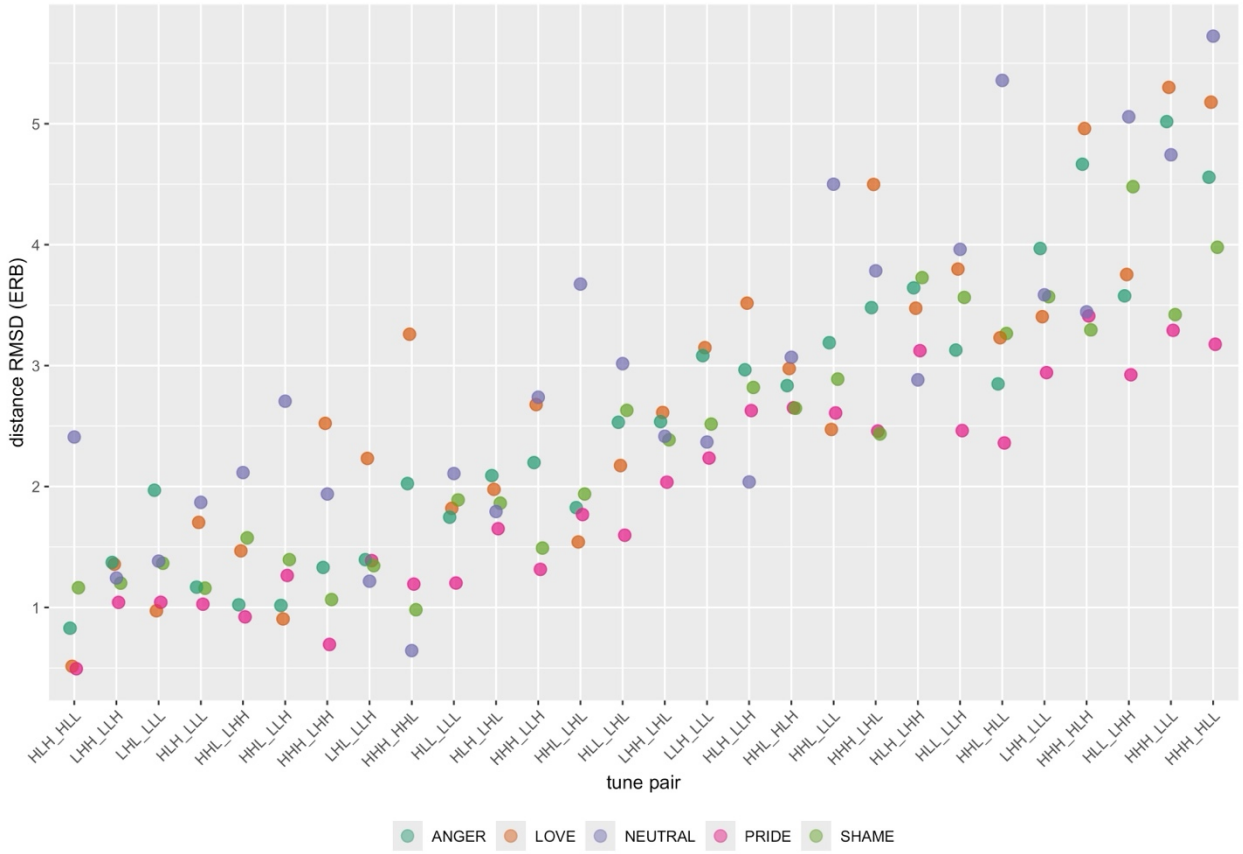


Figure 5: Root mean squared deviation (RMSD) distance between all stimuli, ordered by ascending mean RMSD, coloration by emotion.

The lexical content of sentences that were produced varied depending on the tune and emotion, but each emotion had the unique target word (see Table 1). Target words were all trisyllabic proper nouns with the same lexical stress pattern, to minimize the effect of segmental and stress-related differences between words/emotions. To help focus participants on the tune, target words were presented in isolation ('Marilyn' from "Her name is Marilyn", for example). Each target word (40 total) was manually excised from its carrier sentence using Praat, based on acoustic landmarks of perceptually and visually salient phone boundaries.

Table 1: Tune-Bearing Lexical Items

Emotion	Lexical item
Pride	Melanie
Love	Madelyn
Shame	Lavender
Anger	Gallagher
Neutral	Marilyn

Procedure

Preparation. Before the experiment session begin, participants were asked to choose a quiet and distraction-free time and place to complete the experiment on their computer. Additionally, they were instructed to use comfortable headphones for the whole session. Participants first completed our consent form, followed by a short equipment check to ensure that their headphones were of sufficient quality. The sound check procedure was adapted from Morrison *et al*. (2022) and involved counting steady tones at different F0 and amplitude levels. After

answering the correct number of tones, each participant was shown the instructions in the form of a slideshow.

Instructions. The purpose of the instructions was to convey the goals, keyboard controls, and flow of the study. As explained in the *Experimental task* section, the goal of the experiment was defined in terms of intentional choices a speaker can make to affect the received meaning of a word, ignoring factors like mood and emotion. After explaining the task, participants are shown an illustration of the keyboard controls as well as a sample trial screen with the experimental prompt ("Is the speaker trying to say the words in the same way, or a different way?") and key mappings for "Same" and "Different" (see Figure 6: Sample trial screen), which never changed. To help participants understand how to perform the task, the last part of the instructions involved listening to six example pairs using preselected stimuli produced by a different speaker than the one for the main part of the experiment (a male)[15]. After each pair in the instruction block, an onscreen message revealed whether the stimuli had the same linguistic function (tune) or not—these possibilities were equally balanced like in the main part of the experiment.

[15] The "Same" trials were {HHL_HHL under Love_Pride}, {LHL_LHL under Shame_Pride}, and {LHL_LHL under Pride_Anger}. The "Different" trials were {LHL_HHH under Anger_Pride}, {LLH_HHL under Love_Love}, and {HHH_HHL under Pride_Pride}. Since "Same" trials were not analyzed and most "Different" trials had mismatching emotions, the Emotion Match Model (EMM) was consulted for differences between these stimuli. The discrimination accuracy for these combinations was similar to others; HHH_HHL under Pride was above chance while LLH_HHL crosses chance. See Appendix C for model estimates with combinations of interest highlighted. Overall, it appears that tune discrimination was not impacted by the content of the practice trials.

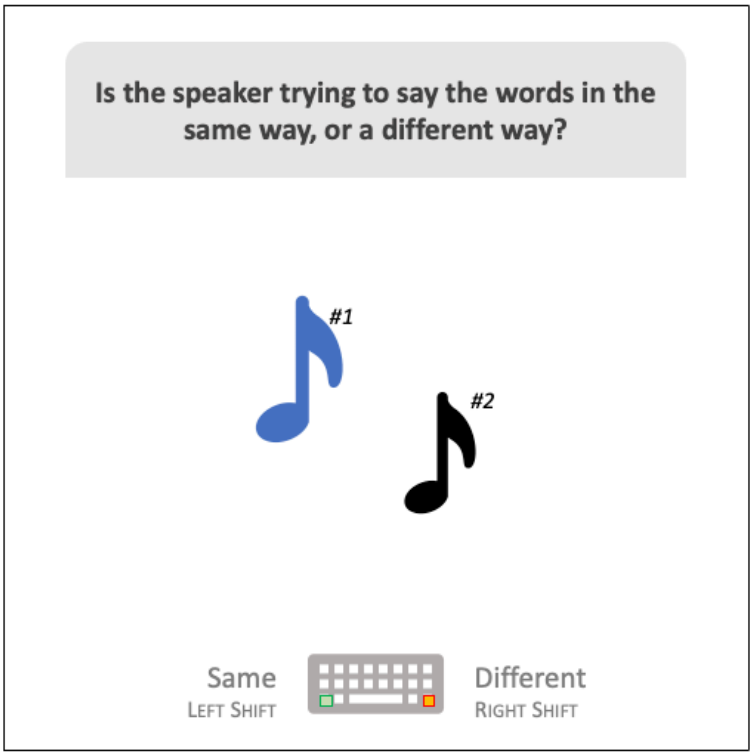


Figure 6: Sample trial screen showing prompt, stimuli icons, and controls. When the two stimuli play, their respective music note is visually highlighted.

Practice trials. To give participants experience using the controls, a four-trial practice block that looked the same as the main task was presented. Stimuli for practice trials came from the same speaker as critical trials (female), but consisted of different recordings[16], to help listeners perceptually calibrate to her vocal properties. These practice trials were identical to critical trials, including lack of feedback, and proceeded as follows:

1. Visual layout of the trial appears onscreen, consisting of two musical note icons to represent the stimuli (A and X), their relative progress ("Practice N of 80" onscreen message), and a reminder about how the keys map to responses (Left Shift = Match; Right Shift = Mismatch). To preview, the 'Practice' text changes to 'trial' for the main part of the experiment.

[16] In the production study, voice actors produced five versions of each tune-emotion combination, so after selecting one version of a tune-emotion combination, there were multiple others to choose from to construct the practice trials.

2. The participant presses a key to start playback of the first stimulus, which is accompanied by the left music note being visually highlighted (see Figure 6).
3. 500ms after the first stimulus finishes, the second stimulus automatically starts playing, which is accompanied by the left music note returning to normal and the right one being visually highlighted.
4. After the second stimulus finishes, the participant has a 5000ms window to respond. When a response is recorded, or the window elapses, the experiment continues.

Critical trials. After the last practice trial, the progress message changed to the form "Trial N of 80" so participants knew they had started the main trial block.

Questionnaire. After the last critical trial, the participant was redirected from the experiment to a short Qualtrics-based questionnaire to confirm their eligibility, to avoid total reliance upon information entered into Prolific.

3D. Quantitative Methods

Data preparation

In order to submit the experimental data to the statistical models, each trial was coded according to response (1 if "different", otherwise 0), tune pair (multilevel factor, deviation coded, order insensitive, identical to Cole *et al*. 2023), emotion pair (multilevel factor, dummy coded, order insensitive), and participant ID (dummy coded factor). In models where the tunes or emotions always match the label is simplified to the tune or emotion label (e.g. 'HLH' versus 'HLH_HLH' or 'Neutral' versus 'Neutral_Neutral'). The probability of tunes matching was fixed at 50% across the experiment, while the probability of emotions matching was 25%. While it was desirable to balance the matching status of tunes to make both the 'same' and 'different' responses equally likely to be correct, the matching status of emotions was left unbalanced for two main reasons. First, it would have made the experiment much longer in terms of the number of total trials to satisfy such a design, which is problematic for a perception experiment since participants may become exhausted. Second, each emotion is conveyed by a particular lexical item, so emotion matching status is transparent to the participant and would be a potential distraction. In order to better understand how matching status of emotions may have impacted the results, an Emotion Match Model (EMM) is run over trials where tunes mismatch but emotions match. Additionally, a Tune Match Model (TMM), where tunes match and emotions mismatch, and a Tune Emotion Interaction Model (TIM), where tunes and emotions mismatch, will be run—see Table 3 for summary. Note that the only untested subset is cases where both the

tunes and emotions match, which for our materials would mean testing whether participants can detect whether recordings are the same, which is likely to be at ceiling[17].

Table 2: Subsets for modeling [proportion of total trials]

		Tunes	
		Match	**Mismatch**
Emotions	**Match**	(Unmodeled) [~10%]	**Emotion Match Model (EMM)** [~10%]
	Mismatch	**Tune Match Model (TMM)** [~40%]	**Tune Emotion Interaction Model (TIM)** **Tune Only Model (TOM)** [~40%]

GLMM implementation

Using the generalized formula given for each model (introduced below) and the subset of trials shown in Table 3, each model was fit using a Bayesian generalized linear mixed model (GLMM) implemented using the R package 'brms`. The GLMM was run using two chains with 10,000 iterations each and first 1000 'warm up' iterations were discarded. The same weakly informative priors as used in Cole *et al*. (2023) was used here for the intercept and fixed effects ([Normal(0,1)]). Models had random intercepts by participant.

[17] This was later confirmed by looking at the mean accuracy for when tunes and emotions match, which was 99.3% (see Table 3 in Results).

Emotion-Matching Model (EMM): Response ~ Tune pair x Emotion pair + (1|Participant)

The purpose of this model is to quantify differences in tune discrimination across emotion conditions, relative to Neutral, for the EMM data (emotions match; tunes mismatch). This is achieved by using data where emotions match and tunes mismatch and setting the reference level for emotion pair on Neutral_Neutral. A strength of this model is that it only considers pairs of stimuli that have the same word (since each emotion condition uses a unique lexical item, see Table 2), which might make comparing recordings on the basis of tune easier. This model will show how tune discrimination varies as a function of specified emotion, which will be especially informative for H2 (testing for a perception-production correlation) with Chapter 2's data as a point of reference. The reasoning is that the data analyzed in this model (mismatched tunes, matched emotion) will include trials where mismatched tunes are incorrectly identified as "Same" driven by the perceived similarity in emotional portrayal. The EMM will be sensitive to any differences in the relationship between the perceptual discrimination of tunes and their differences in production that are due to emotion. Lastly, beyond the present study, this model will be interesting to compare to Cole *et al*. (2023)'s findings, since their study was in essence matched by emotion along the lines of this study's Neutral_Neutral condition.

Tune-Matching Model (TMM): Response ~ Emotion pair + (1|Participant). The purpose of this model is to quantify the degree to which different emotions drove participants to misclassify matching tunes as different for the TMM data (tunes match; emotions mismatch). To that end, the data for the model includes trials where the tunes match but the emotions do not. The

specification does not include a term for tune because there is no prediction about how particular tunes might be affected by emotional variation. Rather, the present study is focused on how emotional variation, modeled here over pairs of stimuli, impacts tune discrimination in general. If Hypothesis H1a bears out, then responses will be more accurate in the Neutral condition compared to the specified emotions because it should be easier to detect a tune match without the presence of emotional variation.

Tune-Emotion Interaction Model (TIM): Response ~ Tune pair x Emotion pair + (1|Participant). This specification includes main effects of Emotion in addition to all two-way interactions between Tune pair and Emotion pair for the TIM/TOM data (tunes and emotions mismatch). This is a model with a large number of interactions (28 tune pairs ×10 emotions pairs (280 total interaction terms), which allow us to understand how specific modes of emotional variation impact specific tune pairs.

Tune Only Model (TOM): Response ~ Tune pair + (1|Participant). The purpose of this model was to evaluate the outcome of emotional variation on the task by tune pair, which is pertinent to testing H1 (general effect of emotional variation on accuracy). The data for this model included all trials where both tunes and emotions mismatched. The null hypothesis is supported if the TOM fit is equivalent (or superior) in fit compared to the TIM, since it would indicate no overall impact of emotional variation.

3E. Results

Empirical means by matching status/data subset

The mean response rate across all different-tune trials (e.g. correct rejection accuracy) is 72.4%, which is far above what would be expected given chance (50%). Table 3 below gives the mean rate for each subset of the data based on the matching status of tune and emotion, complementary to Table 2. The EMM data was found to have a mean rate similar to the overall rate, while the TMM data was markedly lower (63.1%), and the TIM/TOM data was slightly higher (75.0%). Next, heatmaps showing the breakdown of the response rates within each subset are presented.

Table 3: Mean response rate by matching status / modeling subset

		Tunes	
		Match	**Mismatch**
Emotions	**Match**	(Unmodeled) 98.9%	**Emotion Match Model (EMM)** 73.0%
	Mismatch	**Tune Match Model (TMM)** 63.1%	**Tune Emotion Interaction Model (TIM)** **Tune Only Model (TOM)** 75.0%

Empirical means by tune pair within emotion (EMM data)

The following heatmaps show the calculated empirical mean within trials where the stimuli match in emotion but not tune (1 = correct). In Figure 7, the means for tune pairs (x-axis) are broken down by emotion (y-axis). A minority of tune pairs were highly variable by emotion,

such as HLH-LHH, which was accurately discriminated under Shame and Anger, at chance under Neutral, and below chance for Pride and Love. Usually, the response rates for a particular tune pair look uniform across emotions, exemplified by HHH_LLL (highly accurate) and HHH_LHH (accuracy low except at-chance under Anger).

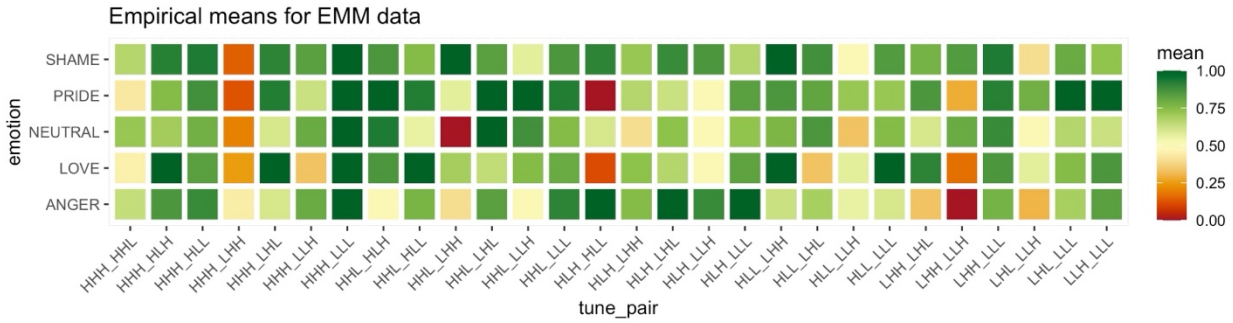


Figure 7: Empirical mean response rate of tune pairs and emotion pairs where tunes mismatch but emotions match (EMM data)

Figure 8 shows the same data aggregated in two ways: over tunes to show means by emotion (left pane) and over emotions to show means by tune pair (right pane). For tune pair (and other tune and emotion pairs in following figures) each member of the pair has a dedicated axis; in this case the first tune is on the y-axis and the second is on the x-axis. To summarize, it appears there is little difference between emotions (Shame is slightly more accurate) and one tune pair that was far below chance (HHH-LHH). Other tune pairs had response rates near chance, and they all had LLH in common: HLL-LLH, LHH-LLH, and LHL-LLH. The most accurately discriminated tune pair was HHH-LLL.

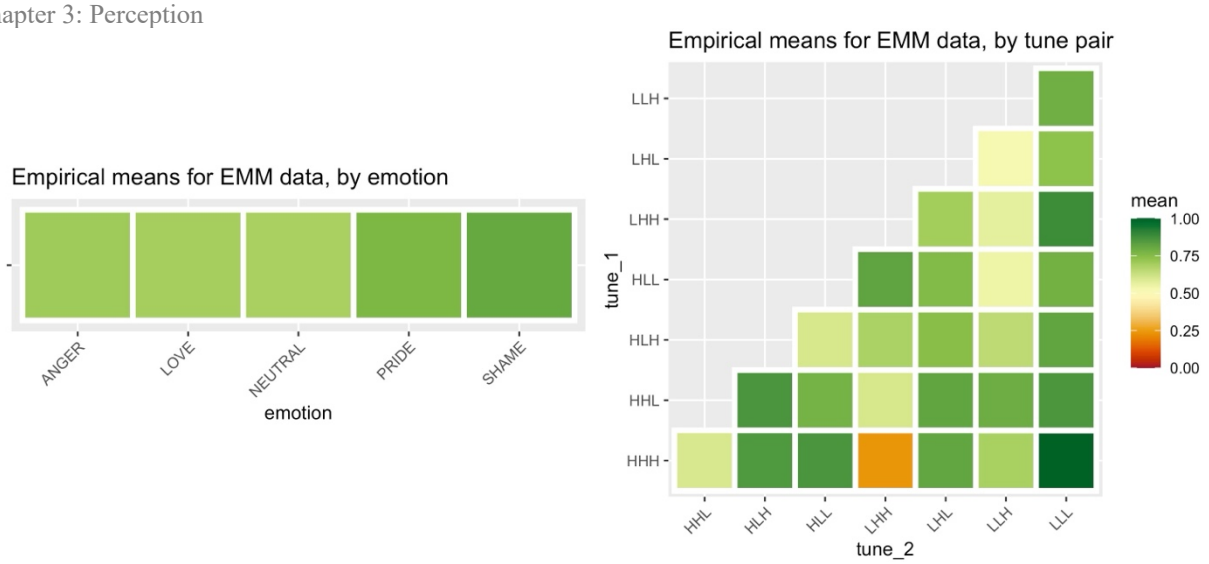


Figure 8: Aggregated empirical mean response rate of tune pairs and emotion pairs for data where tunes mismatch but emotions match (EMM data), calculated over tune pairs/by emotion (left pane) and over emotions/by tune pair (right pane).

Empirical means by emotion pair within tune (TMM data)

The following heatmaps show the calculated empirical mean response rate within trials where the stimuli match in tune (0 = correct) but not emotion. In Figure 9, every tune is at or below chance for at least one emotion pair, and near ceiling accuracy (again, 0 in this plot) for at least one emotion pair. This is taken as an early indication (to be confirmed by modeling) that responses varied extensively within tunes, depending on the type of emotional variation.

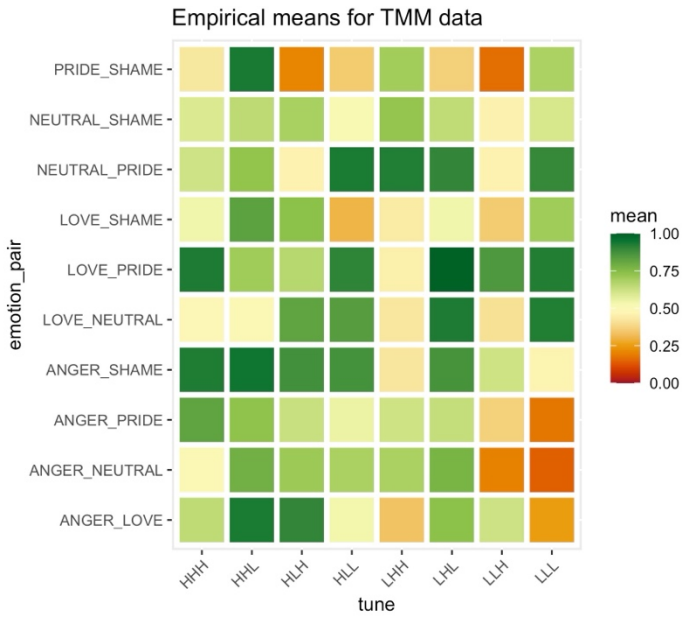


Figure 9: Empirical mean response rate of tune pairs and emotion pairs where tunes match and emotions mismatch (TMM data), so "different" judgements may be based on the emotion rather than the tune pair (errors).

Just as Figure 8 aggregated over results shown in Figure 7 for the EMM, Figure 10 aggregates over results shown in Figure 9, this time over emotions to show mean response rates by tune (left pane) and over tunes to show means by emotion pair (right pane). Recall the correct answer in this plot is zero (red on the color scale). The most accurately discriminated tune based on this plot (left pane) was LLH and the least accurate was HHL. In the aggregated results by emotion pair (right pane) a generalization emerges where like-valence emotions were more often judged as the "Same", which is apparent because negative emotions (Anger, Shame) paired with positive emotions (Love, Pride) or Neutral were darker. One available interpretation given this outcome is that listeners falsely perceive differences due to emotional valence.

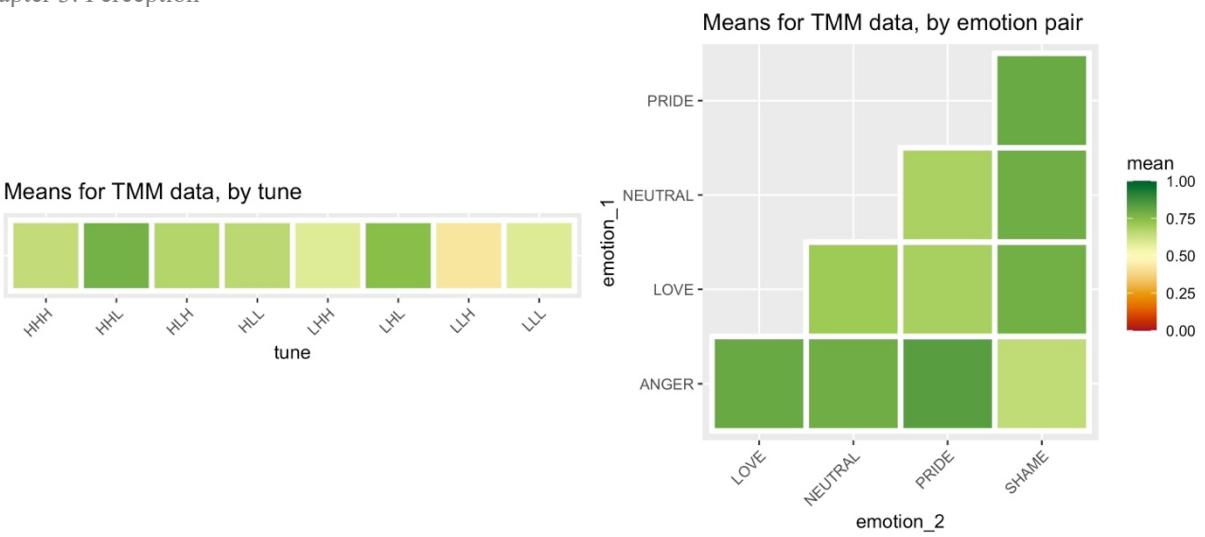


Figure 10: Aggregated empirical mean response rate for data where tunes match and emotions mismatch (TMM data), calculated over emotion pairs by tune (left pane) and over tunes by emotion pair (right pane).

Empirical means by tune pair and emotion pair (TIM/TOM data)

The following plots show the calculated empirical mean response rate within trials where the stimuli mismatched in tune and emotion (1 = correct). Figure 11 shows the mean response by tune pair (x-axis) and emotion pair (y-axis). Based on this figure, although some emotion pairs were particularly challenging (such as Anger-Shame, which was around chance for many tune pairs), there were few cases where accuracy was severely affected. HLH_HLL under Neutral_Pride and LLH_LLL under Neutral_Shame are the only cases with a mean response below 25% (red) according to Figure 12, while many combinations are above 75% (dark green). Accuracy appears slightly higher overall for this subset compared to when emotions matched (EMM; Figure 7), in line with observations in the empirical grand means of these subsets (Table 3), suggesting that the EMM subset of data likely captures the important differences in emotional variation in this task. Moreover, when comparing tune pairs between EMM and TEM, accuracy appears lower for some but not all cases (e.g. HHH_LLL).

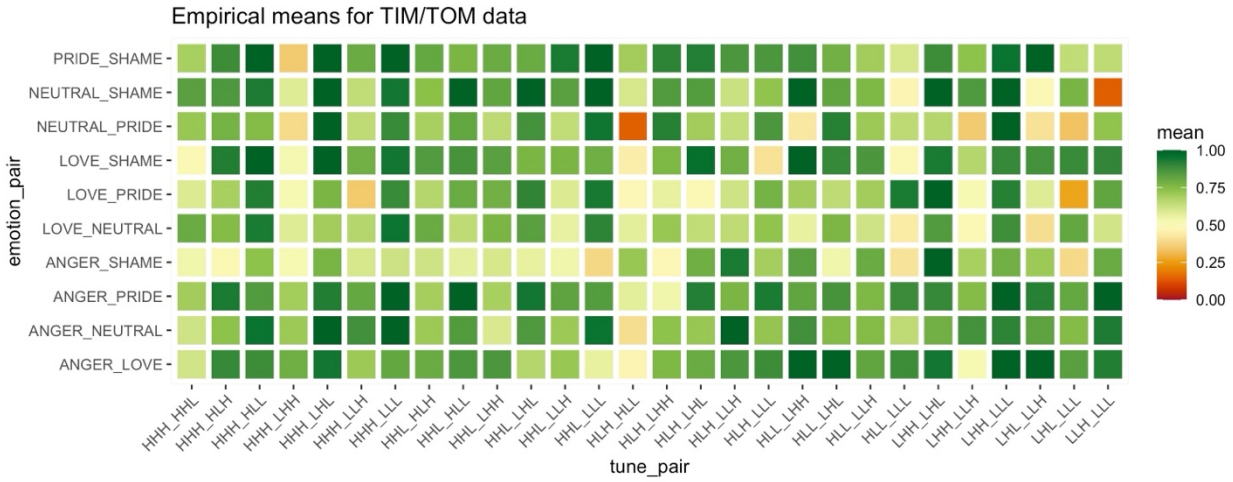


Figure 11: Empirical mean response rate for the data where tune and emotion pairs mismatch (TIM/TOM data).

As was done for the EMM and TMM in Figures 8 and 10 respectively, the next step is to aggregate the empirical mean response rates for the TIM/TOM data shown in Figure 11, which is done in Figure 12 by tune pair (over emotion pairs; left panel) and by emotion pair (over tune pairs; right panel). Several tune pairs, but no emotion pairs, tended to be classified at or below chance (50%). The worst-discriminated tune pairs were HLH-HLL and HHH-LHH. In comparison, accuracy by emotion pair is less variable, with most emotion pairs showing accuracy well above chance. Additionally, the same valence effect noted in the TMM data emerges in the averages across tunes (right pane).

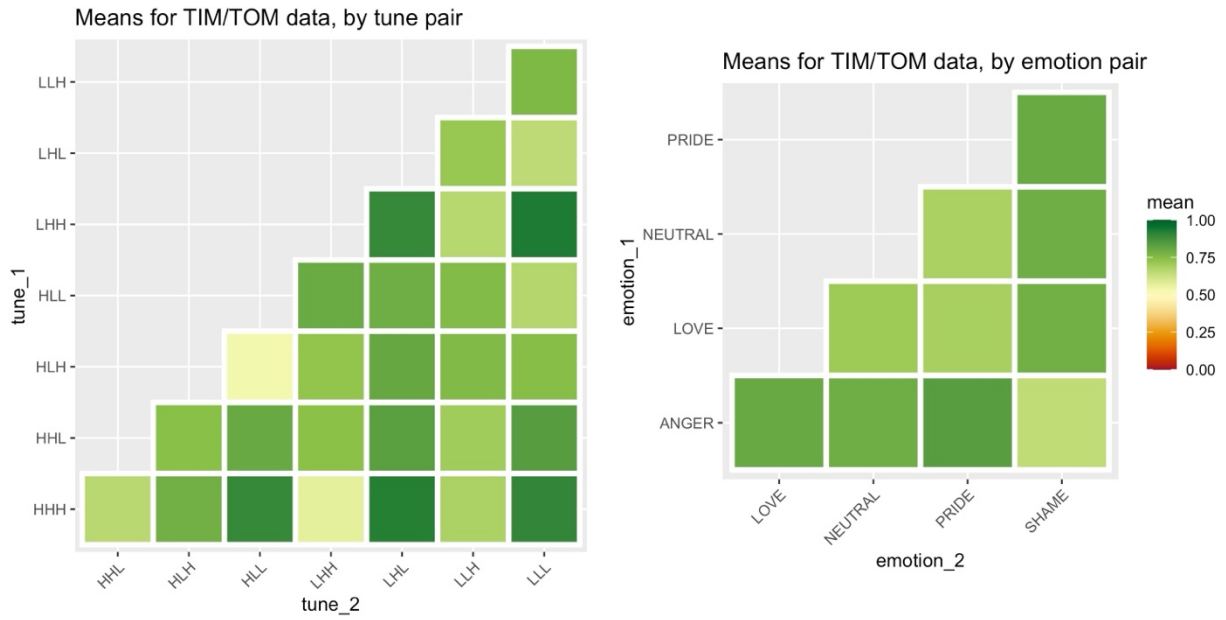


Figure 12: Empirical means of response rates when the stimuli differed in tune and emotion (TIM/TOM), over emotion pairs by tune pair (left panel) and across tune pairs by emotion pair (right panel).

Modeling overview

The advantage of analyzing these data with a statistical model is we can explore all relevant tune-emotion interactions while utilizing a random effects structure by participant to account for individual-level variation. For all GLMMs, R-hat was examined to confirm convergence and results are reported in terms of posterior median estimate in terms of probability. All models converged based on their R-hat being equal to 1, which can be found in the model outputs in each model's associated appendix. Direct model comparisons are made using Pareto smoothed importance sampling (PSIS) leave-one-out (LOO) cross validation to compare fits, specifically the LOO information criterion (LOOIC), which is implemented in the `loo` package (Vehtari *et al*., 2017). Lower LOOIC scores indicate better fits, and for each model below the LOOIC and its standard error are reported.

Cross Validation (LOOIC)

Table 4 below gives the LOOIC results for each model, which shows the following cline in fit quality: TOM > TVM > TEM. Cole *et al*.'s (2023) model was evaluated the same way and included in the table for reference (as 'C23') and proved to be a much better fit to its data compared to the models from the present study. Although TEM had the lowest LOOIC (optimal) it had the greatest variation, TOM had the lowest SE. Note that, due to the types of trials they include, EMM and TMM are not comparable to these models in terms of LOOIC. Specifically, EMM has fewer observations and is expected to include more emotion-conditioned variation, and TMM does not include a tune term, which means it has fewer degrees of freedom than the models in Table 4.

Table 4: LOOIC Estimates & Standard Errors by Model

MODEL	Estimate	SE	Best fit?
C23	2057.323	66.37	*N/A*
TOM	4925.857	71.181	No
TEM	4760.005	76.692	Yes

Emotion Match Model (EMM) results

For the full model output, see Appendix 3A. Recall that one of the benefits of examining cases where emotions match is it makes interpreting differences based on emotional variation straightforward, compared to cases where emotions mismatch (all other models). The model output is shown in Figure 13, which displays the predicted likelihood of a "Different" response

and credible interval for each emotion condition. A credible effect is one where the credible interval does not cross chance. To summarize Figure 13, the cline of response accuracy rates by emotion was: Anger = Love = Neutral < Pride < Shame.

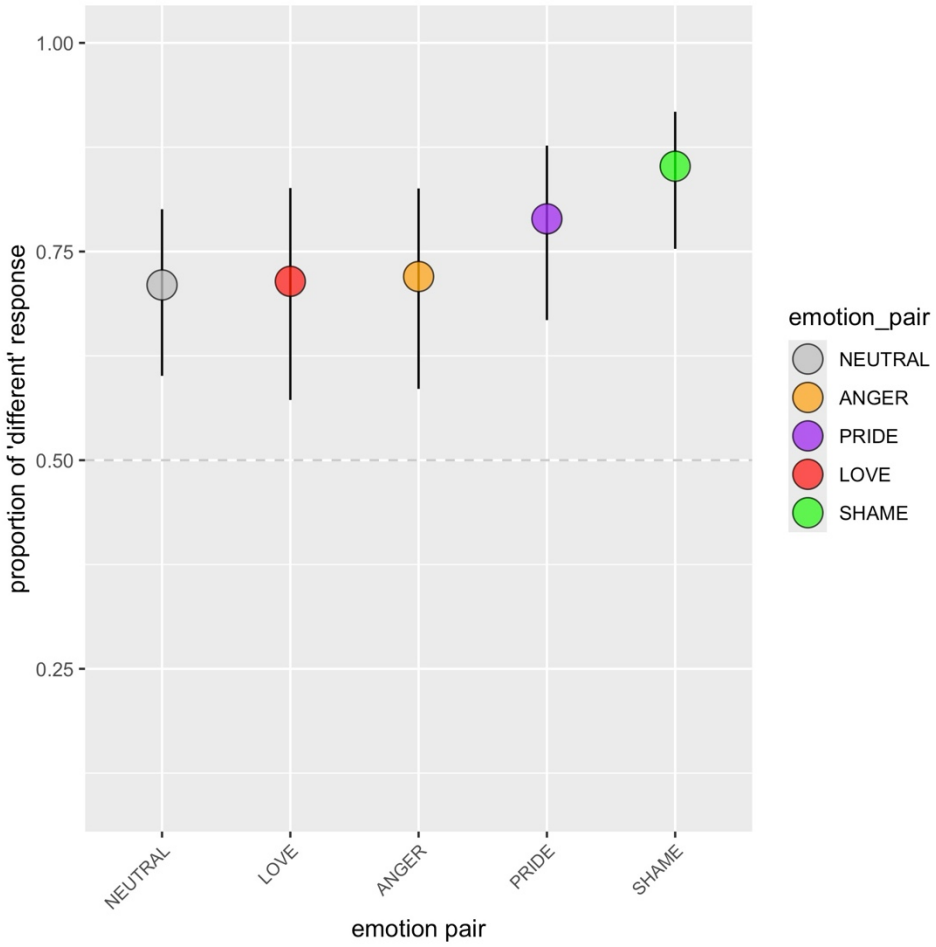


Figure 13: Estimated proportion of "different" (correct) responses in the EMM by emotion.

Figure 14 plots the discrimination accuracy for each tune pair as predicted by the EMM model. The estimated interval of the mean accuracy for one tune pair was below chance, HHH_LHH, indicating a response bias for a Same response for this tune pair. 11 tune pairs had estimated mean accuracy intervals spanning chance, while for 16 tune pairs (roughly half of the total set)

the mean accuracy interval was solidly above chance. The tune pair for which the model estimates the highest discrimination accuracy was HHH_LLL, with estimate accuracy near ceiling. Setting aside the below-chance tune pair, HHH-LHH, the model estimated discrimination accuracy varies more-or-less continuously across the tune pairs, so no evidence for a set of especially accurate or inaccurate tune pairs clearly emerged.

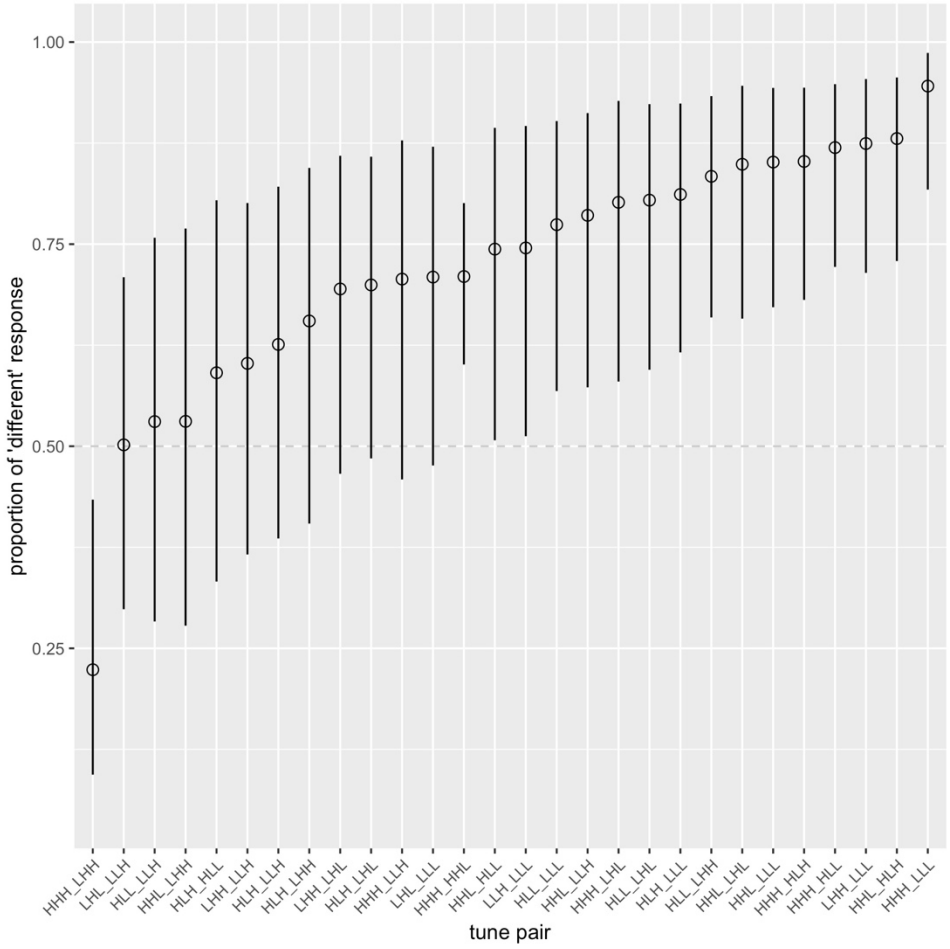


Figure 14: Estimated proportion of "different" (correct) responses in the EMM by tune pair.

Figure 15 compares the estimates from the EMM to those of Cole *et al*. (2023), to examine whether results from these related studies converge. Each tune pair is color-coded for the emotion (the reader is reminded that in the EMM model, the emotion was the same for both tunes in the pair) with trend lines for each emotion from the linear regression model (with reference level as Neutral). Table 5 shows the R^2 value and p-value which indicate the strength of correlation (higher value → stronger correlation) and statistical significance respectively. All regression slopes were significant based on their p-value, and Neutral showed the strongest correlation with Cole *et al*. (R^2 = .344) while Anger showed the weakest correlation (R^2 = .216). The first observation is that the regression lines relating data from the two studies are mostly above the x=y line positive, which means that the estimated discrimination accuracy is generally higher in the present study. That said, the slope of the regression lines for each emotion condition in the present study is <1 (the slope of the x=y line), indicating that the higher discrimination accuracy for tune pairs in the present study is driven more by tune pairs that are poorly discriminated in the Cole *et al*. study, but fare better in the present study (data in the upper left quadrant). Overall, most tune pairs have high discrimination accuracy as estimated by the models in both studies (i.e., most of the data is in the upper right quarter of the graph), although some had low estimated accuracy in both studies (lower left quarter). Only two tune pairs had model estimates with below chance accuracy in the present study but above chance in Cole *et al*., LHH_LHL and HHH_LHH, both under Anger in the present study (lower right quarter). More tune pairs had estimated accuracy above chance in the present study but below chance in Cole *et al*. (upper left quadrant), including LHL_LLH under Pride, HLH-HLL under Anger and Shame, and LHH_LLH under Shame.

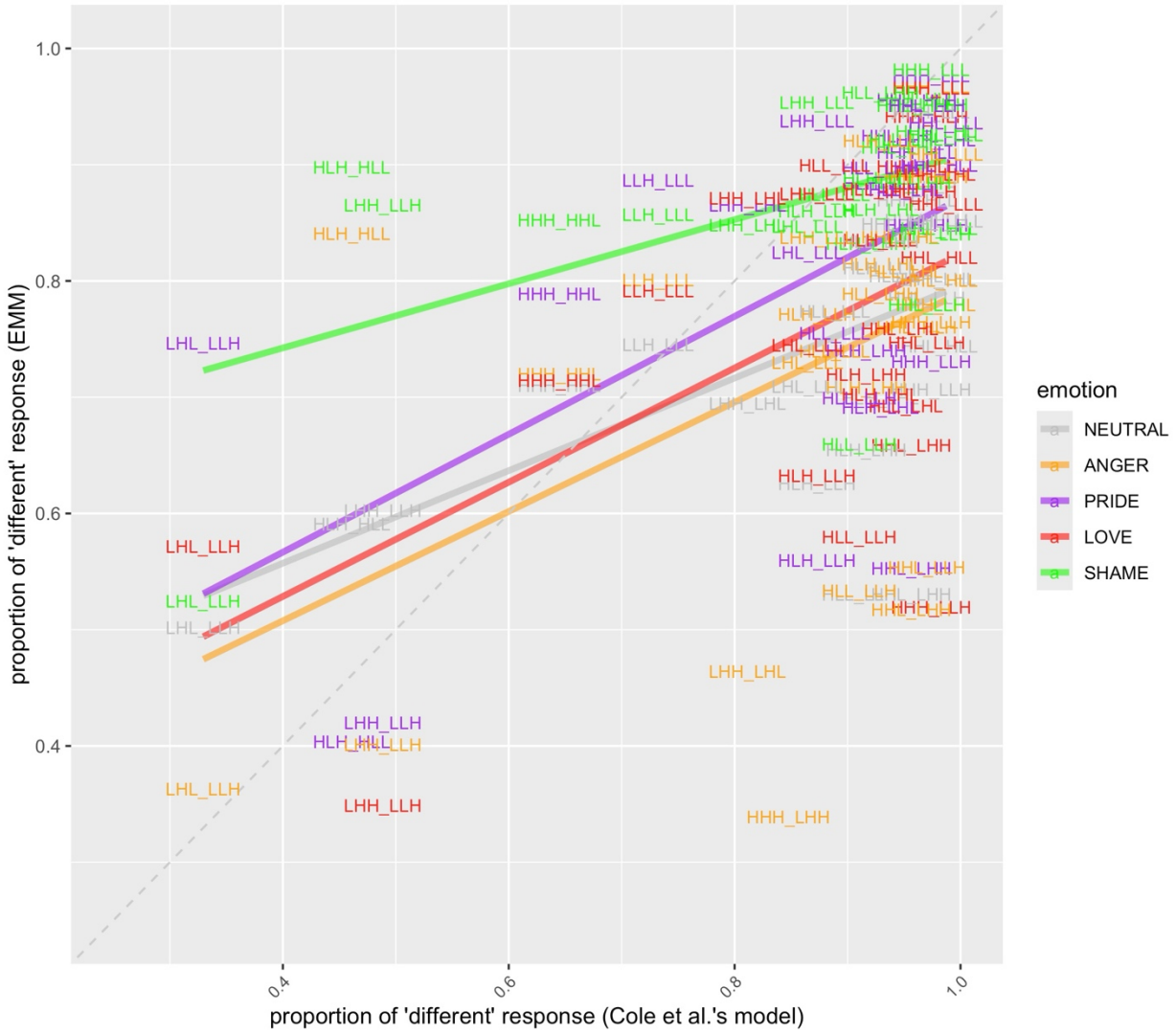


Figure 15: Estimated proportion of "different" (correct) responses in the EMM (y-axis) versus Cole et al. (2023) (x-axis) by emotion (color) with linear regression lines by emotion. Regression model R^2 and p-values are given in Table 5. Dashed line indicates where EMM = Cole et al. (e.g. 'perfect' correlation).

Table 5: Details for Linear Regression Models Shown in Figure 15

Emotion	R^2	p	p < .05?
Neutral	0.344	0.0013	Yes
Anger	0.216	0.0127	Yes

Pride	0.314	0.0024	Yes
Love	0.287	0.0048	Yes
Shame	0.258	0.0068	Yes

Tune Match Model (TMM) results

For the full model output, see Appendix 3B. Figure 16 shows the estimated proportion of "Different" (incorrect) responses when tunes match and emotions mismatch. Recall that in this experiment, participants were instructed to make a Same/Different judgment ignoring "the speaker's expression of emotion in their voice". Emotion pairs with higher estimated proportion of "Different" responses (i.e., more errors, further to the right in Figure 16) could be due to participants mistakenly interpreting differences due to emotional portrayal as a difference in tune. Three emotion pairs led to an at-chance likelihood of a "Different" response (Pride_Shame < Anger_Neutral < Love_Shame).Estimates of inaccurate responses are above chance for the remaining seven pairs, among which three have especially high estimated error rates: Love_Pride > Anger_Shame > Neutral_Pride. This pattern is readily explained by the matching status of the emotional valence, especially if Neutral is counted among the positive valence emotions (with Love and Pride). Whenever both emotions in a pair match in valence (grouping Neutral with positive), the error rate is higher. Likewise, when emotions in a pair mismatch, likelihood of a "Different" judgement is lower.

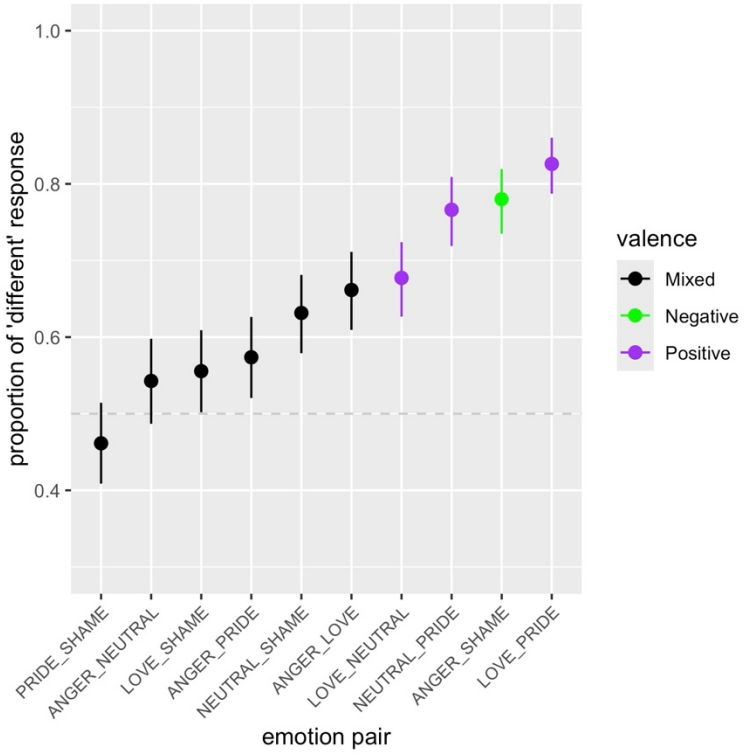


Figure 16: Estimated proportion of "different" (error) responses for emotions pairs in the TMM with coloration by the valence of the emotions as shown in key. Neutral is grouped with positive emotions following observations in the empirical data.

Tune Only Model (TOM) results

For the full model output, see Appendix 3C. The details of this model are not discussed here because its sole purpose was to benchmark the TEM in the cross validation. Given that the TEM had a lower LOOIC, there is little justification to interpret TOM, besides the fact it partially replicates the model from prior work (Cole *et al*. 2023). The model summary for the TOM is in Appendix A.

Tune-Emotion Interaction Model (TIM) results

For the full model output, see Appendix 3D, which also has a plot of all the interactions between tune pairs and emotion pairs. Here, the focus is on tune pair and emotion pair as main effects, shown in Figures 17 and 18 respectively. Figure 17 shows the model estimated discrimination accuracy below chance for only one tune pair, HLH_HLL, with estimated accuracy of most other tune pairs being well above chance (leftmost orange point). Other lower-accuracy tune pairs were HHH_LHH, LHH-LLH, and HHH_HHL, also marked in orange, after which the CrIs of estimated means tend to steadily improve. Model estimated accuracy was especially high for six tune pairs: LHH_LLL, HHH_LHL, LLH_LLL, HHH_LLL, and LHH_LHL, with broad overlap in CrIs among these six. Based on the empirical mean response rates seen in the EMM (Figure 7) and TIM/TOM data (Figure 11), HHH_LLL (green point) could have been expected to be the most accurately discriminated tune pair, but taking variation due to emotional portrayal and individual participant into account, that tune pair has only the fourth highest estimated accuracy, slightly below ceiling.

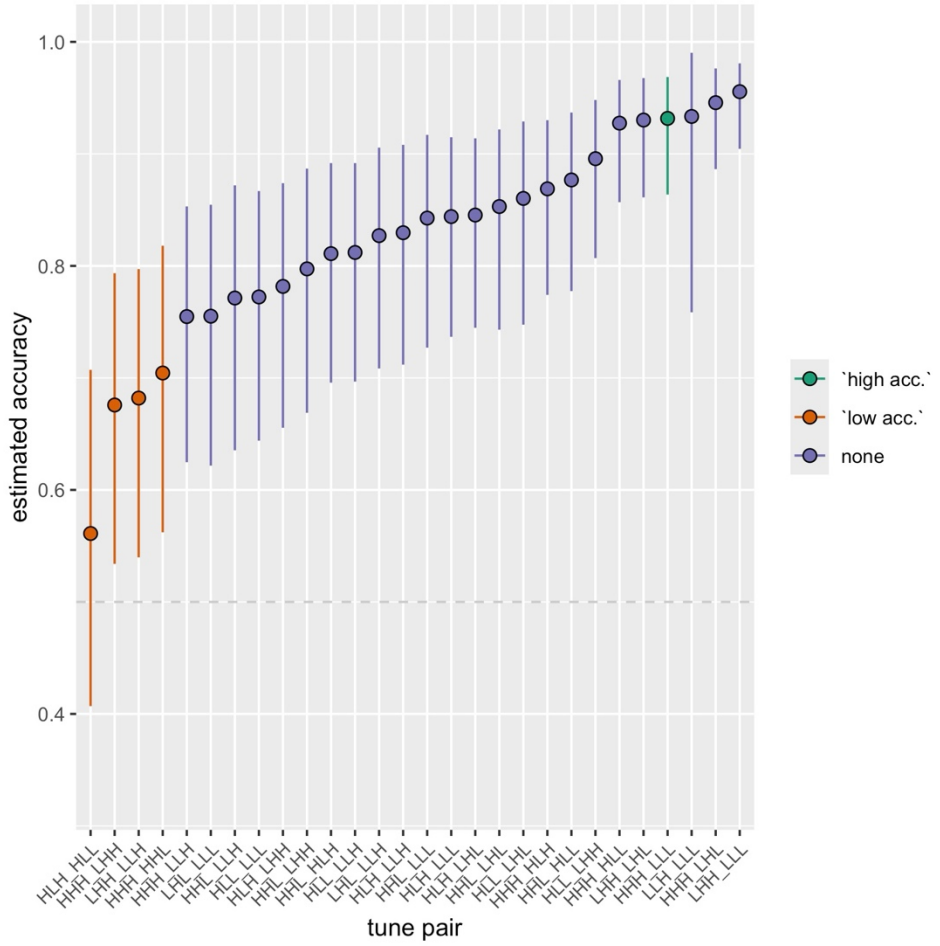


Figure 17: Estimated proportion of "different" (correct) responses for tune pairs in the TIM. The tune pair with the expected highest accuracy, HHH_LLL, is in green and the four tune pairs that were found to have the lowest accuracy are in orange, with other tunes in blue.

Turning to the statistical results by emotion pair, most but not all emotion pairs have CrI's crossing chance, as shown in Figure 18. For the three above chance pairs, emotional valence is different in Anger_Love and Pride_Shame, and also for Neutral_Shame (if Neutral is taken to have positive valence as suggested by the TMM model results discussed above). Likewise, when valence matches, as in Love_Pride and Anger_Shame, accuracy tends to be lower. There is considerable overlap between the CrI of all emotion pairs.

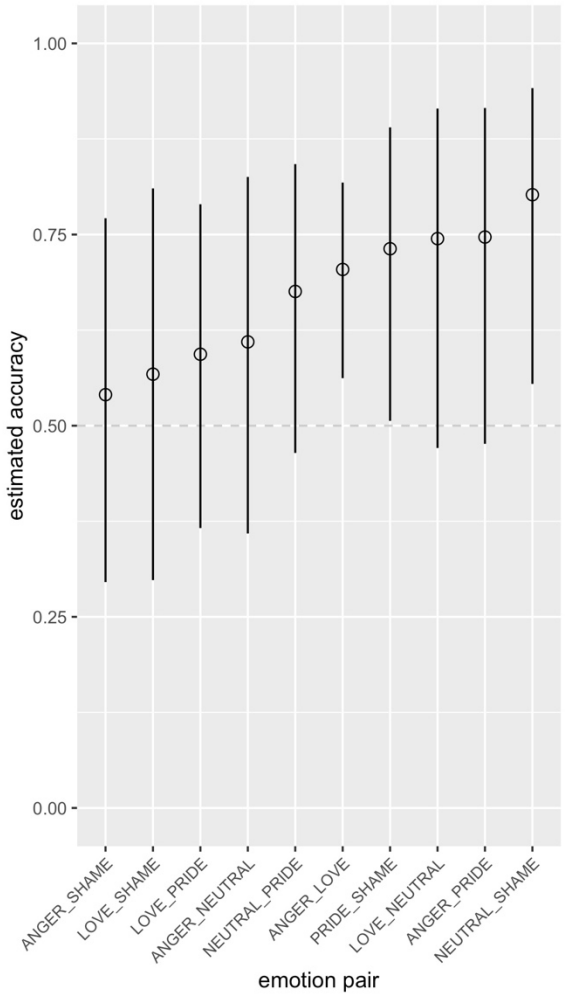


Figure 18: Estimated proportion of "different" (correct) responses for emotion pairs in the TIM.

Comparing perceptual discrimination (EMM) with stimulus acoustic similarity (RMSD)

This analysis involves plotting and modeling the acoustic distance between pairs of stimuli (specifically their RMSD, previously shown in Figure 5) with the estimated accuracy from the EMM model (emotions match; tunes mismatch). The purpose of this analysis is to understand how greater distance in F0 space between the F0 trajectories of a pair of stimuli accounts for "different" tune judgements. Figure 19 plots EMM model estimates on the y-axis and the stimuli RMSD on the x-axis, where each data point represents a tune pair (e.g., HHH_LLL), color coded

by emotion condition. Additionally, linear regression models were fit for each emotion condition, with the formula y-axis ~ x-axis, to see whether the ability to predict perception results using stimuli F0 differences varied by emotion. Table 6 gives the R^2 and p-value for the regression models, which showed differences in correlation although all fits were statistically significant. The emotion with the highest correlation based on R^2 was Pride (.312) and the lowest was Shame (R^2 = .147), while Anger, Love, and Neutral had similar R2 values and slopes.

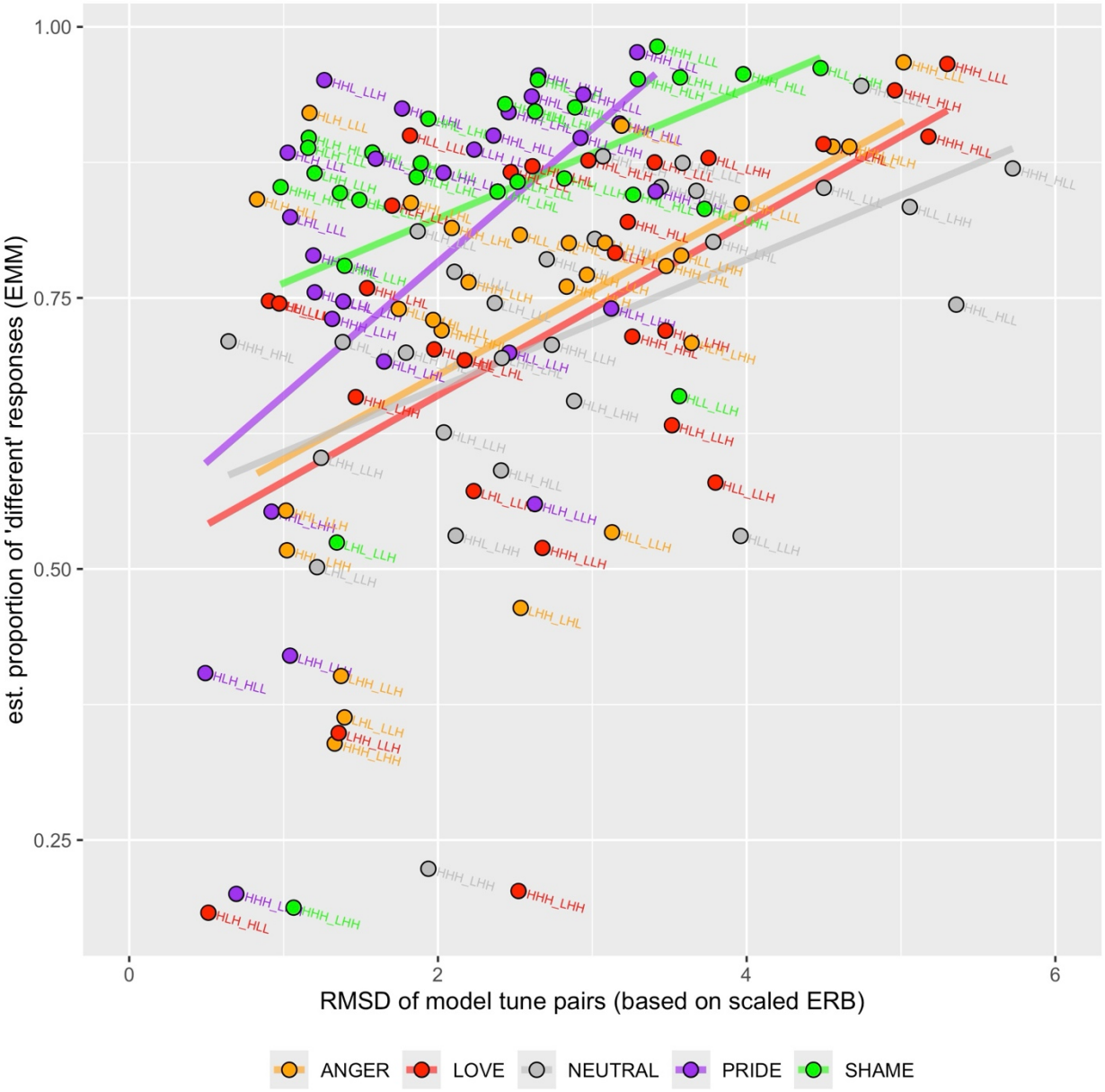


Figure 19: Estimated proportion of "different" (correct) responses for tune pairs in the EMM data (y-axis) compared to the RMSD distance between tune pairs (x-axis) by emotion (color). R-squared and p-values for each regression line is given in Table 6.

Table 6: Details for Linear Regression Models Shown in Figure 19

Emotion	R^2	p	p < .05?
Neutral	0.263	0.005	Yes
Anger	0.270	0.005	Yes

Pride	0.312	0.002	Yes
Love	0.248	0.007	Yes
Shame	0.147	0.044	Yes

Figure 20 compares model estimates of discrimination accuracy for the EMM data by tune pair to differences in F0 trajectories as measured by the difference GAMM analyses reported in Chapter 2. This is arguably a less direct perception-production comparison than the RMSD analysis above, because here, the acoustic difference can be seen as representative of *targeted* distinctions from the imitated production experiment. The purpose of this analysis is the same as the preceding analysis—to compare discrimination accuracy of EMM data to a measure of F0 similarity for a given tune pair. The comparison of the difference GAMMs to the EMM data is appropriate because the difference GAMMs compared tune pairs within a particular emotion, and this model also considers only those trials from the AX discrimination experiment with mismatched tune pairs in the same emotion condition. As with the RMSD analysis above, separate linear regression models were conducted for each emotion, which are detailed in Table 7. The results show that most (but not all) tune pairs had significant positive slopes; Shame was only marginal at $p = .058$[18] and is therefore omitted from Figure 20. Based on R^2 values, Neutral was most correlated with Cole *et al.*'s results ($R^2 = .387$), followed by Pride ($R^2 = .368$) and Love ($R^2 = .324$), which are similar in R^2 and slope, and finally the least correlated emotion was Anger ($R^2 = 152$).

[18] The R^2 value for Shame in this linear model was 0.131, which would have led it to be considered the least correlated condition, if it was statistically significant.

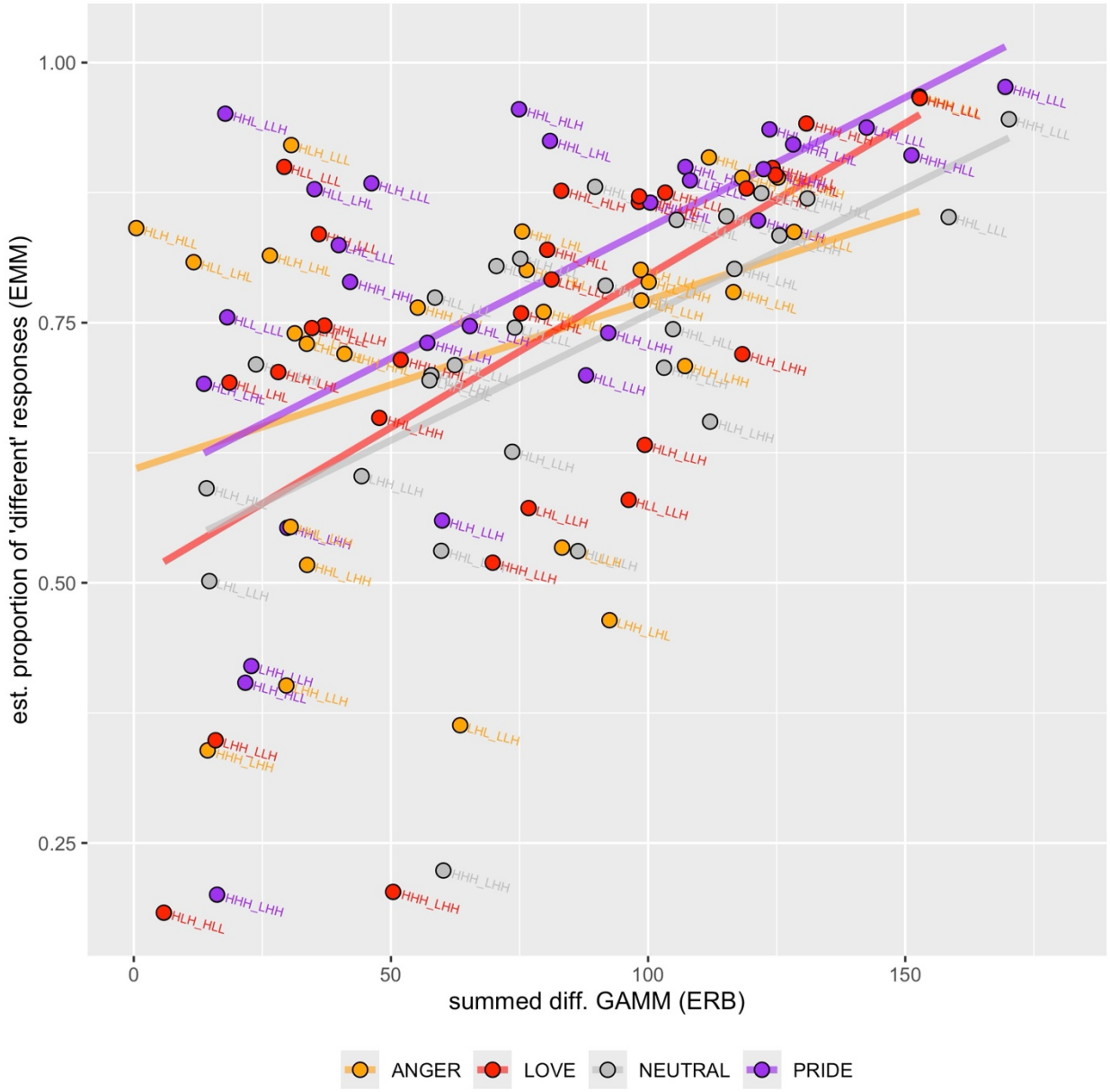


Figure 20: Estimated proportion of "different" (correct) responses for tune pairs in the EMM (y-axis) compared to the summed significant regions based on difference GAMM analysis in Chapter 2 (x-axis) by emotion (color). Shame not shown because p>.05—see details for the regression models in Table 7.

Table 7: Details for Linear Regression Models Shown in Figure 20

EMOTION	R^2	p	p < .05?
NEUTRAL	0.387	<0.001	Yes
ANGER	0.152	0.040	Yes
PRIDE	0.368	<0.001	Yes
LOVE	0.324	0.001	Yes
SHAME	*n.s.*	0.058	No

3F. Discussion

Interim summary

This study aimed to test whether variation in the speech signal due to emotional portrayal diminished the perceptual salience between phonologically distinct intonational tunes, as defined by the AM model (Pierrehumbert 1980, Ladd 2008). From a related production study (see Chapter 2) that crossed an eight-tune inventory that encoded the basic tune distinctions predicted by the AM model with emotional portrayals, a set of stimuli was assembled consisting of utterances with all combinations of tune and emotion, to test the perceptual distinctiveness of tunes as a function of the emotional portrayal. Statistical analysis of the data consisted of Bayesian GLMMs modeling discrimination accuracy, which were used to examine the following hypothesis.

Hypothesis revisited

The hypothesis for this study was framed in a strong-versus-weak dichotomy, where the strong version stated that listeners perceive tune contrasts with similar accuracy under all emotion conditions, including Neutral. The weaker version states that the specific emotion also plays a role in determining tune perceptual distinctiveness, which would emerge as a general discrimination accuracy differential between emotions if it bore out, such as superior accuracy under Neutral.

Evaluating the hypothesis: did emotion have an effect?

Finding A: We begin by considering the strong evidence that emotional variation generally impacts tune discrimination. The speaker's emotion impacted how listeners perceived tunes based on: (i) model cross validation (see Table 4), (ii) systematic variation by emotion portrayal in model results across tunes (TMM, see Figure 16), between emotion conditions (EMM, see Figure 13), and with interactions (TIM; see Figure 18), and (iii) the presence of significant tune-emotion interactions in the model of best fit (TIM; see Appendix D).

Discussion of Finding A. Recall that the LOOIC-based model cross validation showed that adding an emotion term and its interactions significantly improved the fit, seen in the comparison between the Tune Only Model (TOM) and the model of best fit, the Tune Emotion Interaction Model (TIM). In the model that collapses across tunes, the Tune Match Model (TMM), it was observed that error rates were higher when the emotions in a trial were of the same valence (e.g. accuracy for Positive_Positive < Positive_Negative). For example, the emotion pairs with the highest error rate were Love_Pride and Anger_Shame, and Neutral_Pride while those with the lowest were Pride_Shame, Anger_Neutral, and Love_Shame. Interestingly, if this pattern holds for Neutral, which other analyses suggest is the case, it should be considered to have positive valence. In the model that collapses across emotions, the Emotion Match Model (EMM), there was a three-way distinction between the five emotion conditions; Neutral, Love, and Anger led to the same outcome, whereas Pride and Shame led to exclusively different predictions.

Finding B: Differences emerged between the present study and prior work without structured emotional variation (Cole *et al.* 2023, see Figure 15), which was significant for every emotion based on the Emotion Match Model (EMM; see Table 5). The emotion condition most correlated with Cole *et al.*'s model was Neutral (R^2 = .344), closely followed by Pride (R^2 = .314), while other specified emotions were lower (see Table 5). This means that, not only did emotional variation lead to across-the-board effects on tune discrimination, but accuracy was better or worse depending on the emotion. Trends varied by emotion; Shame was the emotion in which tunes were most accurately discriminated, by a wide margin, though it was less correlated with Cole *et al.*'s results (R^2 = .258) which had higher average accuracy. This is most readily explained by higher accuracy for some tune distinctions in the present study, many of which are under Shame (see upper left quadrant in Figure 15). In other words, certain tune distinctions that are apparently challenging to perceive without emotional variation (Cole *et al.* 2023) seem to have been enhanced in certain emotional portrayals, such as tunes under Shame. Finally, there are notable differences between the quality of model fit for the present study compared to Cole *et al.* (2023)—specifically their model fit was far better, based on cross validation (see Table 4). This is taken as an additional indication that emotional variation implicitly increases the noise that participants in perceptual discrimination studies must contend with to be successful in the task.

Evaluating the strong versus weak hypotheses

Given that emotional variation impacts tune discrimination, the question turns to systematic differences between emotions, specifically whether listeners were equally accurate in all conditions (strong) or differentially more or less accurate depending on the condition (weak). Based on slightly more distinctive F0 trajectories produced in Chapter 2, as shown by the difference GAMM analysis, it was important to see whether Neutral was perceived as more distinct, which was not the case. Rather, tune discrimination was better explained in terms of acoustic distance than the specific emotion, as discussed in the following findings:

Finding C: Tune discrimination can be enhanced by emotional variation, of which Neutral is a type. As noted in the discussion of Finding A above, the accuracy rate under Neutral was not superior to other emotions, rather all emotions were in the same range, with Shame being the most accurate (based on y-intercept).

Discussion of Finding C. How do we reconcile this outcome with the fact that accuracy in the present study is generally worse than was found in Cole *et al.* (2023), as noted in the discussion of Finding B? This may have to do with the fact that their participants needed to detect *any* difference between stimuli which *only* differed in tune, whereas in the present study participants needed to detect a difference in the speaker's intended meaning—a truly challenging task. Given that accuracy was generally higher in Cole *et al.* (2023), the emotions in the present study that most strongly correlate with those results are arguably conducive to accurate tune discrimination. Neutral is expected to strongly correlate with Cole *et al.*'s results, and does, but the correlation

with Pride is nearly as strong, so an advantage for tune discrimination under Neutral is not well supported by the data.

Finding D: Tune discrimination accuracy is strongly positively correlated with acoustic distance regardless of emotional variation. This is based on analysis of two separate, but related, comparisons: *(i: EMM × RMSD)* The Emotion Match Model (EMM) was plotted and modeled (by emotion) against the calculated acoustic similarity (RMSD between F0 trajectories) of the audio stimuli, which showed significant correlations across the board. The strength of the correlation between tune discrimination accuracy and RMSD held across accuracy ranges. For example, the strongest correlation was found for Pride ($R^2 = .312$; see Table 6) which was only the second most accurate emotion condition after Shame (see Figure 13), which had the weakest discrimination correlation with RMSD ($R^2 = .147$; see Table 6). This means that, to some degree, tunes were more accurately discriminated under Shame than RMSD would predict for other emotions, including Neutral, although the correlation is barely significant ($p = .044$; see Table 6). If Shame had enhanced tune distinctions, it would have resulted in a stronger correlation with accuracy, so it might be the case that the manner in which Shame diminishes tune distinctions is easier to account for perceptually. *(ii: EMM × GAMM—Ch. 2)* Additionally, the EMM was compared to the F0 predictive model from Chapter 2, a GAMM which captures trends in tune distinctiveness as they are produced with emotional variation, using difference GAMMs. Overall, the outcome was highly similar to the RMSD-based comparison (see Figure 20), with some key differences. First, discrimination under Neutral was the strongest correlation with the GAMMs ($R^2 = .387$; see Table 7), versus Pride (which was close, at ($R^2 = .368$). Second, the

linear model predicting accuracy under Shame compared to the GAMM was not significant (rather than marginally significant, as in the RMSD-based comparison). If we consider the findings of both comparisons together, the experimental evidence strongly supports H2.

Discussion of Finding D. Based on the strong, and mostly uniform, correlations between specific emotions and tune discrimination accuracy, much of the perceptual results can be explained by acoustic differences and participants exercising simple pitch perception. But this does not explain the whole of Finding D, which shows different effects depending on the emotion (like Shame). Nor does the acoustic explanation account for the whole of the data, based on visual observation of the residuals (distance between points and the model prediction) in Figures 19 and 20. Importantly, these residuals are not systematic, so there is no evidence to reject the notion that secondary cues (beyond F0) are necessary to acoustically characterize intonation for the purposes of statistical modeling.

It is notable that for many tune pairs for which the acoustic explanation appears to be incomplete, which leaves open the possibility that other factors may impact the emotion conditions under which particular tune pairs are more perceptually salient. One such factor may be whether or not participants recognize the tune as part of their lexicon or not, which may trigger something like a perceptual magnet effect (first described by Liberman *et al.*, 1957). Recently this effect has been attested to apply to intonational categories like pitch accents (albeit weakly, compared to phonemes), which suggests it may also explain these results. Problematically, since the exhaustive tune inventory for MAE is a topic under active research (unlike its phoneme inventory), so a perceptual magnet effect cannot currently be tested for this

set of data, but it may be possible in the future. A related source of variation that may explain why acoustic distance does not perfectly predict tune discrimination is the perceived meaning of tunes, which may also interact with the perceived emotion (although this is not predicted by the AM model). Tune meaning is beyond the scope of the present study (see Chapter 4), but as phonologically defined units, tunes ultimately exist to convey linguistic meaning, so it would be surprising if it played no role in tune distinctiveness.

Evaluating tune distinctiveness

While a hypothesis about distinctiveness between tunes was not stated, it is highly relevant to the answering the broader question of the robustness of the intonational system to convey linguistic information despite variation due to speaker emotion. This is because, if listeners reliably detect all tune distinctions regardless of emotional variation, it bolsters the phonological framework which is responsible for the rules which generated the tunes, the AM model. If the predictive power of the AM model holds across speaker emotions, which it does not explicitly claim to be able to do (see Chapter 1), it would provide strong support for continuing to develop this model. On the other hand, if emotional variation severely diminishes tune distinctiveness, it calls into question how intonational categories would convey critical linguistic information in everyday situations if the AM formulation is correct. That said, aforementioned evidence for Finding C (tune discrimination can be enhanced by emotional variation, of which Neutral is a type) bodes well for tunes remaining salient despite accompanying emotional variation. The strong case for phonological differences partly determining perceptual differences is summarized in Finding E:

Finding E: Perceptual salience for tune differences is partly predictable based on the phonological representation (a tone sequence). The evidence comes from the statistical models fit to the data with mismatching tune pairs, the Emotion Match Model (EMM) and Tune Emotion Interaction Model (TIM), where the correct answer is "different". The key difference between these data is that in the EMM there one emotion to perceptually account for, while in the TIM *both* the tune and emotion mismatch, which participants found slightly easier based on the empirical mean response rate (see Table 4). In the EMM, about 57% (16 of 28) tune pairs were reliably discriminated above chance response rates, and the cline of response accuracy reflects tonal differences between tunes (see Figure 14). Specifically, the trend shows that the perceptual salience between different tunes increases as a function of the degree of tonal difference. When all three tone positions for a tune pair differ, as in HHH_LLL, accuracy is near ceiling, whereas when only one tone position differs, as in HHH_LHH, accuracy is lower (this tune is even reliably below chance). These trends hold for the TIM, which also shows HHH_LLL to be among the most accurate tune pairs to discriminate, although based on the estimate mean it is in fourth place rather than first, as in the EMM (see Figure 17, green marks). That said, many of the high accuracy tunes around HHH_LLL differ in two tone positions (e.g. LHH_LLL, HHH, LHL, LLH_LLL, and HHH_HLL) with one exception (LHH_LHL). Similarly, tune pairs with lower discrimination accuracy all differed by one tone (HLH_HLL, HHH_LHH, LHH_LLH, and HHH_HHL; see Figure 17, orange marks). When tunes differed by two or more tones in the TIM, they were reliably discriminated with over 60% accuracy. Given that this finding emerges from data with high levels of variation, this evidence seems to substantiate the AM model's formalism, specifically in the phonological representations of intonation.

3G. Conclusion

The present study is part of a research project aiming to interrogate phonological predictions made by the AM model (Pierrehumbert 1980, Ladd 2008), as instantiated within the tune inventory, in light of phonetic covariation due to speaker emotion. To that end, the present study tested the perceptual distinctiveness of tunes as produced with portrayed emotion, to understand whether listeners could disentangle speech cues for intonation from those for emotion, and the perceptual consequences of such variation. Participants made same-different judgements over naturally produced tune-emotion combinations on the basis of whether they perceived the speaker intended a different meaning, ignoring emotion, which was intended to focus their judgement on the linguistically relevant part of the speech signal. The present study inherits the limitations of the production study which the stimuli were drawn from (Chapter 2), as well as particular limitations, mainly due to resource limitations. The details and implications of the study's limitations are set aside for now and revisited in detail in Chapter 5 (General Discussion). The results of this study, based on the synthesis of the quantitative analyses, several experimental findings materialized from the results, which can be summed up by the following general conclusions:

#1. Emotional variation impacts tune perceptual distinctiveness in a manner which is consistent between emotions and does not tend to negatively impact same-different judgements (Findings A, B, & C).

#2. Relative pairwise tune discrimination accuracy is partly explained by acoustic differences, but is also anticipated by phonological (tonal) differences—these factors are interrelated (Findings D & E).

These conclusions have implications for intonational phonology and future research methodology looking at the connection between intonation form and function. Conclusion #1 in particular validates a longstanding intuition within intonational phonology (also partly validated in Chapter 2) that emotion is not a major confounding factor for its predictions. Typically, speaker emotion is avoided as a topic because, possibly because of a perceived connection about how intonational cues may be impacted. These results should allay concerns by researchers who may want to use speech materials which are more naturalistic than those typical of perception studies. It is possible that adopting more naturalistic materials for intonation research could open the door to the analysis of larger datasets, which could help answer longstanding questions, like the exhaustive set of lexicalized intonational tunes for MAE.

Another way in which these results support future empirical work in the area of intonational phonology is robust evidence that F0 is sufficient to characterize intonational phonetic cues in the speech signal. This is based on a significant correlation between relative perceptual discriminability of tune pairs and the magnitude of their F0-based differences. Whiles secondary cues may play some role, there was no evidence in the present study that its importance changes as a function of increasing emotional variation, which would have boosted discrimination accuracy in ways beyond the explanatory power of F0. In sum, this study provides needed empirical validation for the AM model's method of formalizing intonational categories (tone sequences) and provides evidence that listeners can disentangle the predicted phonetic instantiation of those categories (F0 trajectories) from a highly variable speech signal that reflects an emotional portrayal. Additional work is needed to understand how intonational meaning influenced these results, which the next chapter begins to address. Chapter 5 discusses limitations of this experiment and the consequences for interpretation of the results. Based on

this project's understanding of the results, the findings should encourage future development of the AM model based on the impressive way in which the predictions are corroborated by these experimental results.

Chapter 4: Interpretation

4A. Introduction

The present work is concerned with testing fundamental aspects of intonational phonology, specifically as asserted by the AM model (Pierrehumbert 1980, Ladd 2008), vis-à-vis a set of eight phonologically distinct forms called intonational tunes. As discussed in prior chapters, these tunes encode the basic phonological distinctions for Mainstream American English (MAE), which the AM model claims are responsible for conveying linguistic meaning. While tunes generated by the AM model are phonologically well-formed (definitionally) within the AM framework, the status of any particular tune as a linguistic category in MAE, in the sense of being fully lexicalized, is unknown due to the uncertain mapping between phonological forms and meanings. This uncertainty arises, at least in part, to the extensive phonetic variation in production of intonation and, in particular, the influence of paralinguistic factors. Whereas the preceding chapters investigated the production and perceptual discrimination of intonational tunes, the experiments presented here pose the question of whether and how speaker emotion affects the perceived meanings of tunes, whatever they may be. Doing so required positing an underspecified tune meaning possibility space. From a phonological perspective, the representation of sound categories relating to linguistic meaning is not expected to be impacted by the speech correlates of nonlinguistic information, including the speaker's emotion. On the other hand, results from prior chapters suggest that tune and emotion are acoustically entangled, raising the question of whether tunes convey linguistic meaning independently of the concurrent emotional portrayal.

To investigate these possibilities, the current study comprises two experiments with one task each, dubbed ‘Sorting’ and ‘Rating’. Participants in the first experiment SORT audio stimuli consisting of tune-emotion combinations into groups based on what they perceive the speaker is trying to convey while ignoring other factors including the expressed emotion. To be successful in the task, the participant has to disentangle critical linguistically speech cues to decide what function is being conveyed. If each tune is perceived to convey an exclusive meaning, then participants would be expected to create one group per tune, but it is also possible in this design for different tunes to convey the same meaning (a many-to-one mapping) or for any individual tune to convey different meanings (a one-to-many and potentially many-to-many mapping). The critical issue at hand is not how particular tunes convey particular meanings, but the stability of the emergent tune-meaning relationships according to how participants interpret tunes.

In the second experiment, as the name implies, participants RATE tune-emotion combinations according to their perceived ability to convey a specified meaning, one drawn from prior literature. If tunes convey their intonational meanings irrespective of emotional variation, then participants would be expected to rate tune meaning in the same way regardless of the concurrent emotion. Problematically, as the next section explains, prior research on intonational meaning is unbalanced across the tune inventory, because some tune-meaning associations are widely recognized in the literature while the meaning of other tunes is relatively unaddressed. Given the methodological diversity of its tasks, and large tune inventory, the present study is well positioned to shed light on the details of particular tune meanings and guide future tune-based intonation research.

4B. Background

Attested tune meaning

There is a deep well of prior work investigating intonational meaning, but as discussed in Chapters 1 and 2, research efforts have not been equally distributed across the eight tunes being tested in this project. In the literature on intonational forms and meanings, it is common to encounter impressionistic and imprecise descriptions like "rising" or "falling" for the holistic shape of an F0 trajectory, whereas tunes in the present study are formally specified through a phonological tone sequence. To the extent possible, the literature review presented here focuses on analyses where intonational meaning is discussed in relation to intonational forms that have been explicitly phonologically specified. The main advantage of constraining the literature review to work that maintains a phonological understanding of intonation is to enable direct comparison to intonational meaning claims in prior work. Table 1 (reproduced from Table 1 in Chapter 2) summarizes the conventionalized meanings of each tune in the inventory, as previous research has labeled and described them.

Table 1: Attested tune meanings

Tune	Linguistic meaning/function	Source*
HHH	Questioning, possibly when the answer is believed to be positive	PH, Je
HHL	Elaborating on something that's been previously mentioned	PH, Ba
HLH	Non-finality, uncertainty, selecting the addressee	PH, Gu, WH
HLL	Declarative, asserting information, possibly incomplete	PH, Go, Gu, Ba
LHH	Questioning, typical for polar questions, or incredulity	PH, Ho, Je, Gu
LLH	Speaker believes listener should already know this information	PH, BT
LHL	Prompting for the speaker to respond, possibly as a reminder	PH, Ba
LLL	Finality, non-predication	PH

*Key to citations: Ba = Bartels (1994); BT = Burdin and Tyler (2018); Go = Goodhue et al. (2016); Gu = Gussenhoven (2002); He = Heim (2019); Je = Jeong (2018); PH = Pierrehumbert & Hirschberg (1990); WH = Ward & Hirschberg (1985)

With the notable exception of Pierrehumbert & Hirschberg (1990), sources in Table 1 address only one or two tunes at a time, which raises doubts about how the stated descriptions of tune meaning would generalize to untested tunes that are phonologically similar but yet formally distinct. Studies considering multiple phonological neighbors may be better positioned to identify categorical boundaries. In fact, without evidence that a particular tune conveys a certain meaning while its phonological neighbors do not, it is unclear whether a certain meaning is endorsed for that tune, or if the received interpretation is due to phonetic or phonological similarity with another tune.

The present study addresses these limitations by considering not only the attested meaning or meanings of particular tunes ('one-to-one'/'one-to-many'), but ways that multiple tunes could convey multiple meanings ('many-to-many'), as has previously been suggested (Roettger et al., 2019). Certainly, the possibility of many tunes mapping to many meanings is implied by the meaning descriptions in Table I, since there are examples pairs of tunes like LHH and HHH, which both are claimed to convey *questioning*. On the other hand, multiple tunes also seem to convey a type of statement (HHL, HLL, LLL), which logically excludes the *questioning* tunes. The design of the Rating task takes particular advantage of such relationships between attested tune meanings, described next.

Meaning dimensions for ratings

Based on Table 1, some tunes are attested to convey opposite meanings, such as LLL (*finality*) and HLH (*nonfinality*), and other tunes may convey congruent meanings, such as LHH and HHH (*questioning*). The Rating task design exploits these oppositional relationships in its construction of a semantic-pragmatic possibility space that prior literature suggests should encompass (to

some degree) the basic distinctions amongst the eight tunes under investigation. A could be conveyed by the set of tunes under investigation in this work. This is the way the analysis will track whether intonational meaning shifts depending on concurrent emotional speech cues, by identifying and comparing tunes' emergent meaning associations, using attested semantic-pragmatic meaning dimensions. One of the most well attested functions of intonation is conveying the difference between *asking p* and *stating p*, which the present study conceptualizes as complementary meaning. Being in roughly complementary distribution, *asking* and *stating* judgements are potentially highly mutually informative and were therefore modeled together as poles of a common meaning dimension, dubbed *questioning*. Three total dimensions with two poles (or directions) each are proposed, as shown in Table 2. In addition to the *questioning* dimension, the present study tests *floorholding*[19], which is intended to capture the finality/non-finality distinction, and *committing*, intended to capture the speaker's perceived commitment to the tabled proposition[20].

Table 2: Meaning dimensions

Meaning dimension	Pole	Possible Interpretation	Prompt
Questioning	+	**Question**	It sounds like the speaker is asking a question.
	-	**Statement**	It sounds like the speaker is giving a statement.
Committing	+	**Commitment**	It sounds like the speaker believes what they're saying.
	-	**Withholding**	It sounds like the speaker doubts what they're saying.
Floorholding	+	**Floorholding**	It sounds like the speaker wants to keep talking.
	-	**Floor-ceding**	It sounds like the speaker is finished talking.

[19] For the present purposes, 'floorholding' means 'holding the floor' and is written thusly to be consistent with the other meaning dimensions (versus "floor holding" or "floor-holding").
[20] See Farkas & Bruce (2010) for a review about the Table model.

It should go without saying that these dimensions are not intended to represent *all* possible tune meanings in MAE, but rather to capture the key semantic-pragmatic distinctions that emerged in the literature review and thereby provide a rudimentary framework in which to examine possible effects of emotional portrayal on the interpretation of tune meaning. Given this design, each tune is examined twice for each meaning dimension, meaning that participants are, over the course of the experiment, asked both reject or endorse each tune-meaning combination. In order to operationalize the meaning dimensions in Table 2, a positively worded[21] prompt corresponding to each possible interpretation (i.e., both poles of each meaning dimension) is presented to the participant, along with an audio stimulus of a tune-emotion combination, which they respond to with a rating. If participants give a high agreement rating for a certain tune-prompt combination, we can take it as evidence that the interpretation associated with that prompt is available for the given tune, especially if ratings for the prompt associated with the opposite pole are low.

Free classification

Given the unbalanced distribution of prior research across the adopted tune inventory, it can be assumed that gaps exist in the current understanding of tune meaning, raising the possibility that the dimensions in Table 2 might be insufficient to fully characterize underlying distinctions in perceived tune meaning that could arise. Therefore, it was opted to pair the experiment resting literature-attested tune meanings with a more open-ended task that imposed as few assumptions as possible on the intonational meaning space, accomplished by using the Free Classification experimental paradigm. Free Classification involves participants judging stimuli according to an experimental criterion based on perceptually salient features, which is a paradigm in use since at

[21] Which is to say, negation is not used to convey the opposite meaning.

least Imai & Garner (1965). For the present study, the specific methodology is *Auditory* Free Classification (AFC) as implemented in Clopper (2008), which is a type of task that has previously been applied within linguistics, such as the perception of regional dialect (Clopper & Bradlow, 2009), although this may be the first such implementation for intonational phonology.

In AFC, participants judge auditorily-presented stimuli by sorting them into groups based on the experiment instructions and the perceptually salient features, which means the participant ultimately decides how to evaluate speech cues—no particular mapping between form and function has to be assumed. To preview the current study's methods, rather than asking participants to group stimuli based on dialect, as in Clopper & Bradlow's study, participants in are instructed to consider what the speaker meant to convey, independent of the feeling or mood also being expressed (in the speech signal) by the speaker. If participants successfully ignore emotion, and accurately discriminate the underlying tunes, they should create a separate group for each tune (one-to-one mapping of tunes onto groups). Alternatively, if participants judge different tunes to express the same meaning, they may group two or more tunes together in a many-to-one mapping, which could result in fewer groups than tunes. Another possibility is that a single tune will be associated with multiple meanings in a one-to-many mapping, or even many-to-many if this holds for many tunes, resulting in potentially many more groups than tunes. To summarize, the purpose of these experiments is not necessarily to gain a comprehensive view of how specific tunes are generally interpreted, but rather how tune interpretations, whatever they may be, are impacted by emotional variation, which will shed light on tunes generally.

4C. Hypotheses & predictions

This study primarily considers the effects of emotional variation on the perceived meaning of phonologically specified tunes. In broad terms, the utility of a phonological model is to define (for a specific language, in this case MAE) the relationship between sets of linguistic forms (in our case, tunes) and their function in conveying differences in linguistic meaning. From a linguistic perspective, intonational meaning is perfectly distinct from emotions. This makes distinguishing emotional variation from linguistically meaningful variation a critical component of linguistic competence. In order to shed light on whether, and the degree to which, emotional variation impacts how listeners interpret tune meaning, the following competing hypotheses are presented in the present study, and specific predictions for each experiment are given.

Independent hypothesis (H1)

The linguistic meaning associated with a tune and its accompanying emotional portrayal are interpreted independently, such that tune meaning is stable regardless of speaker emotion. This hypothesis is anticipated by the finding in Chapter 2, particularly the GAMM modeling, showing that the effect of certain emotions was (more or less) uniform across tunes, leading to high predictability. Additionally, Chapter 3 showed broad evidence that listeners are able to perceptually factor out effects of emotion on the phonetic implementation of tunes when discriminating tunes along phonological criteria. If the Independent hypothesis is correct, it carries different implications for each experiment:

- For *Sorting*, it should be straightforward for participants to ignore emotion as they organize the experimental stimuli into groups. If so, the groups formed through the Sorting task would be uniform in their tune composition, with each tune assigned to a different group. This would emerge as a lack of significant tune-emotion interactions in a model predicting grouping behavior.

- *For Rating,* participants should evaluate the meaning of tunes the same way regardless of the concurrent emotional portrayal and its effect on the phonetic implementation of a given tune. This would emerge as a lack of emotion effect.

Interacting hypothesis (H2)

The perceived meaning of tunes partly depends on the accompanying emotional portrayal, such that distinguishing tunes that differ in their linguistic meaning is more reliable when emotional portrayal is held constant. The critical difference between H2 and H1 is the importance of tune-emotion interactions, which is motivated by the findings from Chapters 2 and 3 that tunes and emotions are entangled in production and perception. H2 predicts a similar entanglement between emotional variation and tune interpretation, which is motivated by evidence that tune-emotion interactions were critical to modeling the relationship between production and perception in Chapter 3 (see Findings A, B, and C). If true, H2 is predicted to be supported by the following experimental evidence:

- For *Sorting*, many more clusters than tunes would emerge, with most clusters composed of a unique tune-emotion combinations.
- *For Rating*, judgments of tune meaning associations would be more predictable given a model of tune and emotion together, perhaps with significant tune-emotion interactions.

Next, the Sorting methods and results are presented and discussed, which is followed by the same for the Rating experiment. Following the experiments, a general discussion and conclusion are given.

4D. Methods, Sorting experiment

Sorting task—AFC

As previewed above, the experimental paradigm for the Sorting experiment is a novel implementation of AFC. Specifically, this task involves participants creating groups of audio stimuli (tune-emotion combinations) based on "what the speaker is trying to say by how they said it" while ignoring "differences in the recording that have to do with who the speaker is, like their dialect or how they feel at the time". The purpose of the key instructions, given below, is to focus participants' attention on the function implied vis-à-vis the tune, which requires them to disregard irrelevant information in the signal, especially emotion.

a. In this experiment, you will be sorting sound CLIPS[22] of simple phases, like "that's Lavender", based on what you think the speaker is trying to express.
b. While ignoring differences in the recording that have to do with who the speaker is, like their dialect or how they feel at the time, organize the CLIPS into GROUPS based on what the speaker is trying to say by how they said it.
c. An example meaning is asking a question, which is often marked by a rising pitch at the end of a phrase ("I'm going skydiving?").
d. Another possible meaning is making a statement, which usually accompanies a falling pitch ("I'm going skydiving").
e. Nobody knows exactly how many different communicative functions are carried by speech, but by participating in this study, you'll be helping us solve this mystery.

Along with the novel instruction to ignore emotion, there were five major design changes implemented in this experiment which differentiate it from the typical AFC task (using Clopper 2008 as reference). First, participants were allowed to replay tokens and, if desired, change their mind about any classification decision at any point. Second, participants had control over the sequential order in which they classified particular tokens, because multiple unclassified tokens were presented one at a time through a 'bank' (Figure 1 top-center). Third, participants

[22] Note that in the participant-facing materials, the idea of a tune-emotion token (e.g. an audio stimulus) is conveyed by 'clip'.

controlled the number of groups to sort tokens into[23], with tokens in the same group assigned a uniform color as a way to index group affiliation and as a way to code the meaning associated with groups (Figure 1 bottom). Fourth, participants could play any token or group at any time, to monitor the consistency of tune composition within each group (see play '▶' and 'play all' buttons in Figure 1). Fifth, if participants need to temporarily skip a token or remove it from a group without placing it in another, they may temporarily put it in the 'scratch' area (Figure 1 top left), which holds one token but must be empty before the experiment ends (along with the bank, meaning all tokens were grouped). Because participants were given considerable freedom in how they completed the task, a continuously updated 'checklist' was included to help them track their progress (Figure 1 top-right). The task was complete when every token was sorted and every finished group was played in whole, after which a 'submit' button would activate, enabling the results to be sent to the server and ending the task.

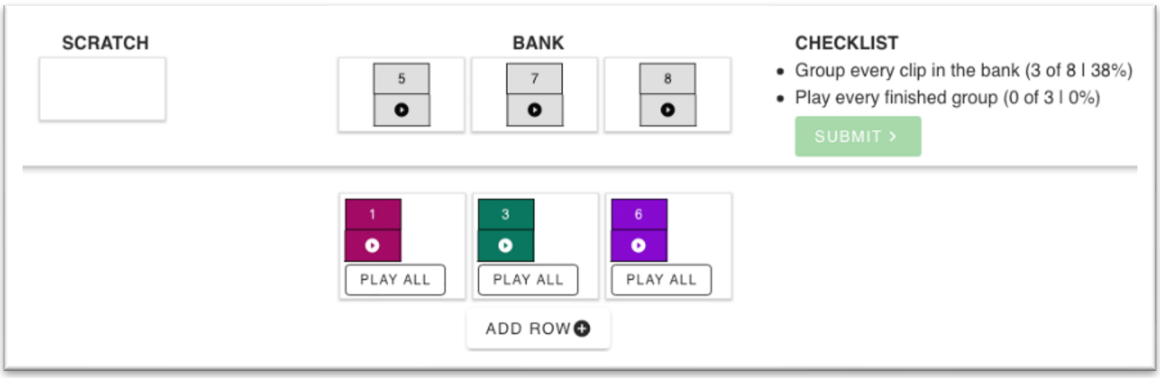


Figure 11: Sorting (AFC) task preview with three groups and one token per group. This same image shown as an example of the task screen as part of the instructions.

[23] The maximum number of groups was capped at 21, which is enough to test all hypotheses, but not the maximum possible number of groups. It is not clear that participants would be able to perceptually track more groups than 21, but it was not tested. Based on the distribution of group counts, this constraint was more than adequate for capturing the variation of interest. This is discussed at greater length in the Limitations (see Ch. 5).

Participants (N=92)

Data from 115 participants were collected from university students recruited from our local Linguistics Subject Pool (N=81) who earned course credit and from remote workers recruited from Prolific (N=34) who were financially compensated for their time. Prospective participants sign up for experiments online and were emailed a link with information on how to begin (preparation instructions and a link to the experiment). Subject Pool members were not screened from participation (due to IRB restrictions), but data was analyzed only from participants who met the screening criteria (below). Prolific participants accessed the experiment in much the same way as Subject Pool participants, although they were matched with the present study only if they met the eligibility requirements, based on screening information provided to Prolific, which was later checked using a questionnaire. To be eligible, participants had to: speak American English as one of their primary languages (bilinguals were allowed); be between the ages of 18 and 65; report not having hearing, speech, or language processing related deficits; and have been raised in the United States.

Exclusions. Based on their questionnaire response, five participants were excluded because they were raised outside of the United States and ten were excluded because English was not their primary language. An additional two participants were excluded due to technical problems. In addition, data from 18 participants was excluded (7 Linguistics Subject Pool, 11 Prolific) who submitted the maximum number of groups allowable by the experiment software (21 groups). Inspection of the results with the maximum number of groups suggested that many of these participants created multiple singleton groups, perhaps treating them as extra scratch areas,

which is counter to the instructions. After exclusions, 86 participants were included in the final dataset.

Demographics. Participant ages ranged from 18 to 55 with a mean of 30.98. For gender, 36 respondents self-reported as female or cis woman, 42 as male, and 2 as nonbinary. In addition to English, participants also self-identified as speakers of the following languages: Spanish (N=16), French (N=3), Algonquian [Blackfeet] (N=1), German (N=1), Hindi (N=1), Korean (N=1), Mandarin (N=1), Punjabi (N=1), Urdu (N=1). Demographic factors were not analyzed in relation to the experimental data in this study.

Design & materials

This study implements a within-participant design, such that all participants grouped all stimuli (tune-emotion recording tokens or "clips") into groups. To covary tunes, emotion, and speaker, the design calls for 64 audio stimuli (8 tunes × 4 emotions × 2 speakers), which were drawn from tune imitations from two voice actors in Chapter 2, from a set of about 1,760 candidate utterances (8 tunes × 4 emotions × 11 speakers × 5 repetitions). Utterances produced in the Neutral emotion condition were not included in order to avoid another distinction in the materials between tunes with emotional portrayals and those without. Two speakers were used was so that multiple versions of the same tune-emotion combination could possibly be grouped together in the task. Another advantage to using two speakers is to expose participants to different implementations of the same tune, especially in different pitch ranges, which is why a female and male speaker were chosen, so the results do not wholly rest on one speaker's rendition of the tunes. The selection of speakers and tokens involved creating multiple potential stimulus sets

based on the following criteria. The primary criterion was congruency between the F0 trajectory of the stimulus and the predictions of the AM model, which involved visual inspection of the F0 trajectory of the tune-bearing words, given in Table 3, which were excised from the sentences for this purpose and task. The reason for not presenting the tune-bearing words in the context of the full sentences they were recorded in was to ensure judgements were not influenced by other lexical items or the F0 trajectories occurring over them. This also saved experiment time, since participants would not have to listen to a full sentence for each item.[24]

Table 3: Tune-bearing words

Emotion	**Word**
Love	Madelyn
Pride	Melanie
Anger	Gallagher
Shame	Lavender

When there was more than one acceptable stimulus based on F0, a secondary criterion was considered, which was how natural sounding the combination of tune and emotion were, as judged by the researcher. Based on these criteria, the stimuli for the experiment were assembled, culminating in a set of recordings that varied by tune, emotion, and speaker, with the F0 trajectories shown in Figure 2. While differences between F0 trajectories are apparent (to varying degrees) across all tune-emotion combinations, based on an informal polling of intonation researchers (the Northwestern University Prosody and Speech Dynamics Lab), and the

[24] To preview, the Rating task does not excise the tune-bearing word from the recorded sentences, because the sentences are presented in context that should help normalize how participants interpret non-tune-bearing words, and participants do not replay items multiple times, so there is no similar time constraint.

judgement of early pilot participants, the perceptual experience of each tune was roughly comparable across speakers and emotions. It must also be acknowledged that, ultimately, perceptual experiences are highly subjective and given that the F0 trajectories for specific tune-emotion combinations are not matched, it is possible that fine-grained acoustic differences could influence the results. See Ch. 5 for more discussion about the limitations of these methods and materials. Nevertheless, it was highly desirable to use natural materials, mainly to be sure that acoustic cues used for the vocal expression of emotion are preserved throughout, whatever they may be.

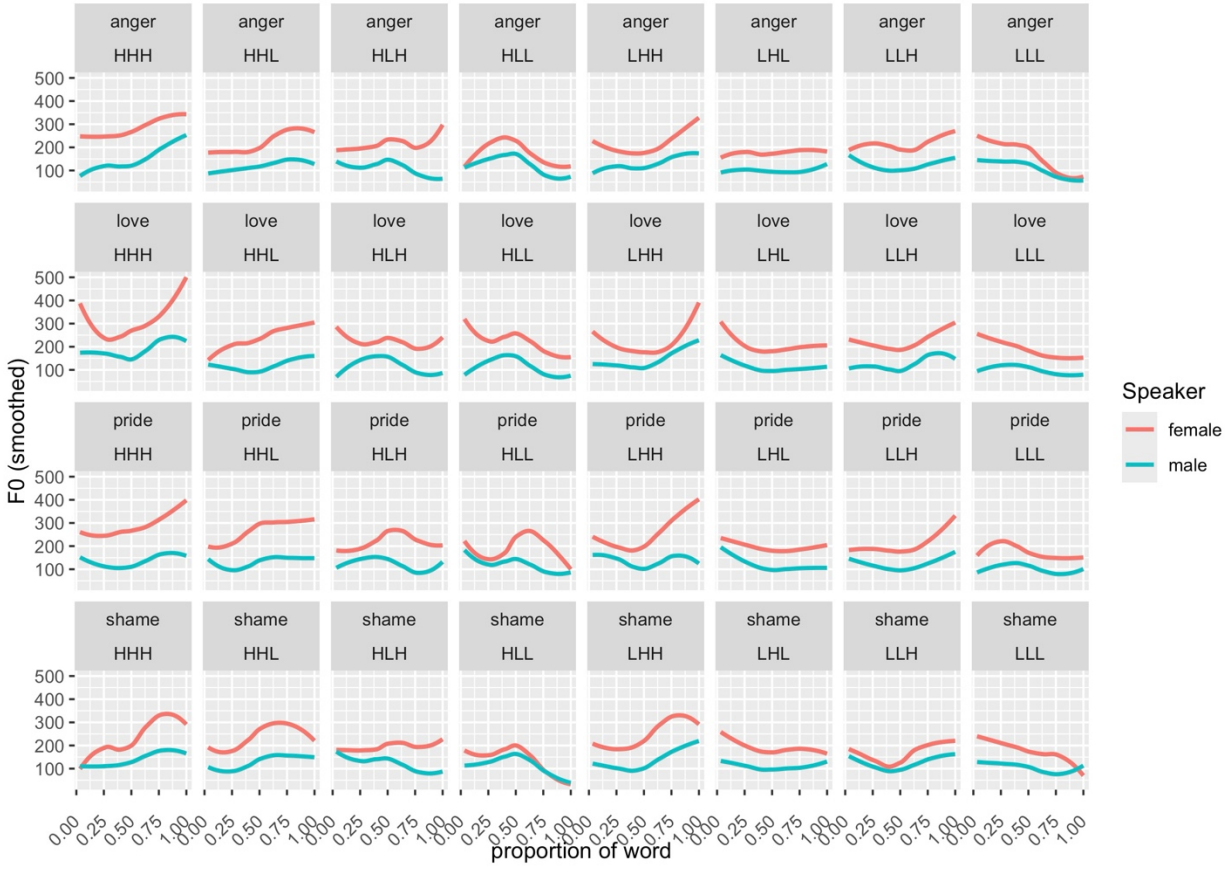


Figure 12: F0 trajectories for all 64 combinations of tune (columns) and emotion (rows) in Hertz over 30 equidistant time samples. Red line indicates one speaker (a female) and blue indicates the other (a male). F0 values were estimated using the STRAIGHT method (Kawahara et al., 2005) as implemented in VoiceSauce (Shue et al., 2009) and were not speaker normalized for this plot to illustrate the speakers' pitch ranges and phonetic details. To reduce the influence of random F0 sampling errors, trajectories were LOESS-smoothed.

Procedure

Preparation. Participants used their personal computers to complete the experiment. To prepare, they were asked to choose a quiet and distraction-free space before beginning and to use quality headphones that were adjusted to a comfortable volume. Upon accessing the experiment, participants had to pass an equipment check that would be tough without wearing properly configured headphones[25].

Instructions. The instructions (given at the beginning of this Methods section) consisted of three parts, (i) a brief introduction to the concept of intonational meaning, illustrated by drawing a comparison between questions and statements; (ii) an explanation of every part of the task screen and its function (e.g. the bank and scratch areas, checklist, how to move and play tokens[26], how to play whole groups); (iii) a final reminder to use headphones for the entire session and to listen to tokens and groups as many times as necessary.

Task screen. Following the instructions, the task screen was presented with three empty groups, an empty scratch area, and a full bank. To complete the task, participants played each token at least once (a prerequisite for moving tokens from the bank), sorted each token into a group (not including the bank or scratch areas), and played each finished group (as a whole).

[25] The equipment check involved counting level tones presented at various frequencies and amplitudes, which was originally developed as a listening test for a prior study by Morrison et al. (2022).
[26] To move a token to a destination on the task screen, participants clicked the top part of a token's icon (its unique ID number), which selected it, then clicked the destination (a group or the scratch area). Before tokens could be moved from the bank, they had to be played, which was accomplished by clicking the bottom part of a token's icon (play button).

Questionnaire. After submitting their finished groups, participants were redirected to a brief background questionnaire to ensure (as a double-check) they meet all eligibility criteria.

Completion time. Because of the less-constrained nature of AFC task, and the novel features specific to its present implementation, a wide range in completion times was expected. Factors impacting the time spent in the experiment were not tested, but likely included: attention to the task, the number of groups being tracked, propensity to refine the solution based on new information, and differences in ability to disentangle the tune from emotional variation. The median completion time was 19 minutes with a range from 6 to 70[27].

Quantitative methods

The final state of the groups generated by participants (e.g. 'solutions') are quantitively analyzed in two ways. The first method is tree-based hierarchical clustering, which is performed to see which tune-emotion combinations tended to be grouped together, and to learn what role emotion played. Clustering has been used previously to analyze participant grouping behavior (Clopper, 2008), specifically to understand what factors distinguish groups. Considering the way participants complete this task, clustering is also a relatively intuitive analysis method for AFC data. The second quantitative method is mixed effect regression (Bayesian GLMMs), which are used to predict whether or not particular tunes are grouped across the data (in line with the preregistration, see https://osf.io/gbk8z/, Experiment. 3), While GLMMs are not as time-tested or

[27] While the experiment did not time out, after 40 minutes a pop-up message would appear to inform the participant about the time lapsed, so that they could adjust their pace if needed.

intuitive as clustering for analyzing AFC solutions, it is desirable to fit a comparable model to those used in prior chapters, and to help determine the statistical significance of effects which arise. This chapter will consider the results of these experiments separately, then draw insights from across the available evidence. Next the quantitative methods are detailed.

Hierarchical clustering. The purpose of this analysis is to use the properties of the audio stimuli that participants sorted into groups in order to understand the factors that predict how participants sorted the items. For example, if participants were perfect in their tune judgements and created one group per tune, the result of this analysis would be eight clusters. The specific clustering algorithm used is the 'partition around medoids' (PAM) method, as implemented in the `treeClust` R package (Buttrey & Whitaker, 2015). In PAM, medoids (centers of clusters) are first iteratively selected to minimize the distance between points ('build' phase), then alternative medoids are tested to optimize the solution ('swap' phase).

Clustering was performed over two versions of the experimental data: one 'simple' version with bare tunes and emotions (8 tunes + 4 emotions = 12 total features), and another 'full' version with pairs of tunes and emotions (36 tune pairs + 10 emotion pairs = 46 total features). Performing 'simple' clustering was desirable in order to directly gauge the contribution of individual tunes and emotions to each cluster and the full clustering was to do the same for tune pairs, which are the data going into the statistical analysis. That said, similar trends are expected to emerge from both analyses of the identical experimental data, but the full model has more degrees of freedom which could allow for a more gradient solution.

Before clustering could be conducted, the optimal number of clusters, *k*, had to be determined calculating the 'gap statistic' for each solution through the `fviz_nbclust()` function

in the R package `factoextra`, which uses Monte Carlo bootstrapping (50 iterations) to compare solutions fit with a range of k (values 2 through 10 were tested). The gap statistic involves summing the distance between all clustered points and the centroid, which measures how 'compact' the clusters are. The optimal value of k is defined as the maximum k whose gap statistic is greater than that of k-1 plus the standard deviation of a random sample, in line with Tibshirani et al., (2001). Gap statistics and determination of optimal *k* are calculated by `factoextra`. The output of clustering is a dissimilarity matrix of the features ('simple' or 'full') based on how the tune-emotion tokens were classified together.

GLMMs. The statistical analysis consists of Bayesian Generalized Logistic Mixed Models (GLMMs), implemented in R using the `brms` package and fit to the same data as the full clustering solution using the generalized formulae given in Table 4 below. The Tune Only Model (TOM) and Emotion Only Model (EOM) are designed to test the statistical significance of tunes and emotions, respectively, on the likelihood that a pair of stimuli (clips) will be grouped together, across all groups and all participants. In other words, these models test whether two clips are more (or less) likely to be grouped together based on tune label alone, or based on emotion label alone The Tune Emotion Model (TEM) includes both tune and emotion as fixed effects. The presence of tune-emotion interactions is further examined using the Tune Emotion Interaction Model (TIM) which expands upon the TEM with added interactions[28].

Table 4: GLMM specifications

[28] The preregistered model closely resembles the TIM, but tunes and emotions were not encoded as pairs, rather as multilevel factors themselves, which led to a considerably more complex model compared to the TIM. The more complex model was run but found to be a poorer fit, and less interpretable in relation to the preceding experiments, than the reported models. Code to fit this model remains in the analysis script but it is not reported.

Name	Abbrev.	Formula
Tune Only Model	TOM	Grouped ~ Tune_Pair + (1|Participant_ID)
Emotion Only Model	EOM	Grouped ~ Emotion_Pair + (1|Participant_ID)
Tune Emotion Model (fixed effects)	TEM	Grouped ~ Tune_Pair + Emotion_Pair + (1|Participant_ID)
Tune Emotion Interaction Model (adds 2-way interactions)	TIM	Grouped ~ Tune_Pair × Emotion_Pair + (1|Participant_ID)

The variables in the model formulae were as follows. The binomial response variable 'Grouped' encoded whether a specific group contained a particular Tune_Pair or Emotion_Pair. The data format is illustrated using a toy example in Table 5, which shows two participants' evaluations of two pairs of tunes and emotions, where they agree that HHH-Love and LHH-Shame should be grouped together, but Participant 1 fails to group HHH-Anger with HHH-Pride unlike Participant 22). Note that stimulus speaker (male or female) is absent from the data format, therefore the comparison is not representing groupings of specific stimuli, but of the tune-emotion combinations themselves.

Table 5: Data format for GLMMs

Participant ID	Group ID	Tune_Pair	Emotion_Pair	Grouped?
1	Group_1_Participant_1	HHH-HHH	Anger-Pride	0
1	Group_1_Participant_1	HHH-LHH	Love-Shame	1
…	…	…	…	…
22	Group_3_Participant_22	HHH-LHH	Love-Shame	1
22	Group_3_Participant_22	HHH-HHH	Anger-Pride	1

The reference level for Tune_Pair was HHH_HHL (to be in line with the F0 models of Chapter 2, which also followed Cole et al. 2023 on this detail) and the reference level for Emotion_Pair was Love_Shame (arbitrarily chosen)[29]. Participant_ID was dummy coded to serve the random

[29] This is the same specification used in Chapter 3 to investigate pairwise perceptual discriminability of tunes.

effects structure shared by all models (random slopes). The full code for models is available through the project data repository (https://osf.io/gbk8z/, see Experiment #3). Convergence of the models was judged on R-hat values, which according to Vehtari et al. (2019) should be at or below 1.01 for all terms if convergence was successful. Additionally, the Bulk and Tail effective sample size (ESS), which mark the 95% credible interval (CrI), should exclude zero to indicate a convincing, likely 'significant', experimental effect (Vasishth et al. 2018). The number of iterations for all models was 10k.

Cross validation method. To determine the model of best fit, the leave-one-out (LOO) cross validation (CV) was conducted using the `loo_subsample()` function of `brms`, which systematically drops observations and refits the model. The same method was used to cross validate the perceptual models in Chapter 3 but using a different function due to issues of computational feasibility[30] when calculating LOO-CV for all observations, given the breadth of this dataset. Therefore, rather than using all observations, the analysis considers 100k samples (arbitrarily chosen, but 250 times greater than the default value 400). The converging model with the lowest LOOIC will serve as the model of best fit.

[30] For the simplest model, calculating exact LOO requires more than 16GB RAM and several days of processor time.

4E. Results, Sorting Experiment

Empirical results

Group count. One of the most straightforward ways to approach these rich results is to ask how many groups participants tended to make. The distribution of group count ranged from a minimum of 3 to the imposed maxima of 20, with a median of 8 and mean of 8.9[31]. A histogram of group count is shown in Figure 3 below. Given the distribution of group counts, the decision to omit groups of 21 seems justified, since fewer than three participants made 17-20 groups each.

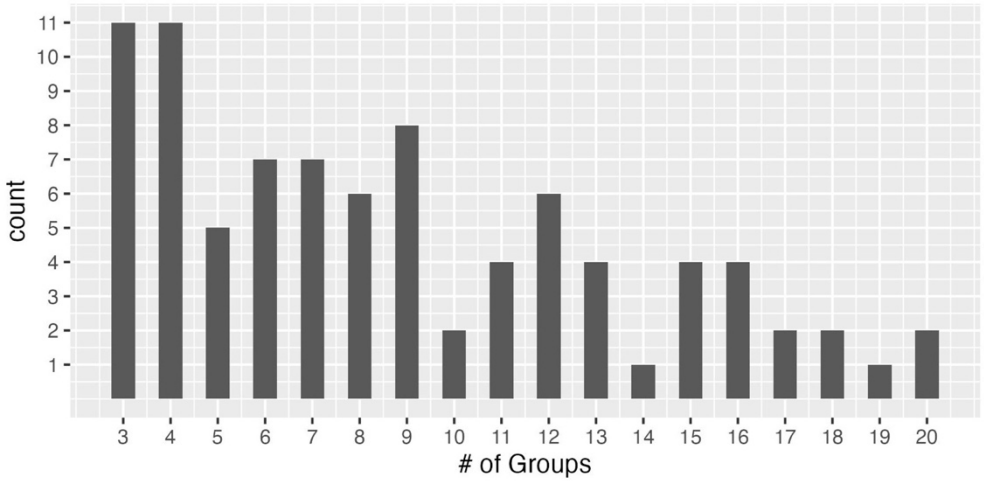


Figure 3: Histogram of participants by the number of groups in their answer

Pairwise frequency. Figure 4 below shows the frequency distribution of tune pairs and emotion pairs identified in participants' groups, which is analyzable on its own merits but also is the basis for clustering (discussed next). Comparing the heat maps in the panels of Figure 4, tune pairs appear to have similar grouping frequencies across emotion pairs. That said, there appears to be a subtle effect, which is more visible in the aggregated version of this data (Figure 5), whereby emotional valence matching status helps predict grouping frequency.

[31] Calculated by calling the R function `summary()` over a vector of every participant's group count.

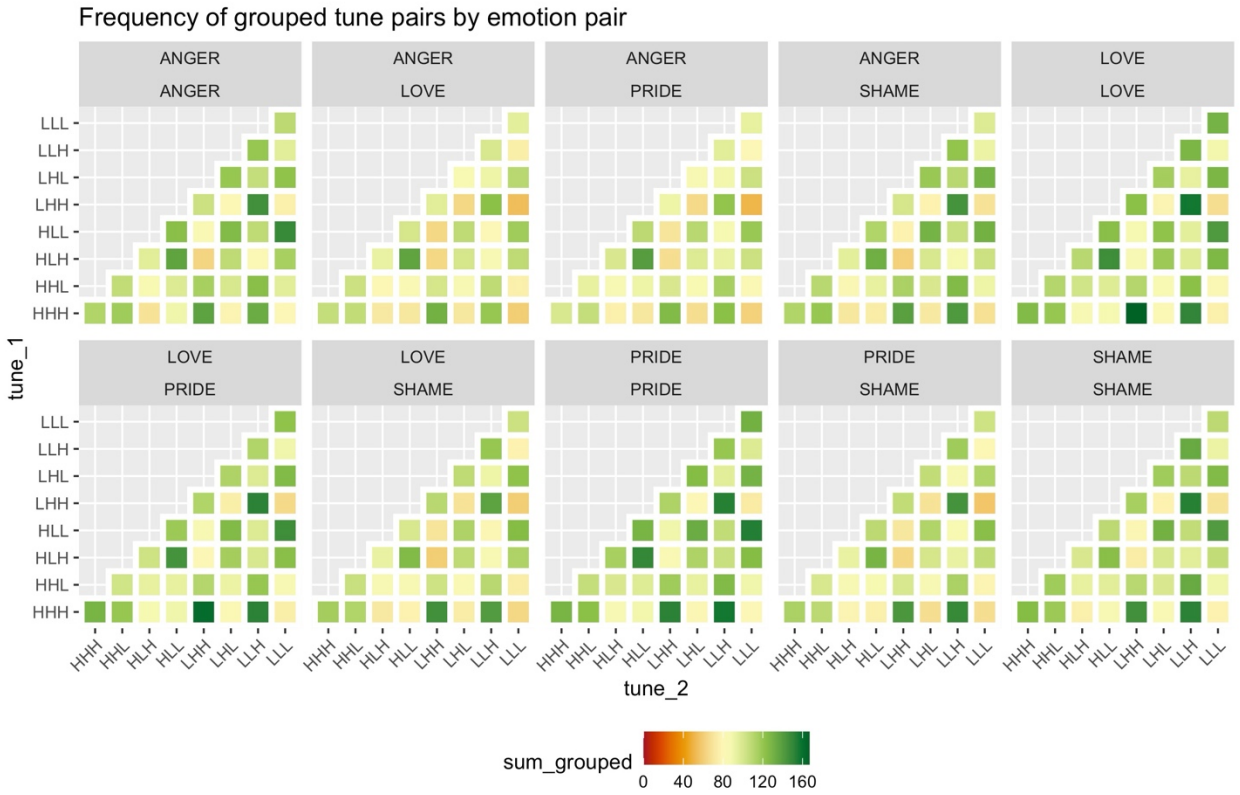


Figure 4: Frequency distribution of unique pairwise combinations of tune pairs (main axes) and emotion pairs (facets). Color scale based on maximum frequency.

Figure 5's left pane aggregates across emotion pair whereas the right pane aggregates across tune pair, which is faceted based on the pair's matching status. Faceting was necessary because there are more ways in which stimuli can mismatch than match, so the pairwise frequency is only comparable within matching status. Continuing the discussion about Valence, the trend mentioned in Figure 4 is easier to see in Figure 5 (right pane) where emotions mismatch (right facet). Note that in this panel, Anger_Shame and Love_Pride are more frequent than Anger_Pride, Anger_Love, Love_Shame, or Pride_Shame. This suggests that similarity in emotional variation (specifically in Valence) may influence perceived linguistic relatedness. Note that when emotions match (Figure 5, right pane, left facet) grouping frequency is uniform, suggesting that any differences are lost after aggregating across tune pairs. Indeed, when considering tune pairs across emotions (Figure 5, left pane) there are visible but modest

differences when tunes match (left facet) versus mismatch, where most variation occurs (right facet). The mismatched tune pairs that were grouped with the highest frequency (dark green) were HHH-LHH, HHH-LLH, HLH-HLL, and LHH-LLH.

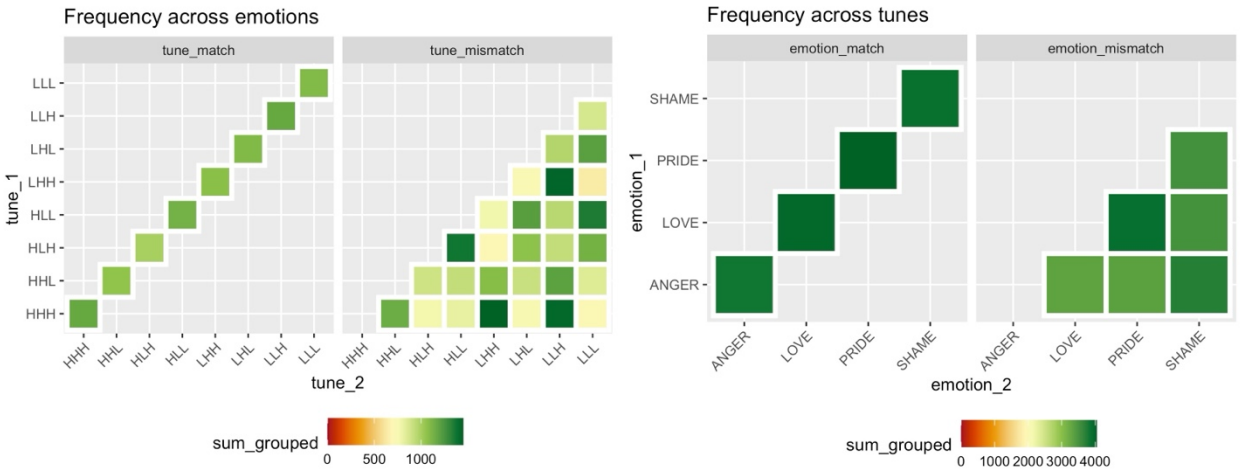


Figure 5: Frequency distribution of tune pairs and emotion pairs, summed across emotions (left pane) and tunes (right pane). Within each pane, data is paneled based on whether both sides of the pair match and colors are scaled based on maximum frequency.

Clustering results

Recall that clustering was performed over two versions of the experimental data: one 'simple' version with *bare* tunes and emotions (8 tunes + 4 emotions = 12 total features), and another 'full' version with *pairs* of tunes and emotions (36 tune pairs + 10 emotion pairs = 46 total features) like previous chapters.

Testing bare or 'simple' features. The gap statistic plot is shown in Figure 6, which shows improvement above k=1 but not $k > 2$, when the standard deviation is considered, which according to the selection criterion means the optimal k=2. In other words, while the value for the gap statistic slightly increases above k=2, those gains never exceed the levels of variation as captured by the standard deviation, suggesting no benefit for increasing $k > 2$.

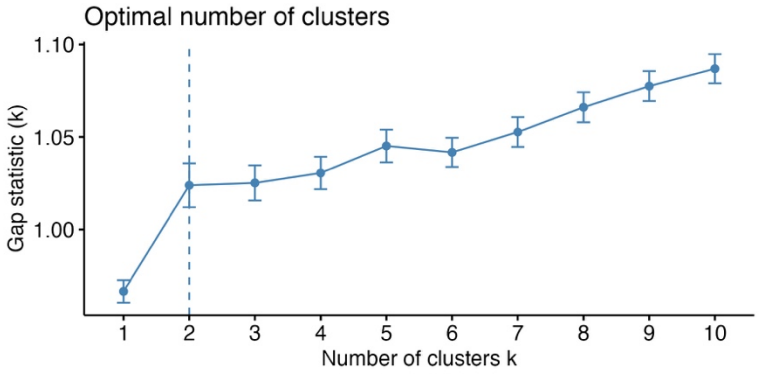


Figure 6: Gap statistic search results for tunes and emotions ('simple' features version). Y-axis indicates gap statistic value, x-axis indicates the number of clusters that was tested, and vertical dashed line indicates optimal k value based on the selection criterion (=2)

Figure 7 shows the proportion of each tune and emotion by cluster as a heatmap, in order to see whether participants created groups of items based on their tune or emotion label. Regarding the tunes (top panel), the figure shows that three tunes dominated Cluster A (HHH, LHH, and LLH), four tunes tended to be in Cluster B (HLH, HLL, LHL, LLL) and one tune was evenly split between clusters (HHL). Within these trends, there is little variation between individual tunes. Regarding the emotions (bottom panel), the figure shows no differences between emotions.

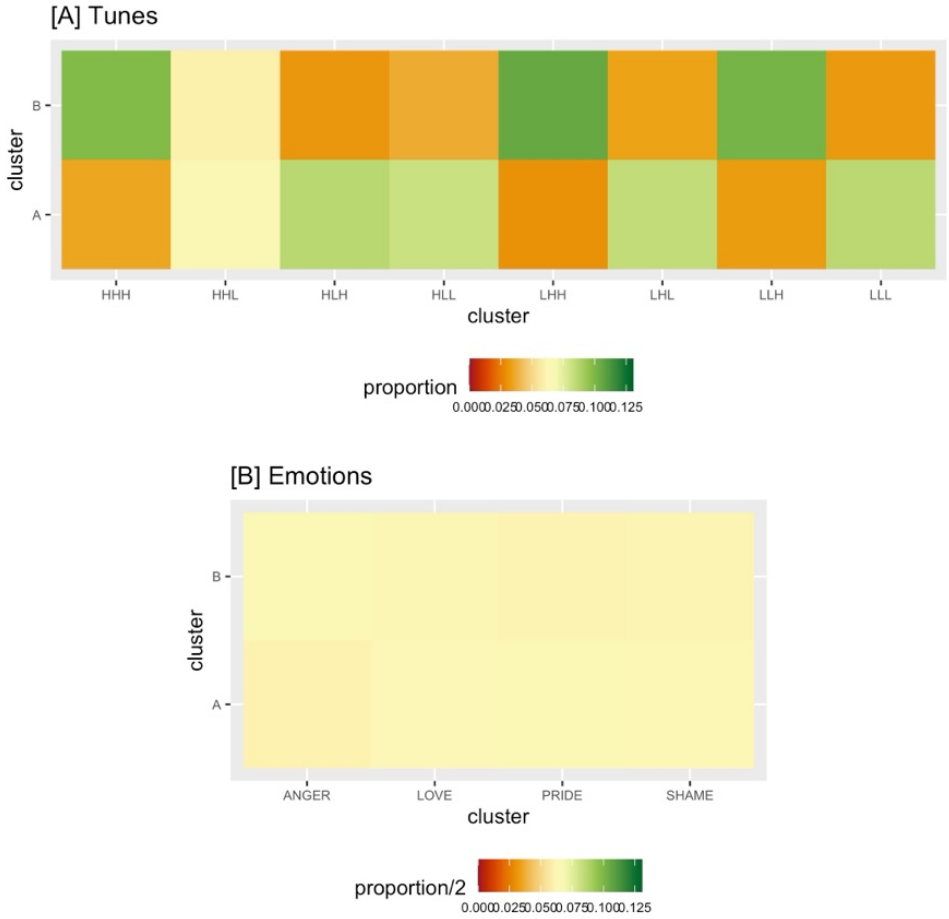


Figure 7: Heatmap of correspondence between pairs of tunes (Panel A; top) and emotions (Panel B; bottom), visualized using proportion of that feature within the cluster. Proportion values for emotions are halved because there are twice as many emotion observations than tune observations due to the experiment design, which makes the color scales comparable. Color is scaled between 0 and the maximum proportion.

Testing pairwise or 'full' features. Using the same selection criterion as the 'simple' features result, the gap statistic for the 'full' pairwise features data was determined to be k=5, as shown in Figure 8. When the same methods as above are applied to the full pairwise data, the optimal k is determined to be five (see Figure 10). Unlike in the simplified model, the gap statistic increases slightly above the optimal k, but note that the gap statistic for k=6 is within the standard deviation of k=5, which is why k=5 is optimal given the criterion.

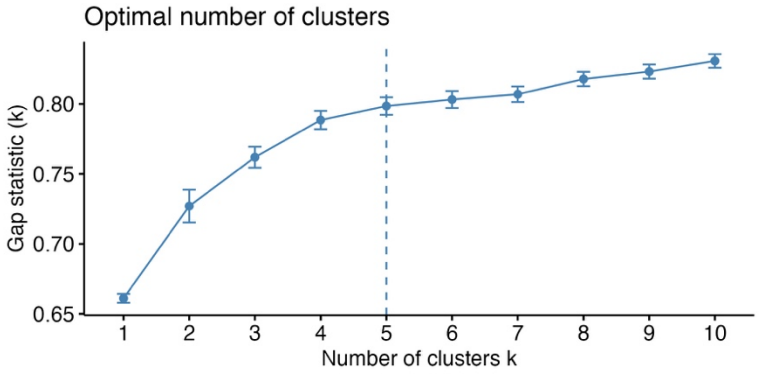


Figure 8: Gap statistic search results for tune and emotion pairs ('full' features version). Y-axis indicates gap statistic value, x-axis indicates the number of clusters that was tested, and vertical dashed line indicates optimal k value based on the selection criterion (=5).

Figure 9 shows the proportion of tune pairs (top panel) and emotion pairs (bottom panel) within each of the five clusters in this result. Unlike the previous clustering result, there is visible variation by tune and emotion pair for each cluster, which makes generalization more of a challenge. That said, Cluster A (HLH, HLL, LHL, and LLL) tended to include tunes that are falling (plus LHL, which does not fall) and end with a low or mid-low F0. Cluster B (HHH, HHL, LHH, and LLH) tended to include tunes that are rising and end with a high or mid-high F0. Cluster C includes all tunes in roughly equal low proportions, but with an even lower proportion of HLH. Clusters D and E include many of the tune pairs that are also frequent in A and B, but in lower proportions, resulting in no green cells for Cluster D and one light green cell for Cluster E. This means that while Clusters A and B can be (in large part) phonologically explained, other Clusters might be better explained by stimulus emotion. This seems to be the case, as Clusters C and D have higher proportions of pairs of the same emotion, and Cluster E is dominated by Love and Pride, both positive Valence emotions.

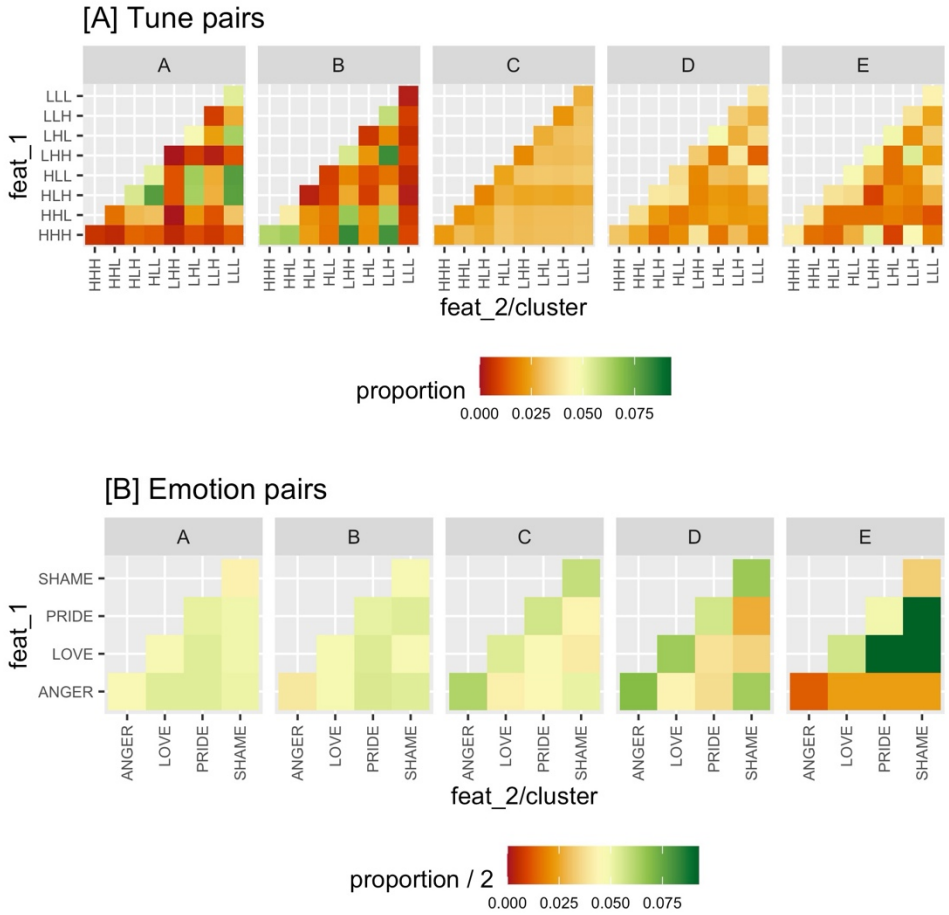


Figure 9: Heatmaps of correspondence across tune pairs (Panel A; top) and emotion pairs (Panel B; bottom) which is halved to make the frequency of tunes and emotions comparable, visualized using proportion of that feature within the cluster. Color is scaled between 0 and the maximum value.

GLMM cross validation results

Recall that the selection criterion for model of best fit is defined as the model that converges with the lowest Leave-One-Out Information Criterion (LOOIC). Table 6 gives the convergence status and LOOIC value (and SE) for the four GLMMs, which highlights TEM as the model of best fit. Based on Rhat values[32], only the EOM failed to fully converge, leading to the following cline of best fit: TEM > TOM > TIM.

[32] EOM's intercept had an Rhat of 1.02 but other terms were lower (converged). Rhat values did not noticeably improve for EOM beyond 2,000 iterations (1/5 of the total) so it is unlikely that convergence problems would be

Table 6: Bayesian GLMM overview

Model	LOOIC	LOOIC SE	Max Rhat	Converged?	Best fit?
TOM	190752.26	747.66	1.00	Yes	No
EOM	192759.03	753.94	1.02	No	NA
TEM	190546.72	746.17	1.01	Yes	Yes
TIM	191033.90	748.31	1.01	Yes	No

Statistical results (TOM & TEM)

Following from the results of cross validation, a detailed look at the results of the TOM and TEM (but not the EOM or TIM) seems well motivated, which is the goal of the remainder of this section. For a full summary of the models, see Appendices B (TOM) and C (TEM) respectively.

Tune only model (TOM) results. The goal of this model is to investigate how tunes alone, setting aside emotional variation, drove grouping behavior—in other words, to predict the organizing factors based on provided solutions. The estimated grouping probability for each tune pair is shown in Figure 10 below. Overall, the probability distribution across tune pairs was graded, with estimates between ~.05 and ~.15 and a group of about five tune pairs that stand out as relatively more likely to be grouped in the same cluster: HHH_LHH, LHH_LLH, HHH_LLH, HLH_HLL, and HLL_LLL. Note all these high-probability pairs have tunes that end in a similar F0 value (high/mid-high, or low/mid-low). Similarly, if we consider the five lowest probability pairs, they tend to consist of tunes that end in a different tone (with HLH_LHH being the

solved by increasing iterations, rather, to preview the position taken in the Discussion, the data might not be well explained by emotion alone.

exception): LHH_LLL, HHH_LLL, LHH_LHL, and HHH_LHL. The matching tune pairs tended to be in the center of the grouping probability distribution.

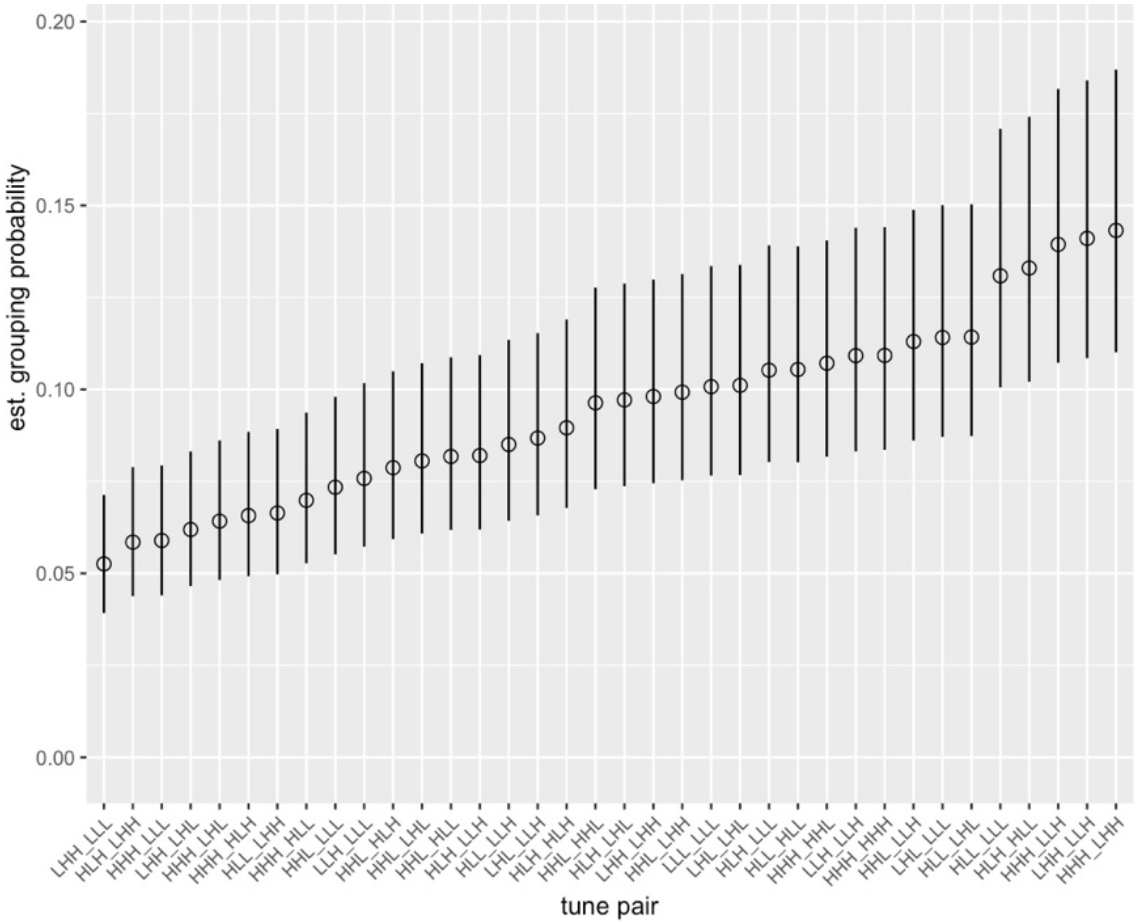


Figure 10: Estimated grouping probability by tune pair (TOM).

Tune Emotion model (TEM) results. The goal of this model is to explore tune and emotion pairs together as main effects. TEM estimates by tune pair are highly comparable to the TOM estimates shown above and (again) are viewable in Appendix C. Differences by tune pair between these models is best explained by the inclusion of emotion predictors in the TEM, the estimates of which are plotted in Figure 11. Specifically, Figure 11 shows that all emotion pairs impacted the grouping probability in a similar positive manner, since all of their CrI values

overlap and are greater than zero[33]. If we choose to interpret the differences between emotion pairs, there is an observable impact of emotion matching, or at least valence matching. The top half of the probability distribution includes Pride_Pride, Love_Love, Love_Pride, Shame_Shame, and Anger_Anger, all of which are pairs of valence-matched (or total match) emotions. Likewise, the bottom half of the probability distribution includes mostly valence-mismatched emotions: Anger_Love, Anger_Pride, Love_Shame, Pride_Shame, and (slightly more probable) Anger_Shame. While the impact of these observations is somewhat tempered by the aforementioned overlap across all emotion pair CrI values, this result leaves open the possibility that some variation is structured by emotion.

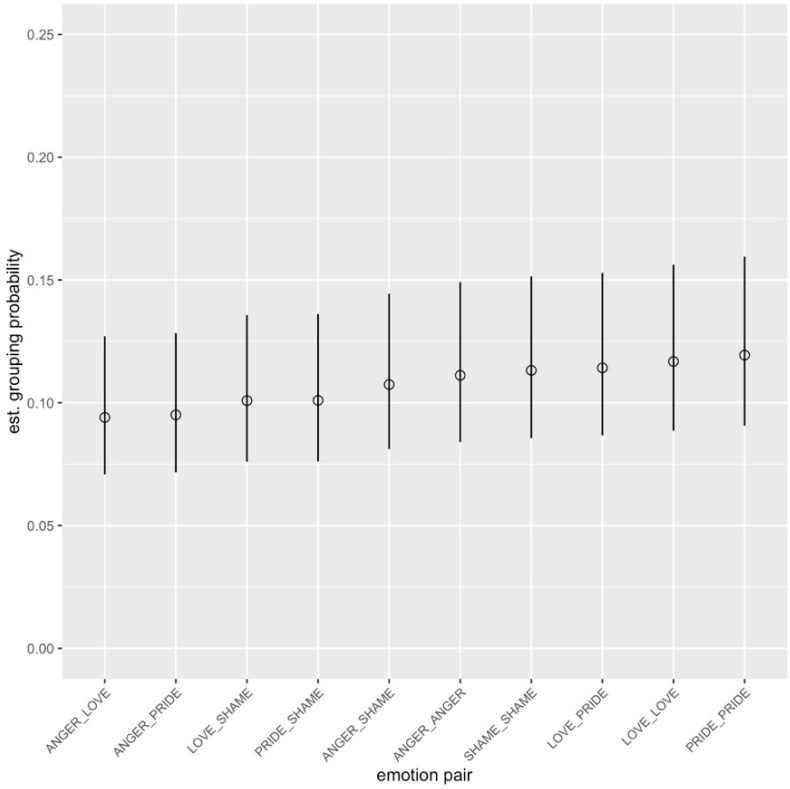


Figure 11: Estimated grouping probability by emotion pair (TEM).

[33] If a particular emotion pair made no impact on grouping probability, it would be expected that its CrI values would extend below zero.

4H. Comparing Sorting and Production results

Recall that k-means clustering over F0 trajectories was one of the quantitative methods used to analyze the production experiment results in Chapter 2 (in addition to GAMMs). Even though the k-means clustering method over the Sorting experiment results is computationally different[34], since the production experiment provided the stimuli for Sorting, their results are naturally compatible. Both clustering analyses speak to the same question of tune distinctiveness given the constraints of emotional variation. Here a metanalysis is conducted to understand tune distinctiveness across the experimental tasks, with the question being how tunes cluster together generally—in this case, both in terms of acoustics and perceived interpretation.

Table 7 below reiterates the relevant clustering solutions by tune, with one clustering solution for Experiment 1, and two competing but similar solutions for Experiment 3. These solutions reveal many parallels which seem most readily explained in terms of F0 characteristics, specifically if a tune has a rising or falling terminus. This trend is most clearly exhibited in relation to Solution II, where tunes that end in a mid-high or high F0 tended to be in Cluster A and tunes that end in a mid-low or low F0 in Cluster B. When a tune in Solution II is in Cluster A, in Solution I it is typically in Clusters {B, C, E}, and in Solution III in Cluster B (also in Cluster C, although it includes other tunes too). Likewise, tunes from Cluster B in Solution II are in Clusters {A, D, F} or {A, F} in Solution I and in Cluster {A} and less consistently also in Clusters {, D, E} in Solution III.

[34] The production experiment k-means clustering was conducted using F0 samples over time which come from estimating instantaneous F0 from across raw audio recordings (see Ch. 2 for details). The k-means clustering for the Sorting experiment does not use time series data, but instead predefined feature sets. In k-means clustering over F0, the distance between clusters is based on F0, and the PAM method was used to calculate distance using feature sets (see Sorting Experiment. methods for details).

Table 7: Comparing tune clusters based on F0 versus participant generated groups									
		Tune/Cluster							
Experiment.	**Solution**	**HHH**	**HHL**	**HLH**	**HLL**	**LHH**	**LHL**	**LLH**	**LLL**
Imitation	I (k=6)	**B, C, E**	**B, C, E**	**A, D, F**	**A, D, F**	**B, C, E**	**A, F**	**C**	**A, F**
Sorting	II 'simple' (k=2)	**A**	**A, B**	**B**	**B**	**A**	**B**	**A**	**B**
	III 'full' (k=5)	**B, C**	**B, C**	**A**	**A, C, D, E**	**B, C, D, E**	**A, C**	**B, C, D, E**	**A, C, D, E**
A tune-cluster correspondence of at least 10% was required for inclusion in this table.									

Taking these correspondences into account, some of the phonologically predicted contrasts from the materials can be reconstructed. Besides HLH, every tune was assigned to multiple clusters in Solution III, and the distribution of tunes among those clusters is highly regular. Tunes show four distributional patterns across clusters in Solution III: A' (Cluster A only; HLH), B' (Clusters B and C; HHH, HHL, LHH, and LLH), C' (Clusters A and C; HLL, LHL, LLL), and D' (Clusters D and E; HLL, LHH, LLH, and LLL). At this level of analysis, when Solution I specifies Clusters {B, C, E}, Solution III is always the B' pattern, and when Solution III is the A' or C' patterns (which are closely related) then Solution I specifies Clusters {A, D, F}, which is another level of complementary distribution. With that in mind, it is interesting to see that, in Solution III's D' pattern, there is no correspondence between how tunes were clustered based on interpretation (Experiment 3) versus F0 trajectories (Experiment 1). Neither the phonological specification of the tunes nor properties of their F0 implementation explain the common interpretation for this set of tunes: {HLL, LHH, LLH, and LLL}. Therefore, it will be interesting to see whether the Rating task separates this particular set of tunes, which will be revisited in the discussion.

4I. Sorting (AFC) interim discussion

Recall that the main goal of this study is to establish whether emotional variation impacts the perceived linguistic meaning of intonational tunes, specifically whether the vocal cues of tune and emotion are interpreted independently (H1) or in terms of a tune-emotion interaction (H2). In the Sorting experiment, participants completed a novel implementation of an AFC task which involved classifying tune-emotion combinations based on perceived meaning, with explicit instructions to ignore speaker emotion. Given that the stimuli comprise different versions of eight tunes, about eight clusters would be expected if H1 is correct.

Empirical findings

Remarkably, the median group size was indeed eight, although the mean was slightly higher (8.59) and the most popular group sizes were 3, 4, and 5. This suggests that while participants did not necessarily reach consensus on the number of distinct linguistic meanings represented in the data, there are far fewer clusters than would expected if the basis was tune-emotion combinations, which would have supported H2. Based on the distribution of group counts in Figure 3, most participants did not track more than nine distinct interpretations. Even in the empirical data, some tunes tend to be classified together (HHH-LHH, HHH-LLH, HLH-HLL, and LHH-LLH), which does not support the idea that tunes were treated as conveying distinctive meanings. To determine the degree to which tunes were grouped together based on phonological versus emotional factors, we turn to the quantitative analyses.

Clustering findings

Two separate clustering algorithm specifications were examined, a 'simple' version based on *bare* tunes and emotions as features, and a 'full' pairwise version that uses tune and emotion *pairs*, in line with modeling conducted in Chs. 2 and 3. The simple version results (k=2) showed three tunes dominating Cluster A (HHH, LHH, and LLH) and four tunes dominating Cluster B (HLH, HLL, LHL, LLL). The tunes in Cluster A all end in a high tone and are characterized by a generally rising F0, whereas Cluster B end in a low or mid-low tone and are generally falling or flat in F0 shape. Based on visual inspection of the emotion-cluster correspondence, clusters had equal proportions of each emotion, which indicates participants successfully ignored (e.g. compensated for) emotion in the speech signal when classifying.

The clustering over the 'full' pairwise features (k=5) resulted in many of the same trends by tune, but gives a more nuanced view of the role of emotional variation for this task. Beginning with the tune pairs, Cluster A contained higher proportions of tunes that end with a low or mid-low F0 (and tend to be falling; LHL was an exception) and Cluster B contained tunes that end with a high or mid-high F0 (and tend to be rising). The other clusters are best characterized in terms of emotion; Clusters C and D have higher proportions of matching emotions (e.g. Anger-Anger) and Cluster E is dominated by positive valence emotions, Love and Pride. While emotions appear to be playing some role in determining how stimuli cluster together, there is no evidence that particular tune-emotion combinations dominate any clusters, which would have supported H2. While the weight of the evidence thus far seems to clearly

support H1, the tune inventory did not clearly emerge from hierarchical clustering, thus the discussion turns to the statistical modeling for a more fine-grained characterization of the experiment results.

Model findings

Based on a cross validation of the four model specifications, the GLMM with tune and emotion as fixed effects without interactions, the TEM, was the best fit, which strongly supports H1. If the model that added tune-emotion interactions, the TIM, was model of best fit, then it would have been taken as evidence supporting H2, but rather the evidence clearly supports an independent understanding of tune and emotion meaning, despite surface level acoustic confounds—more on this in the discussion.

Considering the details of TEM's results by tune, they closely resemble a more interpretable tune-only model, the TOM, so this section presents them together. The TOM and TEM showed that the two tunes in five mismatching tune pairs (HHH_LHH, LHH_LLH, HHH_LLH, HLH_LHH, and HLL_LLL) tended to be grouped together more often than others, which suggests participants did not perceive a robust difference between the two tunes in each pair. Interestingly, all of these pairs of tunes share are similar in their final F0, suggesting that the final F0 is a primary factor guiding participants' judgements. Final F0 also explains the five lowest probability tune pairs, which were mismatched in their final F0 (high or mid-high paired with low or mid-low). Considering the influence of emotional variation on grouping probability vis-à-vis the TEM, there is little distinction between specific emotion pairs, since the CrI values overlap across the board, which means no particular emotion pair (matched or mismatched) boosted probability of grouping. That said, a possible effect of emotional valence may be

emerging in TEM (also visible in the empirical data, see Fig. 7), wherein stimuli with like-valence emotions might be more likely to be grouped together. If the emotion pairs were better differentiated in the model, this effect would be more convincing, and we know from cross validation that adding tune-emotion interactions (TIM) did not improve fit. The final takeaway from the Sorting experiment is that multiple pieces of evidence support H1, while no convincing evidence mounted for H2, which sets the expectations for the next experiment which uses a more constrained methodology (compared to AFC) that ought to provide a more fine-grained perspective on tune interpretation given emotional variation.

4H. Methods, Rating experiment

Task design

The context for the Rating experiment task is a broader study[35] that also included a production task which is not discussed here, due to not being relevant to the present research question about intonational meaning. For the Rating task, participants used a five-point Likert rating scale (Strongly Disagree < Weakly Disagree < Neither < Weakly Agree < Strongly Agree) to respond to the prompts shown in Table 2 given an audio stimulus (a tune-emotion combination). To review, there are six prompts that participants respond to, representing three underlying meaning dimensions often associated with intonation: *questioning*, *committing*, and *floorholding*. Critically, half of the prompts are designed to lead to a positive evaluation of each dimension and half negative in a counterbalanced design, so that rating trends for complementary prompts might be analyzable in relation to one another.

The motivation behind using counterbalanced prompts comes from piloting an early version of the experiment with all 'positive' prompts, which showed strong evidence that participants could accommodate most tune-prompt combinations. Accommodation is a known nuisance variable because pilot participants would endorse clearly contradictory tune-meaning associations, hence the use of 'negative' prompts. Rather than interpret the results for each prompt directly, the analysis will consider both positive and negative pole prompts for each meaning dimension[36] in order to control listener accommodation. There are three possible

[35] In that broader study, the two tasks were interpolated, such that participants would first rate a stimulus according to a prompt using the method below, then produce an imitation of that stimulus using the methods described in Ch. 2. The purpose of the dual-task design (rating and production) was to test if fine differences in the phonetic implementation of tunes correlated with trends in perceived meaning. Due to high degrees of acoustic variation, strong evidence of such a correlation was elusive, but there was limited evidence of structured acoustic variation according to speaker interpretation.

[36] This way of accounting for accommodation bias was developed for the present study and might be novel, based on a literature review, although using counterbalanced prompts seems obvious, and turns out to be quite effective.

general outcomes within each tune-prompt combination that imply varying interpretations about the possibility for tune-meaning associations:

(i) a positive bias (positive > negative), indicating converging evidence (from both prompts) for the endorsement of a particular tune-meaning association
(ii) a negative bias (positive < negative), indicating converging evidence against the endorsement of a particular tune-meaning association
(iii) lack of complementary distribution between prompts (positive = negative), which could indicate that participants are randomly responding to this tune-prompt combination, due to lack of perceived meaning association

Depending on whether or how a particular tune maps onto a particular meaning dimension, these outcomes will emerge through the complementary distribution of responses to the prompts, as illustrated in Figure 12 for the 'questioning' dimension and three tunes. For example, the prototypical questioning tune is LHH according to the literature, so a positive bias expected, meaning that positive responses increase with agreement, complementary to negative responses—see Panel A. For HLL, the prototypical statement, the distribution between prompts is expected to be reversed—see Panel B—and for HHL, which based on prior work on intonational meaning is expected to have no association with the question/statement distinction, a random result is expected—Panel C. Note that emotion is omitted here as a variable in order to simplify the picture, but in the following analysis the same tune might show distributional differences between emotions.

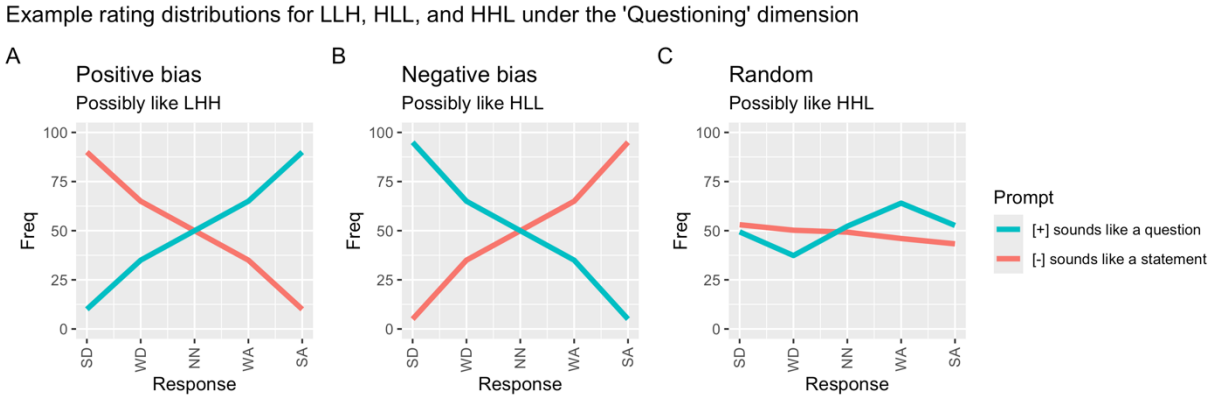


Figure 12: The three main ways that the agreement ratings might be distributed for different tunes, under the questioning meaning dimension. Panel A shows a positive bias, which the literature suggests might arise for LHH for this meaning dimension, panel B shows a negative bias which might arise for HLL, and Panel C shows a lack of a bias (random) which might arise for HLL.

Through the course of each trial, the written context for the target sentence is displayed, the audio stimulus is played, and the prompt is given at the same time the rating scale activates—see Procedure for more details and Appendix A for written materials. In Figure 13, a preview of the task screen in its final state is shown.

Figure 13: Rating task screen in its final state (context and prompt visible; post playback of stimulus; rating scale activated).

Participants (N=147)

All participants in the experiment were recruited from the crowdsourcing platform Prolific, 150 total. As for the sorting task, to be eligible for this task participants had to speak American English as one of their primary languages (bilinguals were allowed); be between the ages of 18 and 65; report no hearing, speech, or language processing related deficits; and be raised in the United States. Those who participated in other experiments (production, perception, sorting) were not eligible.

Exclusions. Three participants encountered technical problems during their experiment session, specifically in the accompanying production task, which prevented their rating data from being successfully uploaded to the server.

Demographics. Based on the Prolific demographic information, the mean participant age was 34.69 years, with a range of 18 to 55 years. More participants self-reported as female (78) than male (69). The vast majority of participants (141 or 96%) were monolingual native English speakers, but six multilinguals were included, namely speakers with English-Spanish (4), English-Chinese[37]-Spanish (1), English-Vietnamese-Ukrainian-Russian (1). Demographic factors were not analyzed for this study.

Materials design

A within-participants design was used such that all participants responded to all 96 items (8 tunes × 2 emotions × 6 prompts) once, blocked by emotion, without repetition[38]. The audio stimuli are a subset of imitative productions collected in Chapter 2 from voice actors, and a subset of the audio stimuli utilized for the sorting task. Whereas the sorting task uses 64 audio stimuli (8 tunes × 4 emotions × 2 speakers), the rating task uses 32 audio stimuli (8 tunes × 2 emotions × 2 speakers). Rather than using an emotion inventory that crosses two levels of Valence and Potency (see Chapter 1) the present study takes the more efficient approach of covarying these dimensions by selecting two emotions that are different along both dimensions. If we consider the Valence × Potency possibility space according to Fontaine et al. (2007)'s

[37] Self-reported as 'Chinese'; this participant is likely familiar with Mandarin.
[38] The lack of repetition of trials was due to the limited time.

model in terms of quadrants, see Figure 14, Love stands alone in the upper left quadrant, which could make it less likely to be confused for other emotions, unlike Shame which has many near neighbors (Sadness, Despair, Fear, Guilt, etc.). Anger was selected to represent the lower right quadrant because of its more extreme Potency value than its neighbors.

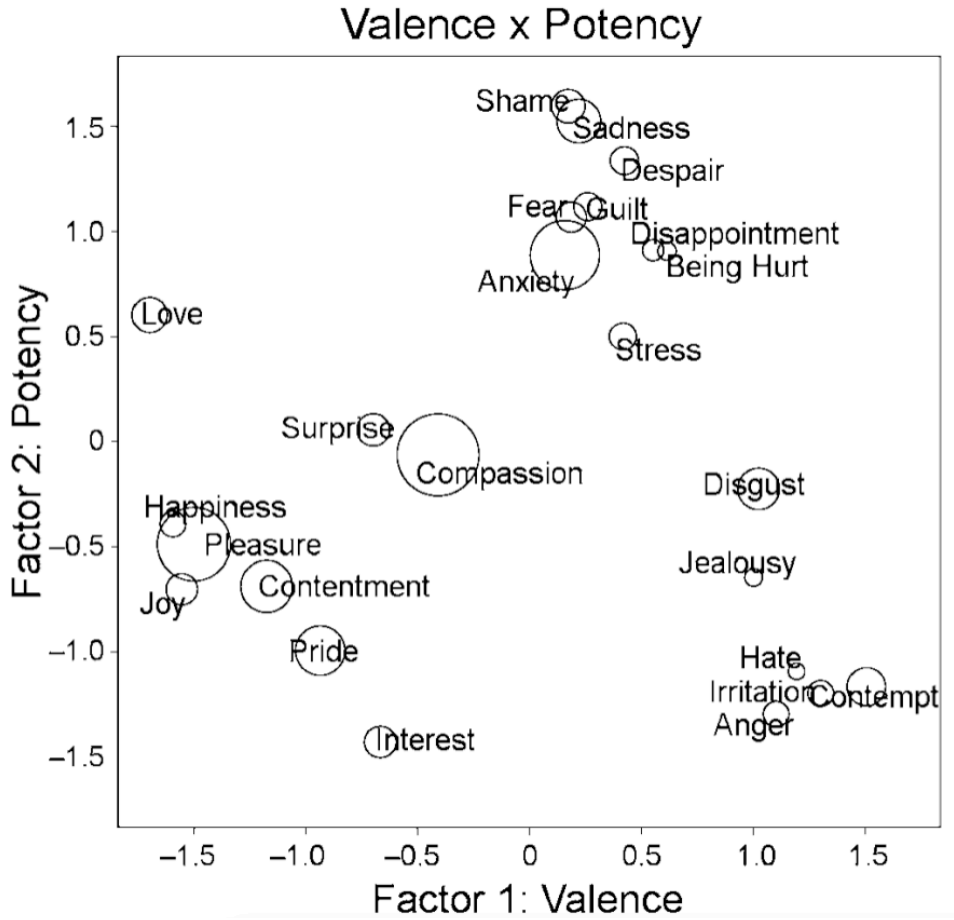


Figure 14: Emotion possibility space (Fontaine et al. 2007). See Chapter 1 for more about the psychometric model that provides the basis for emotion selection.

Summing up, Love and its complement in the emotion design, Anger, were selected to represent the Valence × Potency possibility space. Theoretically, similar results might be achieved with the emotions Pride and Shame, but this was not tested. Therefore, a subset of critical items from the Sorting experiment, consisting of all eight tunes under Anger and Love for two speakers, were

collected. Unlike in the Sorting experiment, the participants heard the full audio stimulus, consisting of a short sentence, rather than only the tune-bearing word (e.g. “Her name is Marilyn”). Figure 15 below shows the F0 trajectories for each stimulus.

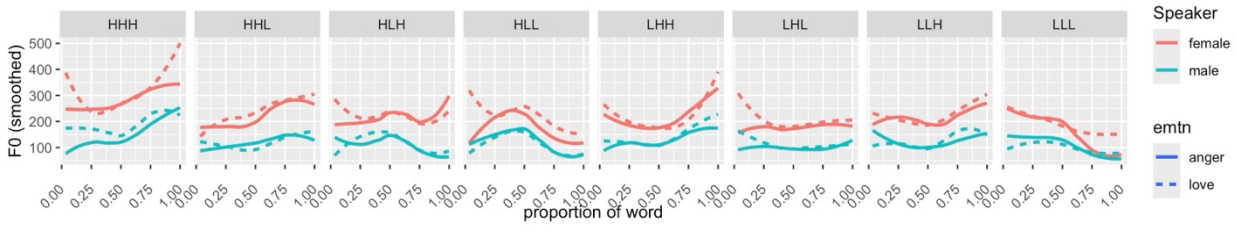


Figure 15: Stimuli F0 trajectories extracted from the tune-bearing target words.

Procedure

Preparation. Before beginning, participants were asked to locate themselves in a quiet and distraction-free place and complete the experiment with a comfortable pair of properly configured headphones (for optimal sound quality, and to aid attention). Additionally, the consent process was completed in this time as well as a short equipment check, identical to the earlier experiment (see Sorting Methods above).

Instructions. After passing the equipment check, participants would continue onto the instructions, the purpose of which was mainly to familiarize participants with the types of judgements they will make in the experiment and how to apply the agreement-based 5-point rating scale. Additionally, it was desirable to focus participants’ attention on the linguistic functions that are part of the token (the tune) over paralinguistic factors (emotion, speaker, etc):

a. This experiment is about the meaning of words in speech.
b. You will hear two speakers (male and female) saying the same sentence in the same way, but in different emotions.
c. Note that the names in the sentences (Melanie, Madelyn, Lavender, and Gallagher) do not convey an inherent meaning, other than to identify a person.
d. However, if said in a specific way, words can convey a variety of different meanings. In written language, some (but not all) of these meanings can be signaled by punctuation:

- That's Mary.
- That's Mary?
- That's Mary…

e. Think about how each of these versions of "That's Mary" sound.
f. In this experiment, first you'll READ the context of the sentence, then you'll LISTEN to the recorded sentence, then you'll RATE it on a scale, and lastly you will SAY your own version.

Practice trials. After the instructions, participants completed four practice trials which use the emotions Shame and Pride, so practice trials excluded critical items. The time limit on practice trials was doubled to ten seconds (which the participant was informed of) in order to lessen time pressure during the familiarization phase. Trial design is identical between practice and critical trails which is addressed in the next section.

Critical trials. After the last practice trial, the experiment begins presenting critical trials. Trials screens are composed of multiple visual elements, as previewed in Figure 13. First the rating scale and prompt appeared alone for 1000ms, then the context appeared, which the participant read silently with no time limit. Showing the prompt first was meant to reinforce the instructions. Importantly, at this point the rating scale is shown as gray to indicate its being inactive. The participant continues by clicking the mouse (anywhere on screen), which initiates playback of the audio stimulus, after which the rating scale activates (indicated by coloration) for 5000ms. The trial ends when the participant either clicks the rating scale or the trial times out.

Questionnaire. After the last critical trial, the participant was redirected from the experiment to a short Qualtrics-based questionnaire to confirm their eligibility, as a double check on screening data the participants provided to Prolific.

Statistical modeling (CLMMs)

Since participants responded using a five-point Likert scale that has an inherent rank order (Strong Disagree < Weak Disagree < Neither < Weak Agree < Strong Agree), the data are statistically analyzed using ordinal regression, specifically cumulative link mixed models (CLMMs). The CLMMs are implemented in R using the `clmm2` function in the `ordinal` package, and were designed to predict the level of agreement that a particular tune conveys a particular meaning dimension. Given the 24 total tune-meaning associations tested (8 tunes × 3 dimensions), and given that each dimension is represented by a positive and a negative prompt, there are six total prompts. Separate CLMMs are specified for each tune-meaning association out of statistical necessity, due to the fact that intonational meanings are not assumed to function independently, but as a system. For example, if *statement of p* implies *commitment to p* for participants, then it would be inappropriate for those two meanings to be different levels of a factor representing the meaning dimension. Another motivation for separately modeling tune meaning associations was the fact the tunes for the current study were generated using phonological rules (from the AM model), which functions to demarcate forms that correspond to contrastive linguistic meanings.

Three models were tested, shown in Table 8, the Tune Only Model (TOM) which models the response based on the prompt (positive or negative) for a given tune meaning association[39]. The Tune Emotion Model (TEM) adds a fixed effect of emotion (Love or Anger) and the Tune Emotion interaction Model (TIM) adds tune-emotion interactions (e.g. HHH × Love). The CLMM results are reported in terms of the estimate, standard error (SE), z-value, and p-value.

[39] Tune is not a term in the CLMM formulae because these models are fit to subsets *by tune* (and meaning dimension, using two complementary prompts), hence the lack of a testable *tune effect* per say. Rather, an effect of phonologically specified tune is expected to emerge or not from this analysis through significance testing the separate models, as described below.

Table 8: CLMM specifications		
Specification name	**Abbrev.**	**Formula**
Tune Only Model	TOM	Response ~ Prompt
Tune Emotion Model (fixed effects)	TEM	Response ~ Prompt + Emotion
Tune Emotion Interaction Model (adds 2-way interactions)	TIM	Response ~ Prompt × Emotion

To compare the quality of fit of the CLMMs, they are submitted to pairwise ANOVA, the result of which is reported in log likelihood value, likelihood ratio (LR) statistic, and Chi-square probability (p-value). The criterion for the model of best fit is the model with the highest log likelihood across tunes and meanings[40].

[40] While a particular model may work better for a certain tune-meaning combination, the goal is to select the model that most accurately characterizes the cumulative results. To this end, the log likelihood will be aggregated by tune-meaning combination (within pairwise model comparisons) to more easily identify the model of best fit for the data as a whole.

4I. Results, Rating experiment

Empirical results

The frequency distribution of responses by agreement level is visualized in Figure 16 below, where tunes are columns, dimensions of meaning are rows, prompts are colors, and emotions are different line types. The x-axis is ordered by agreement level, so if positive prompt responses increase from left to right, it indicates that participants endorse that tune-meaning combination, but if the negative prompt responses do the same, it indicates a rejection. See Figure 12 for an illustration (using simulated data) on how to understand these plots in terms of complementary distribution versus random variation.

An early indication of success, predictions for the example tunes in Figure 12 (LHH, HLL, HHL) are highly recognizable in the aggregated data. For example, consider LHH under the 'questioning' meaning (bottom row), the frequency of positive prompt ratings increases with agreement, in complementary distribution with the negative prompt (positive bias). The pattern we see for LHH is reversed for HLL (negative bias), and for HHL there is a lack of systematicity, indicating more random responses. While the different interpretations by tune is clearly interpretable, there is a lack of a convincing effect of emotion across tunes (note solid and dashed lines usually overlap). That said, there is some visual evidence for emotion-conditioned alternations in some cases, such as LHH-commit and HHL-question which show opposite patterns for the prompts depending on the emotion. This leaves the door open that the perceived function of the tune partly depends on the speaker's emotion—in order to understand whether the observations are statistically significant, the discussion turns to modeling.

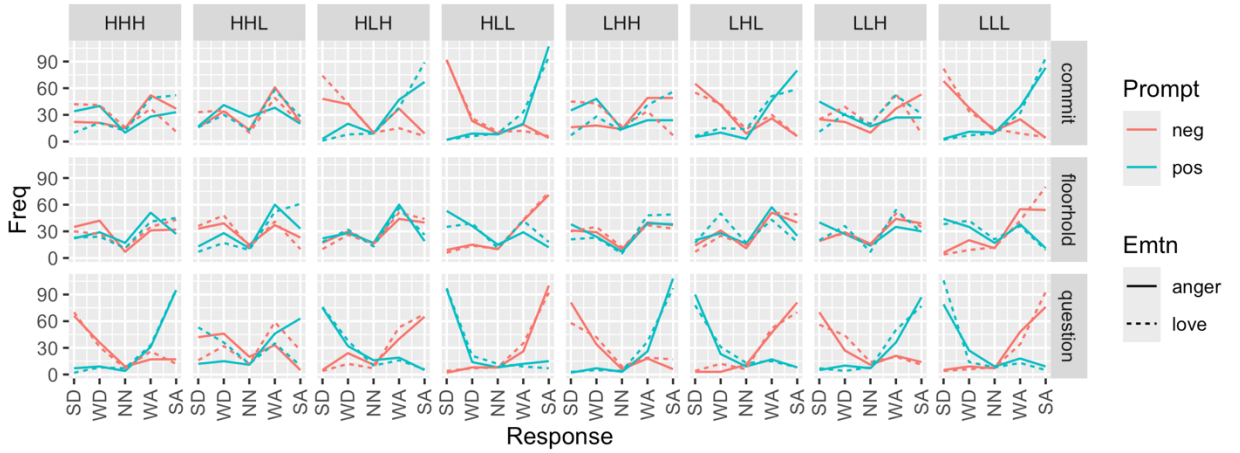


Figure 16 13: Empirical frequency of Likert scale responses by tune (column), meaning (row), prompt (color) and emotion (line type).

CLMM cross validation

To determine the model of best fit, the TOM, TEM, and TIM were run and compared in terms of log likelihood, and by means of an ANOVA (see Appendix D). Both comparisons led to the same conclusion, therefore the analysis focuses on the direct comparison of fit quality on the basis of log likelihood (larger → better). Residual degrees of freedom across model-meaning combinations ranged from 507 to 579, due to small differences in the number of included trials. In Figure 17, the log likelihood (multiplied by -1 to facilitate visual analysis) for each model is given in the form of a stacked bar plot, where each column of bars represents a particular model, each bar (differentiated by color) represents a particular meaning, and each facet represents a particular tune. Based on the overall height of the columns, the quality of the model fit tends to vary by tune, although within tune the differences between models is subtle and for most tunes (five total; HLH, HLL, LHL, LLH, LLL) the model estimates were highly comparable. For the few tunes that exhibited differences between models, TOM and TEM outperformed TIM; note that in the ANOVA comparisons between TOM and TEM show few cases where a statistically significant difference emerged. If Occam's razor is applied to this statistical tie for model of best

fit, the simplest model that adequately explains the data is preferable, which leads to TOM being selected as the model of best fit.

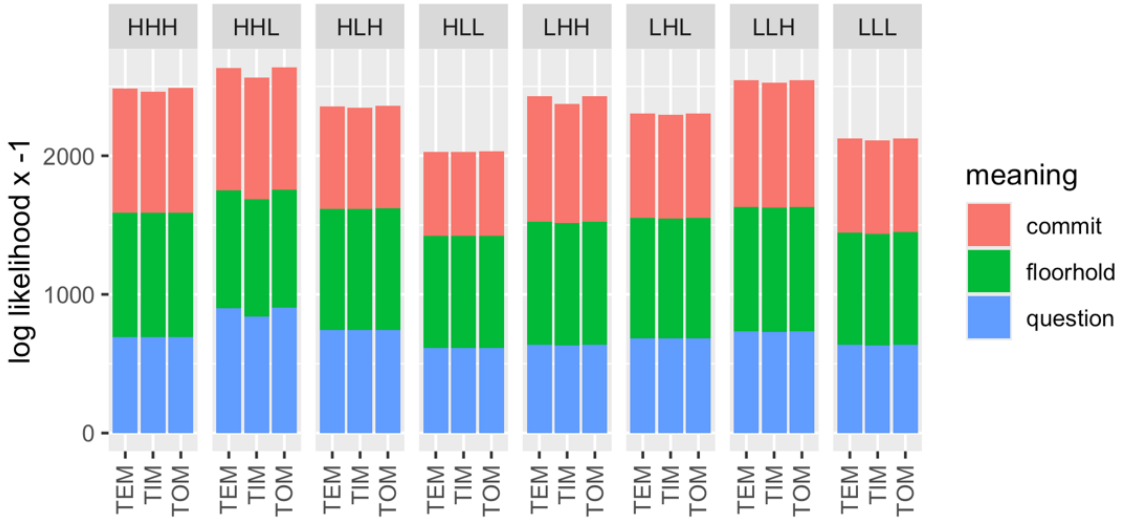


Figure 17: ANOVA results for CLMMs by tune-meaning combination in log likelihood (y-axis). Panels and coloration demark the given tune and meaning, respectively.

Statistical results for CLMM of best fit (TOM)

From the 24 CLMMs (8 tunes x 3 meaning dimensions, with the same model specification), five failed to lead to a significant result, leaving 19 interpretable models which are summarized in Table 9 below. For the 'questioning' dimension, there were significant differences between responses for the positive and negative prompts for every tune, but for the other two dimensions there were five cases where significant differences between prompts were not significant for particular tunes. These 'statically insignificant' cases are potentially instructive, because they could also be a sign of an unendorsed tune-meaning association, which seems possible given how the results align with the meaning literature review, and the random distributions of these cells in the empirical data. For 'committing', HHL, LHH, and LLH did not produce an

interpretable result, along with HHH and LHH for 'floorholding'. Therefore, cases where the model failed to find a significant difference in responses for the positive and negative meaning prompts for a particular tune are reported alongside significant results in the results table.

Table 9: TOM output by tune & meaning

Meaning Dimension	Tune	Estimate	Std. Error	z value	Pr(>|z|)_	p<05
Questioning	HHH	2.15	0.15	14.58	<.001	**TRUE**
	HHL	0.27	0.1	2.54	0.011	**TRUE**
	HLH	-2.13	0.14	-15.66	<.001	**TRUE**
	HLL	-2.58	0.15	-16.97	<.001	**TRUE**
	LHH	2.54	0.15	16.62	<.001	**TRUE**
	LHL	-2.43	0.15	-16.43	<.001	**TRUE**
	LLH	2.04	0.14	15.06	<.001	**TRUE**
	LLL	-2.54	0.15	-16.74	<.001	**TRUE**
Committing	HHH	0.33	0.1	3.18	0.001	**TRUE**
	HHL	0.08	0.11	0.81	0.42	FALSE
	HLH	2.11	0.14	15.56	<.001	**TRUE**
	HLL	2.84	0.16	17.58	<.001	**TRUE**
	LHH	0.18	0.1	1.71	0.087	FALSE
	LHL	1.99	0.13	14.91	<.001	**TRUE**
	LLH	-0.1	0.1	-0.92	0.357	FALSE
	LLL	2.6	0.15	17	<.001	**TRUE**
Floorholding	HHH	0.2	0.11	1.87	0.062	FALSE
	HHL	0.97	0.11	8.69	<.001	**TRUE**
	HLH	-0.32	0.11	-2.96	0.003	**TRUE**
	HLL	-1.43	0.12	-12.01	<.001	**TRUE**
	LHH	0.19	0.1	1.79	0.073	FALSE
	LHL	-0.53	0.11	-4.89	<.001	**TRUE**
	LLH	-0.24	0.1	-2.32	0.02	**TRUE**
	LLL	-1.61	0.13	-12.59	<.001	**TRUE**

The remainder of this section focuses on significant results from the TEM. Interestingly, for 'questioning' and 'floorholding', the significant estimates are both positive and negative, suggesting that participants endorsed and rejected particular tune-meaning associations, but for

'committing' participants did not reject any tune-meaning associations—all significant estimates are positive. Looking at the tunes, except for LHH, which was only significant for 'questioning', every tune had multiple significant meaning associations. Half of the tunes showed effects for all three meanings (HLH, HLL, LHL, LLL) while the others had two (HHH, HHL, LLH). Tune estimates for 'questioning' and 'floorholding' meanings had positive and negative biases, while 'committing' was always positive, which was unexpected but seems to accord with a scalar-based understanding of speaker commitment; there is evidence for *varying degrees* of commitment but not for the absence or failure of speaker commitment. This finding may be of interest for semantics researchers, but deeper discussion about the nature of commitment is beyond the scope of the present study.

In order to help visualize the results in Table 9, the z-values for statistically significant CLMMs were plotted in Figure 18, which provides an overview of tune-meaning associations that were found and their relative strength. This figure is directly comparable to the distributional plots which presented the empirical data; positive bias emerges as positive z-values, opposite negative bias, and larger z-values emerge when the prompts are in more complementary distribution. Likewise, when the bias is not clearly positive or negative ('random' distribution) the smaller z-values emerge, possibly even leading to a lack of significant prediction, in which case the bar is omitted from Figure 18. For example, LHH apparently conveys the meaning of 'questioning' more strongly than HHH based on magnitude of their positive z-values. Additionally, LHH is not associated with other meanings, while HHH has a small positive z-value for commitment. Some tunes have similar very meaning distributions, especially HLL-LLL (visually identical) and HLH-LHL (LHL is slightly less floorholding), both of which are similar to each other (committing > floorholding > questioning).

The other set of similar tunes is HHH-LLH, which differ in the HHH has the additional meaning of (positive) ‘committing’, which is absent in LLH, and LLH has the additional meaning of (negative) ‘floorholding’. HHL was the only tune with a positive value for floorholding. All tunes that had a ‘committing’ meaning also had a ‘questioning’ meaning, which seems in line with Grice (1975)’s Maxims of Conversation, especially Quality, since a speaker *stating p* implicates that they believe *p is true*. Note there are cases where a tune association is significant for questioning but not committing, LHH and LLH.

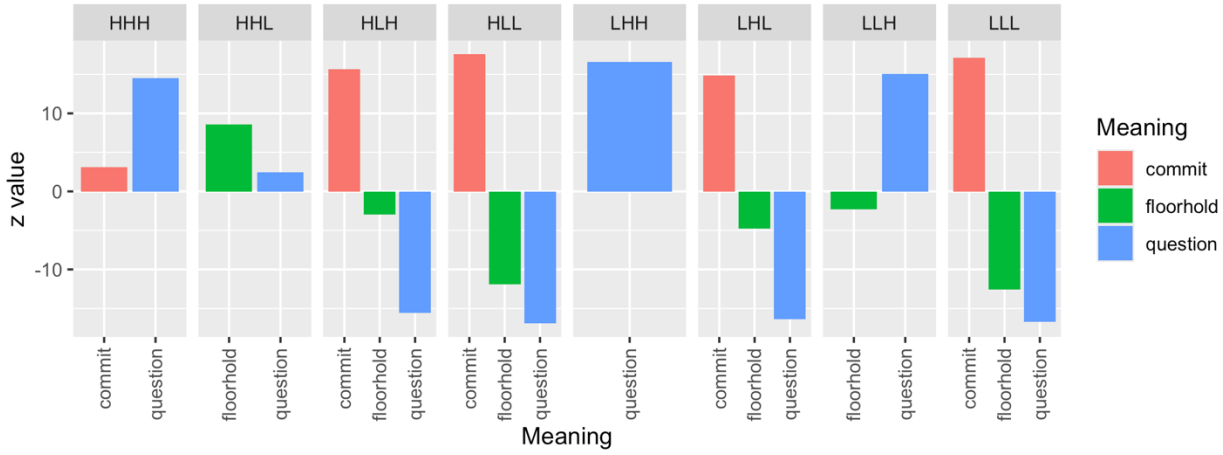


Figure 18: CLMMs of tune-meaning combinations (z values). Displaying significant coefficients only.

Based on the meaning and direction of the effect (ignoring minor differences in strength) five general tune-meaning patterns emerge (ordered by ‘questioning’). In the following discussion section, these patterns are compared to the attested tune meanings from the literature review.

1. questioning only (LHH)
2. questioning > floorholding (LLH)
3. questioning > commitment (HHH)
4. floorholding > questioning (HHL)
5. committing > floorholding > questioning (HLH, HLL, LHL, LLL)

To summarize, while all but one tune was associated with multiple meanings, when we take the strength of the tune-meaning associations into account, half of the tunes impact meaning in unique ways (HHH, HHL, LHH, LLH) and half of the tunes had the same general meaning (HLH, HLL, LHL, LLL).

4J. Ratings interim discussion

Overview

The purpose of the Rating experiment was to gain a more fine-grained understanding (building on the relatively unconstrained AFC task) about whether and now intonational tunes are interpreted in relation to emotional variation in the speech signal. Unlike in the AFC task where participants had to rely on their linguistic knowledge to create the set of possible tune meanings, in this task participants were given three well-established meaning dimensions associated with intonation: questioning, floorholding, and committing. The experiment was designed such that each meaning dimension was represented with two complementary prompts (e.g. question/statement) which participants rated in terms of agreement, based on an auditorily presented tune-emotion combination. The scope of the materials included all eight tunes, and two emotions (Love and Anger) that differed in terms of both key psychometric properties, Valence and Potency. The results were analyzed using visual inspection of the empirical rating distributions which were then submitted to ordinal modeling (CLMMs). If tunes and emotion function independently—H1— then the perceived meaning of tunes will not covary with speaker emotion. However, if tunes and emotions are shown to be interpreted in conjunction with each other—H2— which would emerge in the form of two-way tune-emotion interactions.

Empirical findings

The aggregated empirical results are highly interpretable such that particular tune-meaning combinations clearly have a positive or negative bias in their association. For example, in the 'questioning' meaning dimension, participants strongly endorsed LHH as an exemplar, as evidenced by the strong positive response bias[41], whereas HLL showed a strong negative response bias, and HHL had no clear bias

[41] Recall that 'positive bias' refers to when the frequency of positive prompt ratings increases with agreement level in complementary distribution with negative prompt ratings. 'Negative bias' refers to the opposite pattern; as agreement level increases, so should the frequency of the negative prompt ratings, in complementary distribution with positive prompt ratings. When the prompts for a given dimension are not in complementary distribution, it suggets a lack of tune-meaning association. See Figure 12, which illustrates these possibilities with a toy example.

(random). These specific associations between LHH, HLL, HHL and 'questioning' were predicted by the intonational meaning literature, providing some validation for this novel technique using complementary prompts. Returning to the research question, there was no clear evidence that differences in speaker emotion impacted the ratings distributions in a systematic manner. In a few cases, however, an alternation in the tune-meaning association was observed between emotions, yet the magnitude of these differences raised doubt about interpretability, which helps motivate the following modeling efforts.

Model findings

Given the rating scale response, an ordinal model was appropriate for this data, leading to the adoption of cumulative link mixed models (CLMM) as the method of primary analysis. Each tune-meaning combination was modeled using a CLMM and responses from both prompts together, making the model specification analogous to the previous empirical analysis in terms of positive versus negative bias.

First, the model of best fit had to be determined, which involved comparing models using log likelihood values via pairwise ANOVA (see Appendix D). For most tunes, all CLMM specifications produced similar log likelihood values, but for some tunes the tune-only model (TOM) and tune-emotion model (TEM) performed slightly better than the TEM plus interactions (TIM). Since there were no cases where TOM outperformed TEM or vice versa, despite the relatively greater complexity of TEM, TOM was deemed the model of best fit. This decision was supported by the fact the ANOVA analysis showed few statistically significant differences between TOM and TEM model fits according to a Chi Square test. The lack of an apparent effect of emotion in the statistical model accords with trends in the empirical data, which showed little impact of emotional variation. This finding concords with the outcome of the simplified clustering for the Rating experiment (HHL was an outlier there as well)—see Figure 7. Given the lack of evidence that tune-emotion interactions influenced how participants rated the meaning of tunes using the prompts, the experimental findings thus far clearly support H1 over H2.

While the model of best fit (TOM) had no emotion variable, making it somewhat ancillary to the research question at hand, the details of its results seem to shed new light on particular tune-meaning associations, in terms of direction (positive or negative association, based on which prompt predicts agreement) and magnitude (level of association bias, based on how the prompts together predict agreement). For example, it is interesting that all tunes were associated in some way with the 'questioning' meaning dimension, although the direction and magnitude differed dramatically across the tune inventory. On the other hand, 'committing' was only associated with certain tunes and was reliably positive, and 'committing' only occurred in conjunction with 'questioning'.

Overall, five general tune-meaning patterns emerged in a hierarchical manner: questioning only (LHH), questioning > floorholding (LLH), questioning > commitment (HHH), floorholding > questioning (HHL), and committing > floorholding > questioning (HLH, HLL, LHL, LLL). The first four patterns support a one-to-one or a one-to-many understanding of how intonational meaning is interpreted, while the last pattern supports the many-to-many view. Comparing the first four patterns to the literature review summaries that appeared in Table 1 (italicized below), the level alignment is highly promising:

- LHH is attested to convey *questioning, typical for polar questions, or incredulity*, which is strongly supported by the finding that it was uniquely associated with 'questioning' in a strong positive manner. This is the pattern of results expected if listeners perceived the marker of a polar question.
- LLH is described in prior work as conveying that *the speaker believes listener should already know this information,* which comports with the findings that listeners hear it as primarily marking a question and secondarily as holding the floor. An available interpretation for this pattern is that listeners perceived LLH to mark a rhetorical (non-information seeking) question.
- HHH is attested to convey *questioning, possibly when the answer is believed to be positive*, which seems to perfectly match the finding that the tune primarily conveys questioning and secondarily conveys commitment.
- HHL is attested to convey *elaborating on something that's been previously mentioned*, and the findings showed it to primarily convey floorholding and secondarily convey questioning. Elaboration and floorholding are compatible since the latter implies the speaker is building on previous discourse, but the 'questioning' meaning is unattested.

For the four tunes that failed to clearly convey one of the meaning dimensions, or a hierarchy thereof, the lack of a clear differences in meaning associations seems to be a challenge for the phonological model,

since it shows a lack of meaning distinction between different tunes. Since these patterns persist between both emotions and speakers, the gap between the intonational literature predictions and findings is not due to a particular speaker or recording. Some of the purported meanings for the tunes in this group are contradictory, such as LLL (*Finality, non-predication*) and HLH (*Non-finality, uncertainty, selecting the addressee*). The attested meanings for HLL and LHL are more closely related, being assertive in nature and setting aside secondary associations—incompleteness and reminding, respectively. It is possible that a different formulation of the prompts on which the tune-meaning associations rest would enhance the separation between tunes, which could improve alignment between attested and observed tune meanings, but this was not tested. The only tune with a single attested meaning described in the literature was LLL (Pierrehumbert & Hirschberg, 1990) so the present understanding of its meaning is particularly weak, although it is the topic of active research (Sostarics 2025). Additional work is needed to understand the basis for a significant meaning difference between this set of tunes, and this future direction among others is laid out in greater detail in the following chapters.

4K. Discussion

The goal of the present study was to better understand how listeners interpreted intonational tunes when they occur in the context of emotional variation, which was examined through two experimental methodologies, Auditory Free Classification (AFC) and agreement-based Likert scale ratings. These experiments were radically different in terms of methodology, the types of data they produce, and the available means of qualitative and quantitative analysis for such data, yet their findings converged with respect to the stated hypotheses.

Reviewing predictions

Recall that Hypothesis 1 (H1) posited, "The linguistic meaning associated with a tune and its accompanying emotional portrayal are interpreted independently, such that tune meaning is stable regardless of speaker emotion." For the Sorting task, H1 predicted that tunes would each be assigned their own exclusive groups, and for the Rating task, it predicted no effect of emotional variation. On the other hand, Hypothesis 2 (H2) posited, "The perceived meaning of tunes partly depends on the accompanying emotional portrayal, such that distinguishing tunes that differ in their linguistic meaning is more reliable when emotional portrayal is held constant," and hinges on significant tune-emotion interactions emerging. For the Sorting task, H2 predicted that tune-emotion combinations would be assigned their own exclusive groups, and emotional variation would be an important factor for modeling how tunes were interpreted within in the Rating task.

Hypothesis testing

Based on statistical modeling for both experiments, tune-emotion interactions did not help predict the perceived meaning of tunes, which helps confirm that intonation and emotion are not only conceptually distinct, but are able to function independently as cues to linguistics and paralinguistic meaning, supporting H1. There was little evidence that speaker emotion led participants to classify, or rate, tunes

differently, which suggests they successfully disentangled the relevant cues from the speech signal, and compensated for emotional variation in their linguistic interpretation.

Between the experiments, the Sorting task showed a greater impact of emotional variation, but the tunes were nonetheless the primary driver of participant responses. As a secondary factor in the Sorting task, the influence from emotional variation was best characterized by differences in Valence, which is in line with what was found in production. Another explanation for why Ratings more clearly supported H1 is that the tune-bearing words were presented in multiple layers of context; the tune-bearing words were heard in the context of a sentence, and those sentences were accompanied by written contexts (Appendix A), both of which could help support tune meaning. For the Sorting task, participants heard isolated tune-bearing words and had to rely more directly on their linguistic knowledge, to think of a context where the tune is appropriate, and then infer the generalized meaning—all of which amount to a (broadly speaking) more demanding experiment. Despite such challenges, different tunes were reliably perceived to convey contrasting meaning in both experiments.

Linguistic implications

Given that the present study adopted formalisms from intonational phonology and emotion research, the findings have broad potential ramifications in a variety of research areas, but the present discussion focuses on the linguistic implications (broader implications are saved for Ch. 5).

Intonation theory. The strong showing for H1 within these findings gives long needed validation to assumptions within the field, particularly conceptual distinctions between intonational meaning and emotion. Despite the presence of surface level (acoustic) entanglements that participants do not find challenging to decode, to understand the speaker's linguistic message. For intonational phonology, these findings fill a critical gap between phonological predictions from the AM model and empirical data. Although the tunes predicted by the AM model were observed to be interpreted in a highly robust manner

regardless of speaker emotion, there was evidence that some sets of tunes were not conclusively differentiated in their perceived meaning. Based on the evidence from other tunes that listeners can attend and use fine grained acoustic details that distinguish neighboring tunes, even in the context of emotional variation, the lack of evidence for all predicted tune distinctions in the meaning domain points to the need for more empirical work on intonational meaning. Because the tune inventory was generated from a phonological inventory of intonational features, and not perceived meaning, it is also possible that tunes are hierarchically organized, such that related tunes may convey the same meaning yet be phonologically distinct. While such a conclusion is possible given the experimental results, more research is needed to substantiate such a claim, and it would seemingly necessitate a serious revisit of the AM model. Additionally, there appears to be a mismatch between some patterns of results for particular tunes and attested meanings in prior literature, which deserves clarification through further research.

Intonation research methods. Both experiments in this chapter piloted novel experimental methods that were designed to study intonational tunes that were produced in combination with emotional portrayals, which is a departure from typical intonation experiments. One of the broader goals of the current project is to gather evidence for intonation distinctions in more naturalistic data, and these experiments show a path forward. The Sorting task is highly flexible and can be used to explore other topics within intonation that have been challenging due to the presence of acoustic variation, such as how dialects are reflected within intonation. The Rating task used a novel technique of counterbalanced prompts that was critical for identifying tune-meaning associations while accounting for listener accommodation. The approach taken in the Rating task could easily be extended to other tunes, prompts, and research questions, and is particularly well-suited to crowd-sourced data collection. In the same vein, the present study developed data analysis pipelines and statistical techniques to process and model the diverse (and rich) outputs of these experiments, which involved significant resource investment. Future research can make use of these

stimuli, software implementations, and analysis scripts, which have all been designed to be easily extended to other research questions that deal with noisy and more-naturalistic data.

4L. Conclusion

The present study examined whether listeners evaluate the linguistic meaning of intonational tunes independent of speaker emotion, or in conjunction through tune-emotion interactions. It comes in the context of a broader project that is focused on searching for evidence of intonational contrasts, represented by different tunes, by jointly modeling tunes with speaker emotion. The preponderance of evidence supports the conclusion that a phonologically specified tune generally conveys the same intonational meaning regardless of the speaker's emotion. There was no convincing evidence that tune-emotion interactions contributed to the meaning distinctions participants perceived in the tune-emotion combinations. This underscores how robust the intonation system is, despite high levels of acoustic variation. That said, as in production (Ch. 2) and perception (Ch. 3), certain sets of tunes tended to elicit the same interpretative responses from participants, which emerged as different tunes being classified together in the Sorting task and being interpreted similarly in the Rating task. Taken together, these results point to some weakness in the AM model, since phonologically distinct tunes have obvious overlaps, especially in their perceived meaning. For tunes that overlap in their meaning function, other dimensions of distinction need to be empirically shown to separate the tunes—ideally while also controlling for speaker emotion—or the idea that phonologically distinct tunes also have distinct meaning has to be revisited.

Chapter 5: Discussion

5A. Synopsis

Intonation is one of the main ways pragmatic meaning is conveyed in speech, yet linking the phonological specification of intonation to its phonetic realizations, and linking those to linguistic meaning, has been a longstanding challenge. This problem runs deep in the field of intonation, because it impedes our efforts to evaluate, and thereby improve, our scientific understanding of intonation and by extension intonation theory. This thesis adopts the view the problem stems, in large part, from the level of variation that emerges in the key acoustic correlates of intonation, primarily F0. Based on this rationale, this research explored a novel approach to model the sources of acoustic variation, through joint consideration of the contributions from intonational phonology (phonetically instantiated) and the vocal cues of speaker emotion. Specifically, this thesis tested for distinctions amongst intonational patterns of Mainstream American English (MAE) predicted by the Autosegmental-Metrical (AM) model of MAE, which posits a system of phonologically contrastive tone features at the end of a prosodic phrase, called intonational tunes, phonetically realized in the form of distinct F0 trajectories (Pierrehumbert 1980, Ladd 2008).

Though speaker emotion itself is not the scientific focus of this project, a key feature of the present work is the formalization of speaker emotion within a mainstream model of psychology (Fontaine et al., 2007). The emotion formalism was central to the design of the experimental manipulations, in order to efficiently explore the possibility space, and in the analyses, since it contains a unifying framework for relating the observed effects of different emotions. The basis of the emotion model is a small set of psychometric dimensions that define

the possibility space for emotion distinctions, therefore, the structure of the present thesis can be correctly conceptualized as crossing the set of fundamental phonological contrasts for MAE intonation with a set of key dimensions of emotion. The research, once appropriately motivated (Ch. 1), was divided roughly[42] into three phases, production (Ch. 2), perception (Ch. 3) and interpretation (Ch. 4), amounting to four total experiments which are described in Table 1. Across the experiments, the total number of participants was N=424 and five distinct quantitative methods were applied to the highly varied data types (audio, discrimination judgements, interpretive classifications, ordinal rating scales). To the greatest extent possible, the analyses were accomplished using open-source software combined with custom processing, modeling, and data visualization scripts which are being released as open source, for scientific transparency, to facilitate replication, as well as to contribute to the available code base for future empirical intonation research. The remainder of this section gives a broad characterization of the work and findings of each phase.

Table 1: Overview of experiments

Phase	Experiment.	Method	Description	Participants *(N=424 grand total)*	Quant. Analysis
I: Production	**#1**	Imitative production with emotion portrayal	Participants imitate pitch-resynthesized model tunes while portraying a specified emotion and in a condition with unspecified (Neutral) emotion.	• Voice actors: N=13 • Subj. pool: N=19 *N=32 total*	• k-means clustering (time series) • GAMMs

[42] 'Rough' because of dependencies between these separate modalities of intonation. For example, *production* involved speakers' imitation of *perceived* tunes as instantiated in F0-resynthesized model recordings. *Interpretation* also involved perception, but of tune differences as they were naturally produced with an emotional portrayal.

II: Perception	**#2**	AX Perceptual Discrimination (2-alternative forced choice)	Participants classify tunes produced with emotional portrayal (stimuli drawn from Experiment. 1) as same or different, ignoring emotion.	• Prolific: N=153	• Bayesian GLMMs
III: Interpretation	**#3**	Auditory Free Classification (AFC)	Participants 'sort' tunes produced with emotional portrayal (from Experiment. 1) into groups based on perceived meaning function, ignoring emotion.	• Prolific: N=24 • Subj. pool: N=68 *N = 92 total*	• k-means clustering (PAM) • Bayesian GLMMs
	#4	Likert scale rating of tune meaning (five-alternative forced choice)	Participants rate how well particular tunes produced with emotional portrayal (from Experiment. 1) convey previously attested meaning functions, ignoring emotion.	• Prolific: N=147	• CLMMs (ordinal regression)

Phase I: Production

Objectives. This phase was concerned with the acoustic-phonetic implementation of intonation, considered through the lens of variation due to emotional portrayal, in order to empirically test predicted distinctions from intonational phonology, specifically the AM model. Special considerations were required in order to control intonation while varying the portrayed emotion and to analyze them in conjunction, in particular: (a) joint modeling of linguistic and emotional variables to predict F0 trajectory variation, (b) adoption of a formalized emotion model, and (c) recruiting trained voice actors as participants.

Experiment. In a nutshell, the experimental task involved participants reading a motivating context for the tune-emotion combination before (critically) giving their imitation of a model tune, while portraying a specified (by name) emotion. This method extends earlier successful

efforts (Cole *et al.*, 2023; Steffman *et al*., 2024a, b) to analyze tune distinctions while holding emotional variation at a minimum—akin to the Neutral condition in this study's design. The main justification for dedicating resources to recruiting voice actors as participants was the desire to acoustically capture some of the conventional vocal cues of emotion, so as to better understand its possible interaction with linguistically specified intonation. Quantitative analysis relied on extracted F0 trajectories, measured in terms of similarity between F0 trajectories of imitated tunes (k-means clustering), and through direct modeling of the contributory factors that predict variation in F0 trajectories from intonation and emotion as factors (GAMMs).

Quantitative analysis. First, the clustering analysis showed that F0 trajectories produced as imitations of specific pairings of tune and emotion (eight tunes and five emotional conditions yields 40 distinct combinations, total) were best modeled as representing six distinct F0 patterns, corresponding to the six emergent clusters shown in Figure 1. This finding points to a reduction in the number of tune distinctions predicted by the AM model, similar to what was found in the prior work by Cole, Steffman and colleagues (*op. cit.)*.

Evidence for the robust nature of tune encoding across emotions comes predominantly from the difference GAMM analysis showing that the phonetic implementation of tunes conforms to AM model predictions. It was concluded that differences in the phonological specification of tunes are the primary determinant of F0 trajectories; emotion portrayal exerted a secondary effect on F0 variation, which was largely uniform across tunes for a given emotion. This means that the F0 trajectory for a given tune-emotion combination was largely predictable despite random (e.g. at the level of speaker) variation, which is also typically problematic for empirical intonation research. The level of distinctiveness in F0 between tunes was similar across emotions, with weak evidence for slightly more distinctive tunes within Neutral versus the

emotions, but there was no other supporting evidence to suggest specific emotions led to specific impacts. An informative outcome from this analysis is that Neutral was not necessarily the condition that resulted in the most numerous and clearly distinct tunes; rather, certain tune distinctions were stronger under certain emotions.

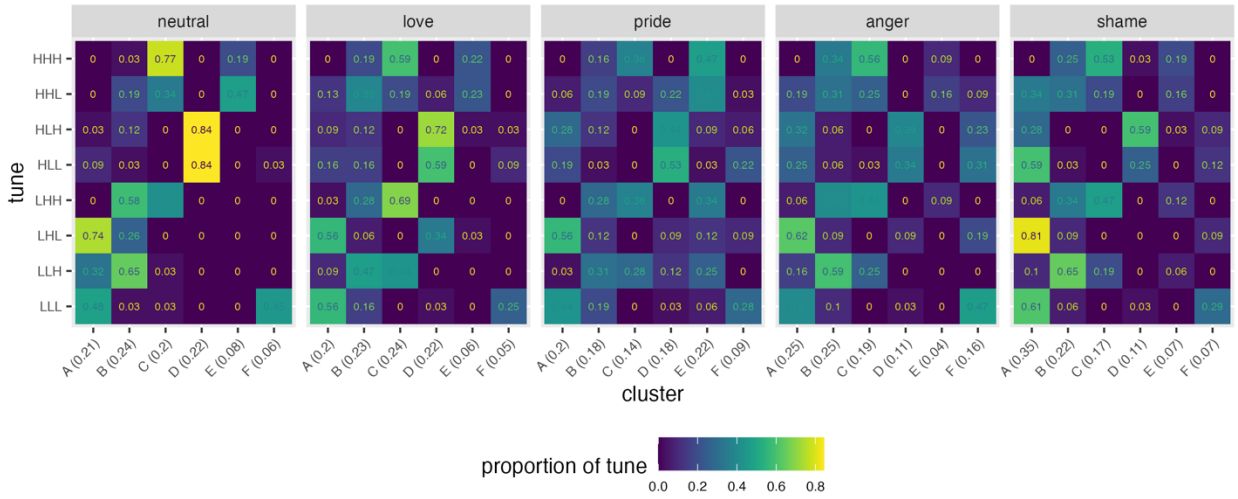


Figure 1: Heatmap showing the composition of each cluster from the clustering analysis in terms of the tune label of the stimulus that was the intended target of imitation for each token. Panels display results from separate clustering analyses for each emotion condition. The number and shading for each cell indicates the proportion of imitated tokens of the given tune (row) grouped in that cluster (column). Within each emotion panel, each row (a tune-emotion combination) sums to 1 . Cluster labels shown with number indicating the proportion of total tokens in the specified emotion condition that were assigned to that cluster and sum to 1 within emotion. Repeated from Fig. 7B in Ch. 2.

Intonation and emotion codetermine F0. Considering the predicted F0 trajectories for each tune by emotion, shown in Figure 2, a notable pattern emerges in which the F0 trajectories are (fairly transparently) modulated by the Valence of the emotion being portrayed. For negative emotions (Anger, Shame), trajectories were lower than for positive emotions (Love, Pride). By analogy, if tunes and emotions were part of a function determining F0, it is as if the intonational tune is the result of a function that returns F0 (y) for a given time point (x), and valence helps determine the relative height in F0 space (y-intercept). This gives us an unexpected view into how perceptual

compensation for emotional variation might work that seems to bode well for explaining how listeners may be able to recognize underlying tune distinctions despite emotional variation, which the next section elaborates upon (e.g. perceptual discrimination).

By and large, the strategy of using emotional variation as a lens into the phonetic realization of intonation appears to be successful, which advances the idea that, although intonation and emotion are entangled in the speech signal, critical intonational distinctions such as those predicted by the AM model tend to be preserved. In other words, although intonation and emotion are conveyed through a common acoustic parameter, F0 (among others), their phonetic implementation tends to shift F0 values in predictable ways (given a particular emotion), rather than creating second order forms through tune-emotion interactions. For example, more positive Valence tended to elevate the mean F0, contra negative Valence, which is a type of modulation that, in account of applying to the whole F0 trajectory, generally leaves critical tune distinctions well preserved.

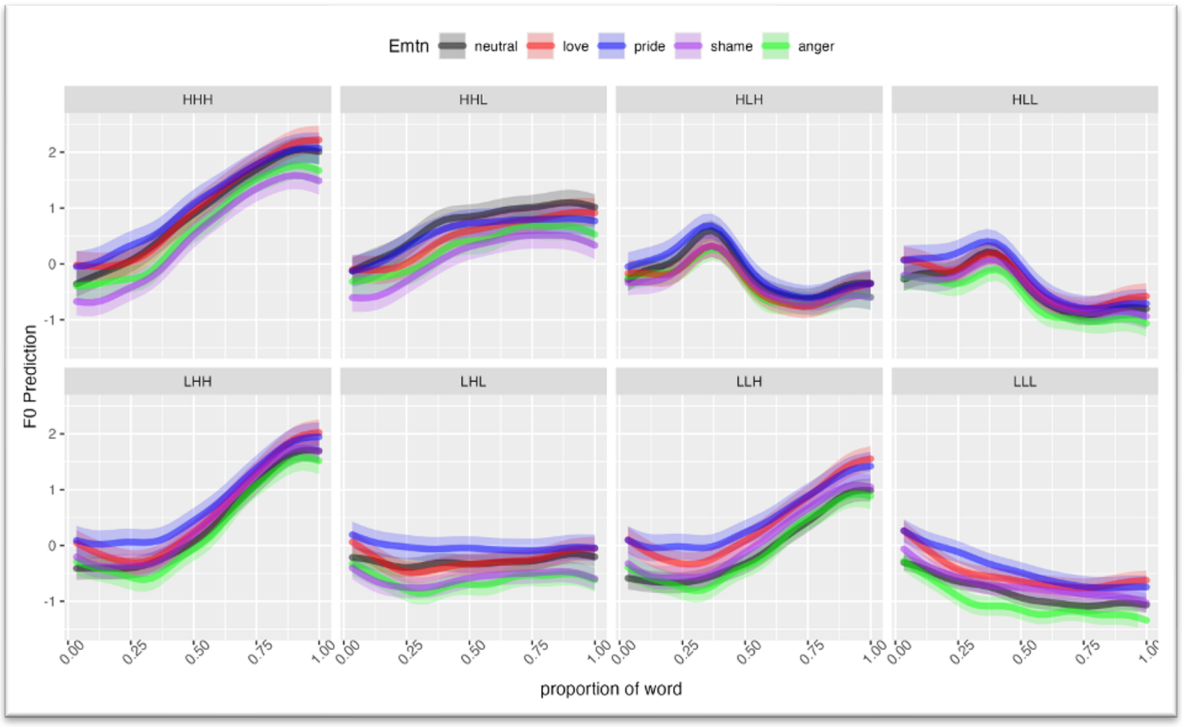


Figure 2: GAMM predictions for tunes by emotion, relative to Neutral. Repeated from Fig. 8A in Ch. 2.

Phase II: Perception

Objectives. Listener expectations can shape their perceptual judgements, leading to potential disparities between measurable acoustic differences and perceived distinctions among intonational tunes. This consideration motivated a dedicated phase of this study for assessing the perceptual distinctiveness between tunes, as produced with emotional portrayals. Based on acoustic evidence collected in the production phase, at least a five-way distinction among intonational tunes appeared to be robust despite variation due to emotion portrayal. This finding rests on state-of-the-art signal processing, clustering analysis and mixed effects modeling, and may differ from the manner in which F0 trajectories are processed and perceived by human listeners in ways we do not currently fully understand, but which this project elaborates upon.

Since this phase of the study has access to the imitated speech materials of the previous phase, it is possible to rigorously test the predictions of the AM model against a background of naturalistic variation conditioned (in part) on the specified emotion for each utterance, as well as idiosyncratic variation from the speaker. Therefore, in the perception experiment, participants were tasked with judging pairs of audio stimuli (naturally produced tune-emotion combinations from Phase I) based on the perceived linguistic function, and (critically) ignoring emotion. Specifically, for each pair[43] of stimuli they answered the question "Is the speaker trying to say the words in the same way, or a different way" using the keyboard. The data was primarily analyzed using Bayesian GLMMs.

Quantitative analysis. The goal of the quantitative analysis was to gauge whether—and how—emotional variation impacts perceived differences between phonologically distinct tunes, as a test of the perceptual salience of tune distinctions predicted by the AM model. The results showed that listeners appear to be sensitive to predicted phonological distinctions despite their co-occurrence with vocal cues of emotion in the signal. Most tune distinctions were at least as salient in the 'Neutral' condition (no specified emotional portrayal) as they were when phonetically encoded in combination with cues conveying a particular emotional portrayal. This comes out in the statistical model comparing tune classification accuracy by emotion condition—see Figure 3.

[43] More specifically, every possible pair of tune-emotion combinations in every order was tested, which required an across-participant design due to the large number of necessary comparisons. The order of presentation of a given stimulus in the AX pair of a given trial was dropped from the models as it did not improve fit.

Comparing perception and production. Although not all the perceptual distinctiveness results are explained by measurable F0 differences between stimuli, there were significant correlations between perception and production, based on simple linear regression. This is true based on comparisons with (i) the root mean squared distance (RMSD) between the actual stimuli being judged in this study, see Figure 4, and (ii) the aggregated difference between tunes according to the difference GAMMs from the production study that was the source of the perceptual study stimuli.

The aim of the first comparison was to understand what share of the results is simply driven by acoustic differences; after all, if the correspondence is perfect there would be no need for future perceptual discrimination experiments. For the second comparison, the connection between the perceptual discrimination stimuli and the acoustic differences under consideration are one degree removed, which allows us to compare *trends* in perceptual distinctiveness to *trends* in produced (F0) distinctiveness. Both comparisons showed strong significant correlations between perception and F0-based differences, yet there was variation that emerged in the perceptual results which suggests there is room for other factors (including speaker emotion) to play a role. Note that in Fig. 4, while there are many cases where acoustically similar tune pairs are, perhaps counterintuitively, perceived as highly distinct (upper-left quadrant), practically no cases show the opposite (dissimilar tunes not perceived as distinct; lower-right quadrant). Moreover, the tune pairs that are discriminated accurately despite their acoustic similarities tend to be from conditions with specified emotion portrayals, particularly Shame and Pride.

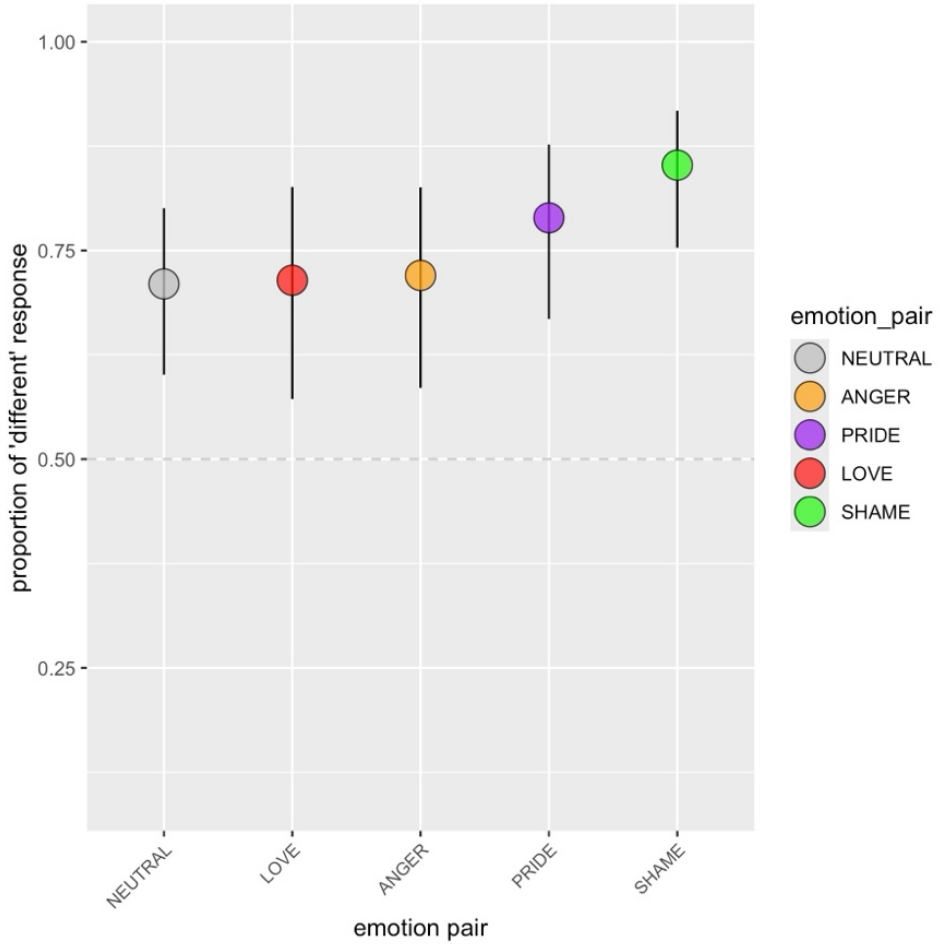


Figure 3: Estimated proportion of (correct) "different" responses in the EMM model (includes Emotion and Tune as factors), by emotion. Repeated from Fig. 13 in Ch. 3.

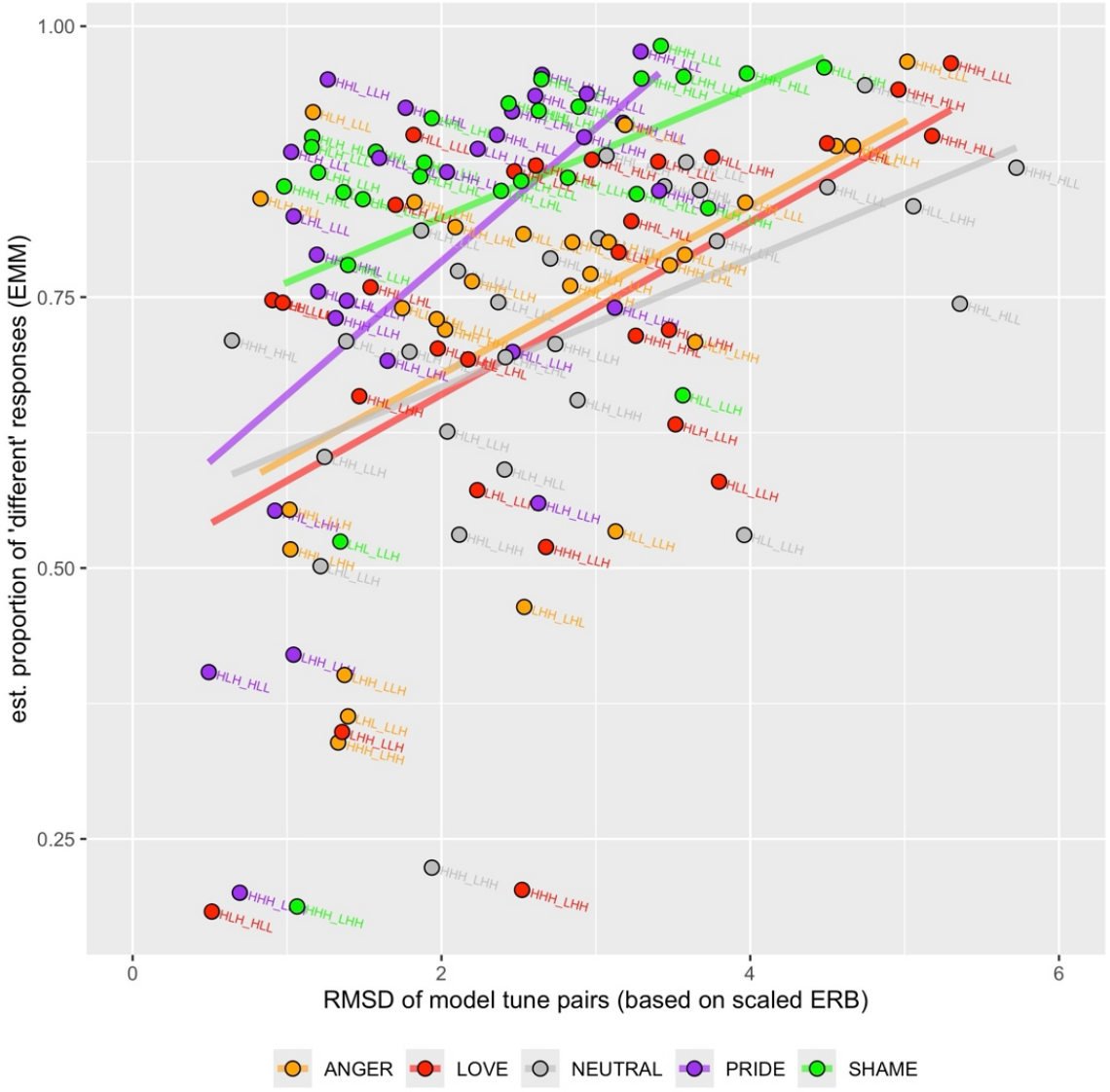


Figure 4: Estimated proportion of "different" (correct) responses for tune pairs in the EMM data (y-axis) compared to the RMSD distance between tune pairs (x-axis) by emotion (color). Repeated from Fig. 19 in Ch. 3.

A surprising takeaway from these results is that the predicted tune distinctions from the AM model are often more perceptually salient when produced with emotional portrayal than without (Neutral), although this is not universal across tunes. In Fig. 4, Neutral is amongst the lowest accuracy of the linear models (purple trend line), and is tied with Shame for the shallowest slope, indicating smaller improvements in perceiving pairwise tune distinctions as the acoustic

difference between a given tune increases, compared to results from other emotion conditions. One possible explanation is that participants are hesitant to interpret the differences between tunes produced in the Neutral condition, uncertain about whether salient acoustic differences relate to linguistic or emotional factors. If listeners are parsing emotion-based variation in the speech signal in parallel to parsing variation due to linguistic differences, it would make sense for intonation judgements to be more confident (if not more accurate) when emotional portrayal is present and straightforward to account for. In sum, the perception study provided new evidence of the AM model's predictive power, specifically that listeners easily disentangle the basic set of intonational forms despite variation in their phonological implementations due to emotional portrayal.

Phase III: Interpretation.

Objectives. Whereas the prior phase of this study focused on the perceptual basis for tune distinctiveness in the speech signal, here the question is how listeners interpret different tunes, specifically the stability of those mappings despite emotional variation. This means that the experiments were designed to find out whether tunes were perceived to convey different interpretations depending on speaker emotion, as tune-emotion interactions revealed in the production and perception findings from the prior two phases suggest is possible. Rather than setting out to precisely characterize a distinct perceived meaning of each of the eight tunes, insights from prior work on intonational meaning in MAE were used to design complementary experiments that take the literature-attested meaning of these tunes into account (see Chapter 1 for that review).

Sorting experiment findings. In the Sorting experiment, participants grouped tune-emotion combinations based on perceived linguistic function, ignoring emotion, without being provided a specific set of tune meanings. The experimental data was through k-means clustering and Bayesian GLMMs. Clustering results suggested that phonological tune differences mainly drove participants' grouping behavior, such that no impact of speaker emotion was found in the coarse-grained ('simple' feature set)—see Figure 5. The two clusters that emerge from the tune-based clustering solution are generally well explained by the phonological specification of the tune (final H versus L tone) which correlates more or less with acoustic differences (F0 trajectory ends with rise versus fall). Additionally, based on the GLMM, there was little evidence that tune-emotion interactions predicted grouping behavior, although including an emotion term slightly improved model fit. Together, the Sorting experiment findings support the view that intonation and tune are independently (or interdependently, considering phonetics) signaling meaning in their respective domains, rather than interacting and co-determining the perceived meaning.

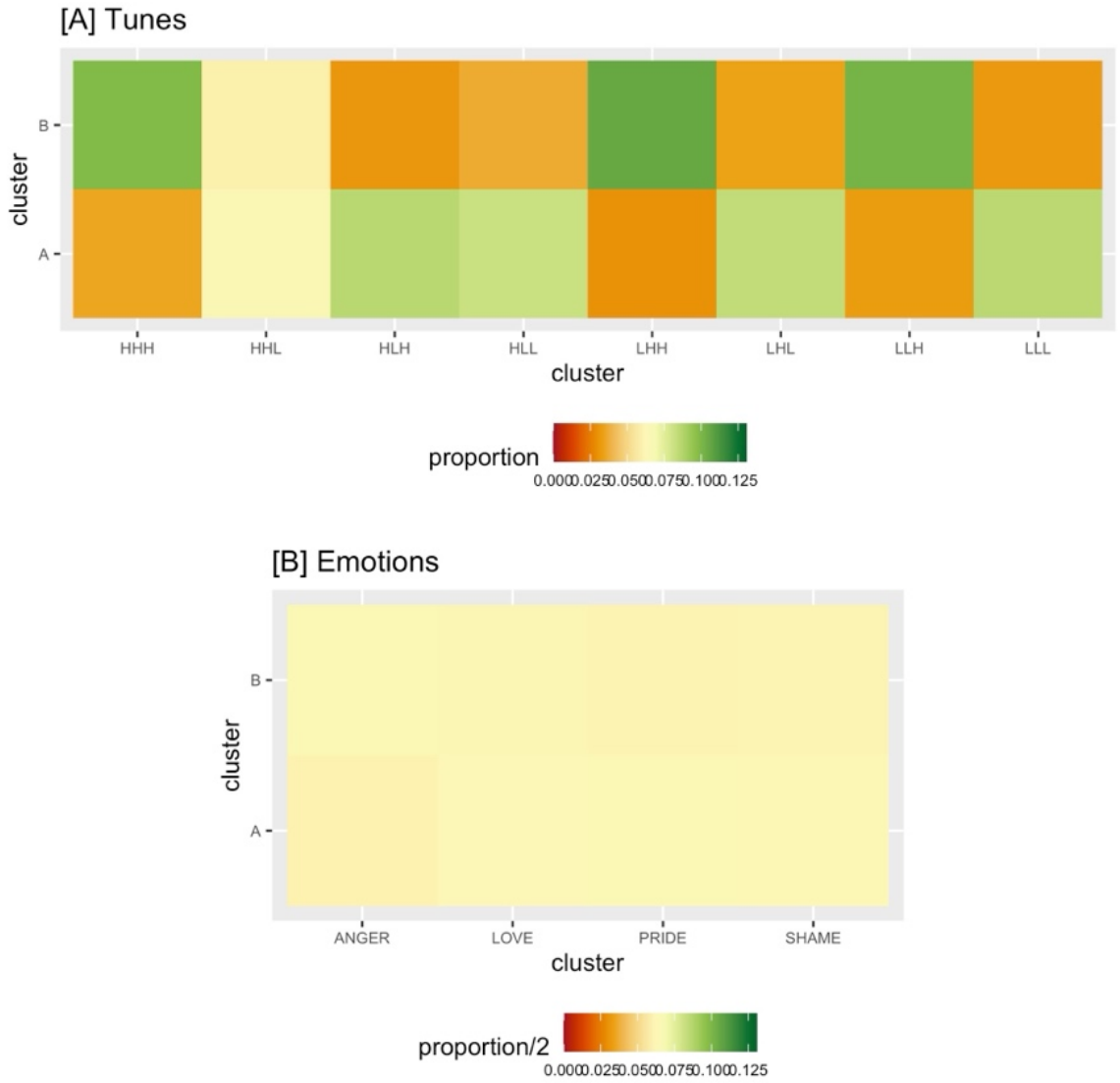


Figure 5: Heatmap of simple clustering solution for tune pairs grouped together in the Sorting experiment, with cluster composition by tune (Panel A; top) and by emotions (Panel B; bottom), with color coding showing the proportion of that feature (tune or emotion) within the cluster. Proportion values for emotions are halved because there are twice as many emotion observations than tune observations due to the experiment design, which makes the color scales comparable. Color is scaled between 0 as the minimum proportion of tokens in a given cluster and just over 1.2t as the maximum proportion. Repeated from Fig. 7 in Ch. 4.

Rating experiment findings. In this experiment, participants were asked to rate a token with a particular tune-emotion combination for a specific meaning on a given trial on a 5-point agreement scale. They did so based on how well the meaning is conveyed by the speaker's

rendition of the tune-emotion combination in that stimulus while attempting to ignore any and all vocal cues of emotion from the speaker. This finding is visible in the empirical data where different response patterns arise for positive and negative prompts (poles of the same meaning) in each meaning category (e.g., “commit”), but highly similar response patterns for tunes produced in Anger (solid lines) and Love (dashed lines). Figure 6 and Figure 7 below show an example of how to interpret the empirical distribution of ratings between prompts by tune-meaning combinations, and the results, respectively. In line with Phase II and the Sorting experiment, participants here were highly successful in perceptually compensating for emotional variation, as evidenced by ratings of meaning associations being highly consistent across the two emotion conditions compared, for every tune.

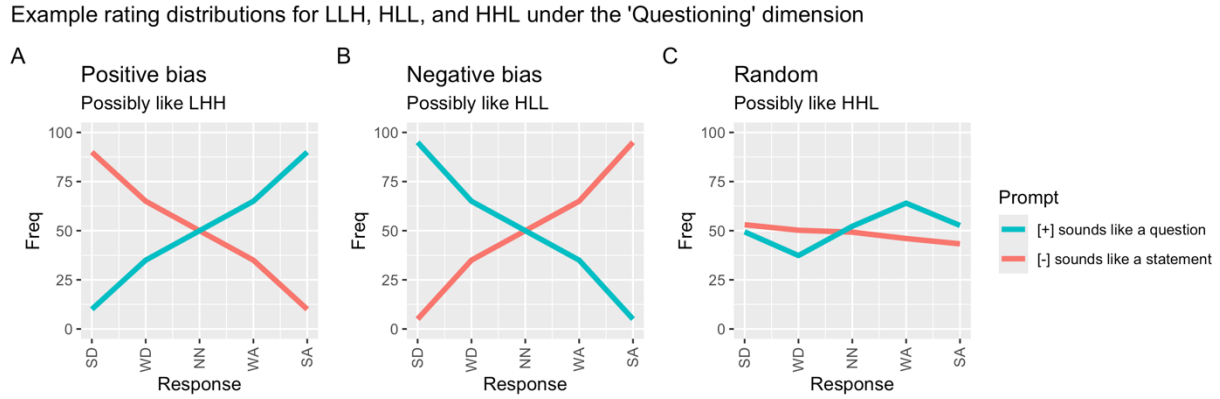


Figure 6: The three main ways that the agreement ratings might be distributed for different tunes, under the questioning meaning dimension. A shows a positive bias, which the literature suggests might arise for LHH for this meaning dimension, a negative bias which might arise for HLL, and a lack of a bias (random) which might arise for HLL. Repeated from Fig. 12 in Ch. 4.

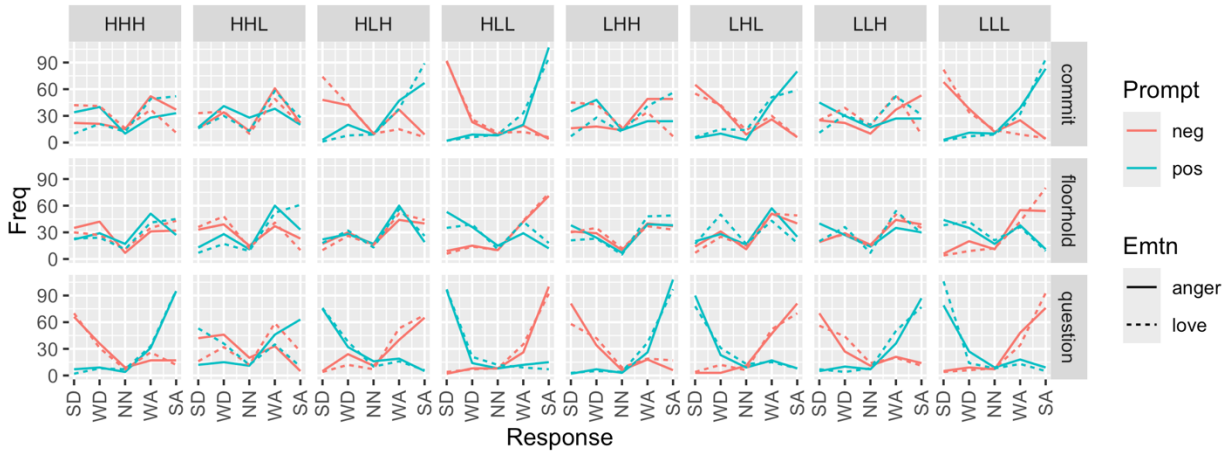


Figure 7: Empirical frequency of Likert scale responses by tune (column), meaning (row), prompt (color) and emotion (line type). Repeated from Fig. 16 in Ch. 4.

Although the above empirical results are highly interpretable, in order to gain a deeper insight into the relative availability of various tune-meaning combinations, responses were submitted to ordinal mixed modeling, specifically CLMMs. First, based on a cross validation of CLMMs, it was determined that the model of best fit did not have an emotion term, which accords with the visual analysis of the empirical results. In other words, there was no systematic impact of speaker emotion on how participants evaluated tune-meaning associations, when all factors were considered. While this finding resolved the primary research objective for this phase, these data

provide a rich view into the set of intonational meaning associations found in this study, shown in Figure 8. Most tunes are shown to be associated with multiple meanings, in fact there is one exception, LHH which only appeared to map to 'question'. Besides LHH, three others tunes appeared to be mutually distinctive in meaning, HHH (questioning > commitment), HHL (floorholding > questioning), and LLH (questioning > floorholding). For the other half of the tune inventory (HLH, HLL, LHL, LLL) the ratings point to a roughly shared meaning involving all tested dimensions (committing > floorholding > questioning).

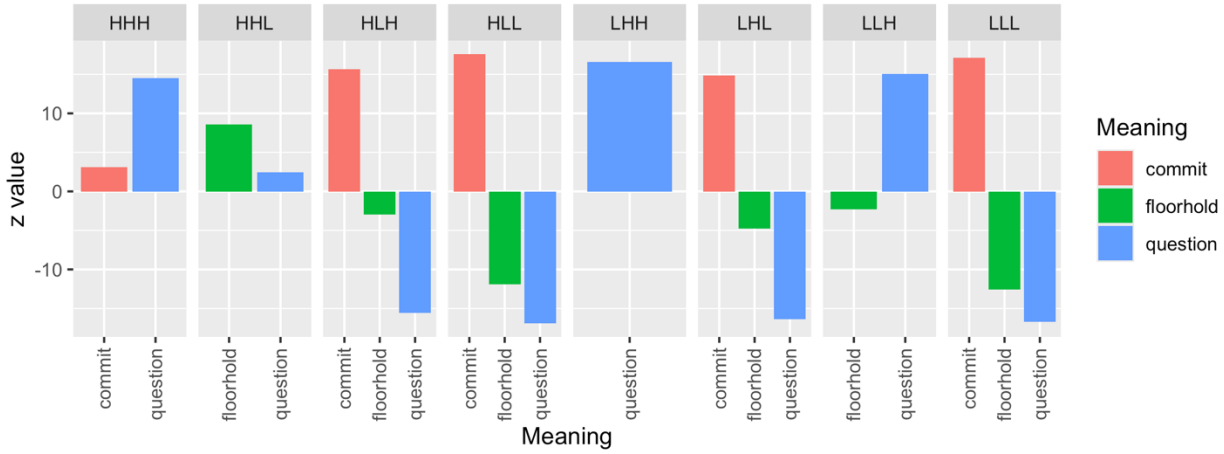


Figure 8: CLMMs of tune-meaning combinations (z values). Displaying significant coefficients only. Repeated from Fig. 16 in Ch. 4.

The five distinct interpretation clines emerged from the model, according to the direction and magnitude of the tune-meaning association shown in Figure 8, are:

1. questioning only (LHH)
2. questioning > floorholding (LLH)
3. questioning > commitment (HHH)
4. floorholding > questioning (HHL)
5. committing > floorholding > questioning (HLH, HLL, LHL, LLL)

For the four tunes that showed distinctive meaning associations (LHH, LLH, HHH, HHL), the correlation with the intonational meaning literature is strong, granting some level of validation for the novel technique of using counterbalanced prompts and multiple meaning dimensions. That said, the fact that half of the tune inventory apparently overlap along these supposedly critical meaning dimensions is a challenge for the AM model. This is because phonological compositionality does not explain the apparent overlap in perceived meaning for half of the tune inventory (HLH, HLL, LHL, LLL)[44]. These four tunes were mainly different in how strong they conveyed floor *ceding* (negative floorholding), but it is unclear whether this is a sufficient distinguishing feature. It is possible that meaning distinctions among these tunes is more convincing along untested dimensions of interest beyond the three from this study, but based on the current evidence there appears to be striking overlap in the meaning domain of tunes.

Summary

Overall, different emotions appear to systematically shift the phonetic implementation of tunes (in their F0) in predictable ways that were shown to be easily accounted for in production (based on modeling F0), in perception (based on perceptual discrimination), and interpretation (based on free classification and rating studies). This is taken as an encouraging sign for intonational phonology, since the theoretical predictions mostly bear out, but nonetheless persistent gaps in the expected evidence for tune distinctiveness remain.

[44] From these four tunes, six unique (non-identical) pairs can be created: HLH-HLL, HLH-LHL, HLH-LLL, HLL-LHL, HLL-LLL, LHL-LLL. This set of tunes is discussed in relation to perceptual distinctiveness in Ch. 4, which is reviewed in the 'Synthesis' section below.

5B. Synthesis

The purpose of this section is to take advantage of the parallel design implemented across all three phases of research, in order to draw additional insights about intonation that are available given the breadth of evidence. This dissertation sought to leverage a wide variety of experimental paradigms, in order to build a broader understanding of the object of research, intonational tunes. While differences between the types of data being collected and analyzed across the studies makes some quantitative comparisons challenging, commonalities in the materials and modeling opens the door for many cross-experiment analyses.

The comparison of interest for this section is a synthesis of all three phases of research, including how intonation is phonetically impacted by emotion (production), whether listeners can disentangle intonational tunes from emotionally variable speech (perception), and whether listeners recover distinct intonational meaning despite such variation. If it is assumed that for intonation to convey linguistic meaning that information must be phonetically encoded by the speaker, then those cues must be perceptually salient to the listener against the background of irrelevant variation (e.g. linguistically uninformative, as we imagine emotion to be). In a nutshell, the listener must comprehend the speaker's intended meaning by decoding the relevant perceptual cues. In this comparison of tune pairs, the acoustic distinctiveness of their phonetic encoding is represented by the Difference GAMMs fit to F0 trajectories in Chapter 2 on the x-axis, while their relative perceptual salience is represented by the Emotion Match Model (EMM) of Chapter 3, and relative differences in perceived meaning is represented by the Tune Only

Model (TOM) of Chapter 4's Sorting experiment, on the y-axes[45]. Simple linear regression models were fit comparing the Production results to the Sorting and Perceptual experiment results, the details of which are given in Table 3.

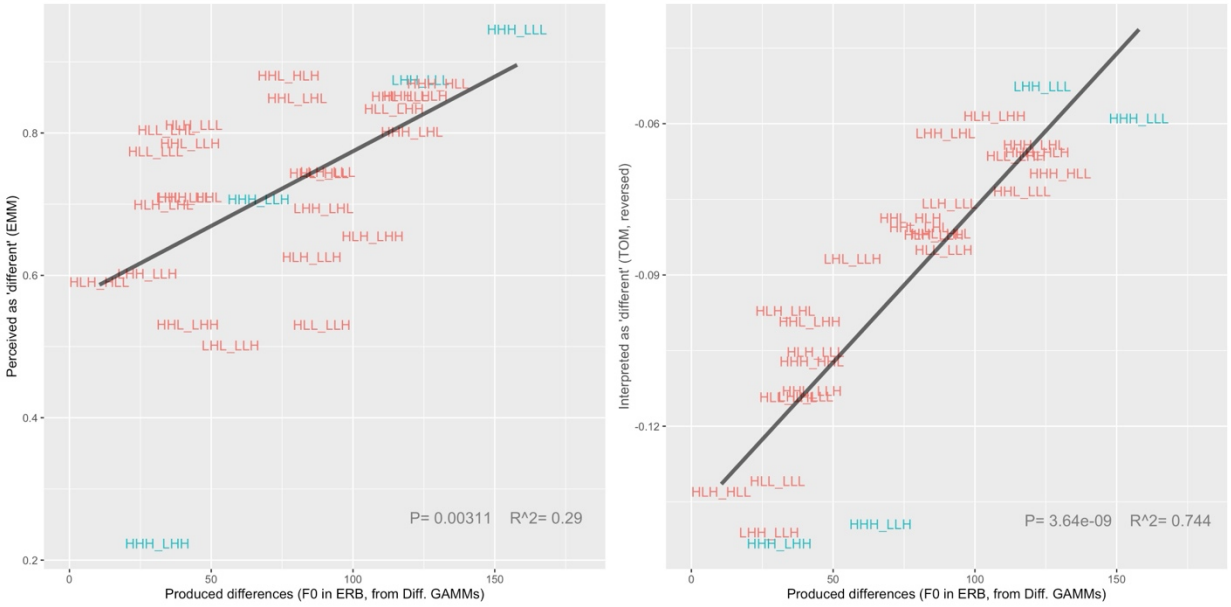


Figure 9: Scatterplots comparing production results to those from perception (left) and interpretation (right), based on the statistical model results, with independently scaled y-axes (slopes not comparable). Blue tune labels indicate pairs discussed in text; default is red. Linear model details are given in Table 3 and repeated in the corner of the plots.

Simple linear regression models were fit comparing the Production results to the Sorting and Perceptual experiment results, the details of which are given in Table 5. The linear models show agreement between results from the phases of Production and Perception, as well as Production and Interpretation, which were both significant and positive in slope. This indicates

[45] Note that the axes for the Sorting experiment data are reversed relative to the Perception experiment because the former judged on the basis of perceived (meaning) similarity and the former on perceived (acoustic) distinctiveness. Additionally, the range of values for Sorting is narrower than for the Perception experiment, which is due to task differences, since in Perception the probability of tunes matching or not was balanced between trials. On the other hand, in the Sorting experiment there are no trials, no balancing, and consequently more ways for tunes to mismatch than match, ultimately leading to a compressed possibility space.

that as F0 differences increase in the acoustic signal, participants are more likely to perceive said differences and interpret them as linguistically meaningful. The correlation between Production and Perception is weaker based on R^2 value, which could mean that on one hand, listeners are highly sensitive to tune distinctions in the perceptual discrimination paradigm. On the other hand, the perceptual discrimination results are less correlated with produced distinctiveness than interpretation, suggesting that the perceptual salience of tune distinctions may be partly task dependent. When listeners were asked to attend to the linguistic function of the tunes, as in the Sorting experiment, listeners' ability to judge tune distinctions tracked very closely with trends in tune implementation from the production data.

Table 3: Details for linear regression models shown in Figure 9

2-Way Comparison	R^2	p	p < .05?
Production × Perception	0.290	0.003	Yes
Production × Interpretation	0.744	<0.001	Yes

Given that the Sorting experiment results are highly correlated with those from the Production experiment, there were few opportunities for common outliers when also considering the Perceptual experiment results, which showed more variation. For instance, the left plot in Figure 9 shows HHH_LLH (blue) to be well-predicted by data from Production x Perception (middle of regression line), but it is a clear outlier based on Production × Interpretation data (bottom left quadrant). This means that while HHH_LLH's perceptual discriminability is commensurate with acoustic distance, perceived differences in interpretation were less than one would expect. Another tune pair, HHH_LHH, shows roughly the opposite relationship. Based on Production × Perception, HHH_LHH was perceptually discriminated less reliably than other tune pairs, yet in Production × Interpretation it was in the range of other (relatively indistinct) tune pairs—

arguably less of an outlier than expected. One straightforward explanation for these minor differences between experimental results is the linguistic knowledge of the participants, all of whom use MAE as their primary language. It is assumed that participants are guided in practically every linguistic task by their expectations and with a deeper understanding of intonation variation—which this research contributes to—broader evidence of language (or dialect) specific variation is likely to be found across the data collected in these studies.

5C. Limitations

The insights and contributions of this research project should be considered with its known limitations, constraints, and blind spots. The purpose of this section is to consider what ramifications may be associated with the research limitations that are known, such as theoretical and methodological choices, and generalizability of empirical and analytical observations to other data.

Production (Phase I/Chapter 2)

Since the recordings collected in the production phase were used as stimuli for all following studies, limitations are inherited by subsequent experiments, and therefore are considered general to the project.

Linguistic/phonological model. The sole unifying design features across all experiments is the set of eight phonologically distinctive tunes generated by the AM model, given two tones (High/Low) and three featural components to the intonational tune (see Chapter 1). This leaves a considerable untested intonational possibility space, which includes shorter and longer tonal sequences, as well as variants like bitonal pitch accents and downstepped High tones. The present study is not the first to exclude such cases from consideration; it follows in the steps of Cole *et al.* (2023). It is left to future work to explore the relationship between production, perception and interpretation for these other intonational forms.

Participants. There is yet limited research on dialectal variation in MAE intonation, though prior studies indicate differences in the inventory of pitch accents (Arvaniti & Garding, 2007) and the

frequency of use among pitch accents (Burdin *et al.,* 2018). A priori, one has no reason to think that dialectal variation would systematically impact the tunes investigated in this study, where only native English speakers from the Inland/Northern Cities dialect were tested, but questions of dialectal variation in intonation needs further exploration. The prospect for expanding the current research project to investigate these issues is discussed further in the following section.

Acoustic modeling. In the acoustic domain, the data consisted of time-normalized F0 measurements versus other possibly relevant dimensions, like amplitude/energy, voice quality, duration, spectral cues, etc. While F0 is a highly appropriate choice to model tunes, important questions about how intonation and emotion interact in the speech signal in other acoustic dimensions remain unaddressed in the present analysis. That said, by making the raw data and F0 analysis files from this study freely available, the groundwork exists for a straightforward extension to consider tune-emotion effects across the speech signal. In the future, especially as mixed multinomial time series models become available, simultaneous consideration of F0 with other dimensions will shed light on unaddressed components of the speech signal.

Psychometric/emotion model. Due to the many tunes under consideration, the design of the emotion inventory required simplification, such that each of the four emotion words represented an extreme on both dimensions. This is the minimum to represent two dimensions of emotion, but it makes disentangling the effect of particular lexical items from their associated emotions impossible. That means that, even though the selection of emotion words was done based on relative position in Valence x Potency space, not denotation or connotation of the lexical item itself, the particular word could have played a role.

Perception (Phase II/Chapter 3)

Materials. The perceptual discrimination experiment was designed around a set of auditory stimuli consisting of one token (a recording from Chapter 2) per tune-emotion combination from one speaker. The justification behind using one model speaker was to constrain the number of unique trials in order to test all orders of tune-emotion combinations—modeling another level for speaker would have required considerably more data collection, for unknown gain. That said, because only one speaker was used, even though the recordings were chosen based on F0 features, the salience between stimuli may partly reflect idiosyncratic elements of the speaker's production that cannot be phonologically explained. The present study took the strategy of trusting the speaker to preserve critical intonational cues, since there was no way of telling how tunes would be produced when combined with emotion. This was confirmed by measurements of F0 trajectories across tunes and emotions for the audio materials selected as stimuli for Experiment 2.

Participants. The choice to recruit perception participants who had a language history like the speakers (native English speaker of the target dialect) was driven by the fact that the participants in the production experiment (Experiment 1) were also engaged in a perceptually guided task, i.e., imitation. This way, participants in both production and perception are expected to share similar linguistic expectations that undoubtedly shape the experiment results. It would therefore be instructive (as in production) to examine tune perception for individuals who were not eligible for the current research project because of their different language background—this is discussed further in the following section. The comparability of participants in both studies in terms of perceptual expectations would be strengthened if the production participants also performed a tune discrimination task.

Analysis. A shortcoming of the analysis is that it subdivides the experimental data into multiple subsets based on the matching status of the tune and emotion, ultimately including 90% of what was collected. The 10% that was not modeled included trials where the two stimuli being discriminated matched both in tune and emotion, where performance was near ceiling. It is possible that a different analysis could be implemented that adheres to different statistical assumptions to ultimately unify these different models. If the statistical analysis could be met by a single model, the main advantage would be better estimation of participant-level random effects. While this modification could improve model fit by reducing variation, it would not be expected to change the direction of results, especially considering Phase III's findings using two model speakers, discussed next.

Auditory Free Classification, a.k.a. "Sorting" (Phase III/Chapter 4)

Materials. The stimulus diversity was expanded upon what was used in Ch. 3 to include a second model speaker, a male with a lower mean F0. Given that using a sole model speaker was noted as a limitation of Phase II, the use of two model speakers should be seen as an improvement. This is because participants experience listening to the same tune distinctions in a different parts of the F0 space, which may improve their ability to retrieve useful linguistic knowledge to help with the task.[46] That said, stimuli from additional speakers, or different speakers, could have been used, but this was not tested. Yet another approach that was considered but not executed was to average the F0 trajectories for tune-emotion combinations from Experiment. 1, or their predicted

[46] Model speaker was tested as a factor in the statistical model but a significant effect failed to emerge from this analysis, suggesting participants were successfully ignoring irrelevant information, as they did with emotion, according to the instructions.

F0 trajectories from the Experiment. 1 GAMM analysis, due to the possibility of informative secondary cues. Given the close correlation between F0 trajectories and linguistic interpretation found in this project, there is scant evidence for a role for secondary acoustic cues, so aggregating over F0 trajectories might be a useful strategy for future work. *Task.* Developing the novel software implementation of the AFC task involved solving technical challenges, such as how participants could interact with and visually manage multiple groups. Participants began with three empty groups and were allowed to add more as desired, up to a limit of 21 groups, that was reached after much consideration. The main justification for this cap is that it is sufficient for testing the specific hypotheses laid out for this experiment; far above tunes in the inventory (8) and large enough to accommodate a distinct group for each tune-Valence combination (16). However, based on the distribution of group counts, and the number of distinct intonational meanings that are expected, given the literature review (Chapter 1), it is likely that for larger solutions that not all groups are equally informative. This may reflect differences in task strategy, depending on whether participants tried to sort new tokens into existing groups, or created a new group every time a difference in perceived meaning was detected. If the latter, participants may inadvertently create more groups than they can possibly track, which affects the informativity of such solutions, hence the exclusion.

Group labels. Another factor impacting the interpretability of the groups in the AFC task is the fact that the present study only considered the data in terms of the emergent properties of the solutions that participants provided (based on tune, emotion, speaker), without also directly asking participants about meaning. In retrospect, it could have been highly insightful—and simple—to also collect participant-generated group labels, although (in the opinion of the author) it is too challenging to ask untrained participants to briefly characterizing intonational meaning

in prose. An alternative that was not tested was to ask the same participants to complete both the Sorting task and the Rating task, ideally within the same session, ideally in that order (lest the meaning dimensions influence classification).

Five-point Likert scale experiment, a.k.a. "Rating" (Phase III/Chapter 4)

Materials. Due to time limitations, the materials were limited to one speaker (identical to the Perception experiment) and just two (of five possible) emotion portrayals. The selection of the critical emotions, Love and Anger, was mainly justified on the basis of observed differences by Valence, which suggested that simplifying the design of the emotion manipulation would not affect the overall results. This was not tested, however, leaving open the possibility that Neutral, Pride, and Shame may show different outcomes with the tunes for this task, although it should be noted that in the AFC task, which included Love, Anger, Pride, and Shame stimuli, there was no emergent influence of Potency. If the Rating experiment had been conducted independently, more participant time could be allotted to it, allowing for a greater variety of emotions to be included. That said, participants were successful in ignoring emotion in this experiment, to the extent the model of best fit did not have a term for emotion, so it is likely that testing different emotions is tantamount to additional repetitions.

Modeling. Recall that the primary quantitative analysis for this experiment was ordinal modeling, but specifically a separate model for each tune-meaning combination (8 tunes × 3 dimensions = 24 total CLMMs) for the same analysis. This was justified because, from a phonological perspective, tunes are capable of conveying multiple distinct linguistic functions which may be interdependent or even hierarchical, which motivates separate models due to violating statistical

assumption of independence. Post-hoc, it is possible to reanalyze the data by combining rated-alike tunes through a unified ordinal model, wherein tune is a multilevel factor, which may improve estimation of random effects and overall effect sizes. While the use of a Likert response allowed a fine-grained analysis, it necessitated leveraging ordinal statistical models, which were difficult to reconcile with the types of data collected elsewhere in this research project. An alternative version of the experiment might use a procedure akin to the AX discrimination paradigm in the perception phase, with pairwise presentation of tunes-emotion stimuli and a two-alternative forced choice (2AFC) response to whether they both convey a particular prompt (the same prompts could be used). Such data could be analyzed using binomial regression, like the Perception and Sorting experiments, which may lead to a deeper understanding, but given that such a methodology does not readily exist and is unvetted, considerable resources would have to be invested.

Machine learning. Yet another alternative route of analysis is via machine learning (ML), since the CLMMs attempt to model a simple behavior which (as argued in this chapter) is largely predictable based on F0 trajectory differences. With ML, part of the experimental results can be used to train an ML classifier that assigns a continuous acceptability rating to tune-meaning combinations, then the other part of the results can be used to test how accurately ML mimics human judgements. If the ML's representation of different tunes converges in training, it suggests that the tunes may have the equivalent perceived meaning for listeners. While ML was not leveraged in this project for a variety of reasons, due to the flexible way it has been shown to accommodate highly variable signals, it may prove increasingly useful in the future

5E. Implications

Before this project, the question of whether and how emotional intonation would interfere with the expression of linguistically determined intonation was unaddressed. It was generally assumed, but not confirmed, that phonological contrasts are uniformly informative across diverse emotional contexts that involve other, sometimes conflicting, acoustic cues. This project found widespread evidence for the stability of linguistically determined intonational cues in the speech signal, and of their form-meaning mapping. The details of the findings, along with the diverse methodology used and developed to obtain this newly transparent view of intonational phonology, has major implications for phonological theory, research methodology, and development of speech technology.

These results support the AM model, though further consideration of that model is also justified. Research in this area has long been hampered by challenges in separating the linguistically informative features of the speech signal from other factors, but the method employed in this study demonstrates away forward. Specifically, this project highlights the strength of interdisciplinary approaches to linguistics research, which may eventually lead to more cognitively informed phonological models. Acoustic variation is complex but structured, according to our findings, which opens the door to analyzing more naturalistic speech data in the future. This research project lays the groundwork for a more complete understanding of intonation from its phonetic implementation to the perceived linguistic functions its basic forms are able to convey. Given the noted variation in the phonetic expression of intonation from prior work, a phonology-centered approach was adopted here, but given the tight link between interpretation, production, and perception, there appears to be an opportunity to attempt a meaning-centered approach. For example, according to the Rating experiment, every tune

conveyed some meaning on the ‘questioning’ dimension (question/statement distinction), which was not true for all meaning dimensions. In other words, while some dimensions of semantic-pragmatic meaning were not associated with certain tunes, the ‘questioning’ dimension appears to be available across tunes, universally. Furthermore, in the Rating experiment, participants compensated for emotional variation to such a degree it did not improve the statistical modeling of rating responses.

As intonation researchers learn to decode F0 trajectories to identify their source in phonological specifications, despite variation due to non-linguistic factors like speaker emotion, the ability to compute the speaker’s intended linguistic meaning should improve. Given the current underutilization of intonational cues in speech technology, together with the implications of our findings, intonation seems to be a likely next frontier for natural language understanding.

.

Chapter 6: Conclusion

6A. Objectives

Acoustic variation surrounding intonational forms has been a major contributor to the paucity of evidence for robust phonological distinctions, as predicted e.g., by the Autosegmental-Metrical (AM) model (Pierrehumbert 1980, Ladd 2008, among others), even for such well-documented languages as American English. This problem obstructs a deeper understanding of intonation from the listener's perspective as well, for perception of intonational contrasts and their interpretation, intonational meaning. The unique approach tested here was to constrain acoustic variation that may blur tune distinctions by jointly considering sources of variation in phonological specification and emotional portrayal. Importantly, emotion was also formalized through adoption of a recognized analytical framework from psychology (Fontaine *et al.* 2007). The scope of this project included four total experiments, one for production (imitating tunes while portraying emotions; Phase I), one for perception (distinguishing tunes produced in combination with emotional portrayal; Phase II), and two for interpretation (judging the meanings of tunes produced with emotion; Phase III). The research objective of considering multiple methodological perspectives was to thoroughly test the hypothesis that intonational distinctions will be enhanced by eliciting, then controlling for speaker emotion. With a clearer picture of intonational form as it relates to distinctions predicted from the phonological model, a further aim was to examine the association between intonational form and pragmatic meaning. A secondary objective was to use emotional variation to better understand how listeners perceive and interpret the linguistic meaning encoded by intonation in conditions of emotional variation.

6B. Outcomes

The research findings contribute to a deeper understanding of how the phonological contrasts of intonation are encoded and decoded by jointly controlling for speaker emotion. Effects of emotion on intonation, measured in the F0 trajectories that implement phonologically specified tunes, was observed mainly in production. Listeners seemed to easily account for emotion conditioned variation, both in terms of perceptual distinctiveness and the range of meaning associations listeners endorsed. The phonological specification of intonational tunes was found to be the primary driver of acoustic variation in F0 trajectories, and a strong predictor of whether listeners perceived tunes to convey contrastive meanings. This bolsters a common assumption within linguistics the expression of linguistic content can be independent of speaker's emotion, to a large extent. Speakers tend to preserve contrasts predicted to be critical for intonational distinctiveness according to the AM model, within specific emotional contexts. This suggests that, despite emotion-conditioned variation in the phonetic realization of intonational contrasts, the linguistic system remains robust. Evidence from the perception and interpretation of tunes builds on these findings to further demonstrate that listeners can cope well with emotional variation in their linguistic evaluations of tunes. That said, fully half of phonologically distinct tunes were often treated as functionally interchangeable, based on the findings from the interpretation phase of this study. Comparisons across all phases of research (perception, production, interpretation) show that perceptual discriminability and distinctiveness of tune-meaning associations tend to correlate with F0 trajectory similarity.

6C. Future directions

Interdisciplinary research

This project was conceptualized, in part, to resolve longstanding barriers to empirical work that might deepen our scientific understanding of intonation, particularly by its interdisciplinary design. The main benefit of exploiting the connection between intonation and emotion was it allowed for a fine-grained characterization of tunes, made possible due to their entanglement in the speech signal, and therefore by extension for the perception and interpretation of tunes.

Dialectal variation

The influence of speaker dialect on the phonetic implementation of intonation would be topic worthy of future research. Given that dialect was controlled in the present study, the results presented here serve as a possible baseline for future work testing dialectal variation in American English intonation. For example, the special AFC task would be a logical starting point to test whether listeners can classify speakers' dialect based on their intonation. This is also an invitation to emotion researchers to adopt these methods and results to explore questions beyond linguistics, particularly to problems that would benefit from phonologically specified intonation.

The AM model and intonation theory

This project's findings generally support the AM phonological model of MAE intonation, but also point to the need for continued development and refinement of that model as it makes

predictions about meaning distinctions. Phonologically distinct tunes that are perceived to convey the same linguistic function across emotions seems to be a deep challenge that will require empirically guided reconsideration of the current AM model. Another weakness in our current understanding of intonation, as highlighted by this project's findings, is the possibility space of intonational meaning, and the structure of the intonational lexicon. Currently it seems that phonologically distinctive tunes may be interpreted to mean the same thing, despite perceptually salient differences. This research found a mix of evidence for one-to-many and many-to-many associations between tunes and pragmatic meaning, which suggests this is a ripe topic for future exploration. It is of particular interest for intonational theory to establish whether tune-meaning mappings are flat, as assumed here, or hierarchical, which might elegantly explain the emergent patterns of tune meaning.

Emotion research

Beyond linguistics, this project's results and findings may be useful to better understand how emotion expression is accomplished vis-à-vis speech cues, independent of linguistic intonation. This would seem to benefit important ongoing work where paralinguistic factors are the research focus, such as how the phonetic encoding of emotion in speech conveys speaker culture, a topic recently discussed by van Rijn & Larrouy-Maestri (2023). The present findings suggest that similar future research would benefit from an understanding of the linguistic formalisms underlying intonation.

6D. Closing

This research fills a gap between a formal model of MAE intonation, and by extension the AM theory on which it is based, and empirical evidence from a series of four experiments testing perception, production, and interpretation of intonational tunes. The findings broadly support the AM model. By considering how tune-emotion combinations were treated by speakers and listeners, it was concluded that despite acoustic confounds, phonological and emotional factors function separately in the production and perception of intonation. Thanks to this relationship, accounting for emotional variation in the results led to a more accurate picture of how phonology conditions the phonetic implementation of intonation. Listeners were highly sensitive to critical tune distinctions despite emotional variation, to such a degree in some experiments speaker emotion made no statistical difference in tune interpretation despite measurable effects of emotion on the F0 trajectories that implement tune contrasts. On the other hand, listeners' judgements about tune-meaning associations only partly validated claims about the pragmatic function of tunes claimed in prior literature, pointing to an area for additional research. Finally, while the goal of this work was not to study emotional prosody *per se*, the robust takeaways for intonation researchers is the demonstrated predictive power of emotion on F0 variation, where emotion is modeled based on validated theory as well. This success clears the way for future interdisciplinary research that can serve to build and refine our linguistic understanding.

Chapter 7: References

Anikin, A. (2019). Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behavior Research Methods*, *51*(2), 778–792. https://doi.org/10.3758/s13428-018-1095-7

Arvaniti, A. (2020). The Phonetics of Prosody. In A. Arvaniti, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.411

Arvaniti, A., & Garding, G. (2007). Dialectal variation in the rising accents of American English. *Laboratory Phonology*, *9*.

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion*, *12*(5), 1161–1179. https://doi.org/10.1037/a0025827

Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2012). Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology*, *3*(2). https://doi.org/10.1515/lp-2012-0017

Bartels, C. (1999). *The intonation of English statements and questions: A compositional interpretation*. Garland publ.

Bartels, C., & Kingston, J. (1994). Salient pitch cues in the perception of contrastive focus. *The Journal of the Acoustical Society of America*, *95*(5), 2973–2973. https://doi.org/10.1121/1.408967

Boersma, P., & Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.

Bolinger, D. (1986). *Intonation and its parts: Melody in spoken English* (Nachdr.). Stanford Univ. Press.

Bolinger, D. (1989). *Intonation and Its Uses: Melody in Grammar and Discourse*. Stanford University Press.

Braun, B., Kochanski, G., Grabe, E., & Rosner, B. S. (2006). Evidence for attractors in English intonation. *The Journal of the Acoustical Society of America*, *119*(6), 4006–4015. https://doi.org/10.1121/1.2195267

Brooks, M., E., Kristensen, K., Benthem, K., J. ,van, Magnusson, A., Berg, C., W., Nielsen, A., Skaug, H., J., Mächler, M., & Bolker, B., M. (2017). glmmTMB Balances Speed and

Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, *9*(2), 378. https://doi.org/10.32614/RJ-2017-066

Burdin, R. S., Holliday, N., & Reed, P. (2018). Rising Above the Standard: Variation in L+H* contour use across 5 varieties of American English. *Speech Prosody 2018*, 354–358. https://doi.org/10.21437/SpeechProsody.2018-72

Burdin, R. S., & Tyler, J. (2018). Rises inform, and plateaus remind: Exploring the epistemic meanings of "list intonation" in American English. *Journal of Pragmatics*, *136*, 97–114. https://doi.org/10.1016/j.pragma.2018.08.013

Büring, D. (2016). *Intonation and meaning*. Oxford University Press.

Buttrey, S., E., & Whitaker, L., R. (2015). treeClust: An R Package for Tree-Based Clustering Dissimilarities. *The R Journal*, *7*(2), 227. https://doi.org/10.32614/RJ-2015-032

Clopper, C. G. (2008). Auditory free classification: Methods and analysis. *Behavior Research Methods*, *40*(2), 575–581. https://doi.org/10.3758/BRM.40.2.575

Clopper, C. G., & Bradlow, A. R. (2009). Free classification of American English dialects by native and non-native listeners. *Journal of Phonetics*, *37*(4), Article 4. https://doi.org/10.1016/j.wocn.2009.07.004

Clopper, C. G., & Pisoni, D. B. (2007). Free classification of regional dialects of American English. *Journal of Phonetics*, *35*(3), Article 3. https://doi.org/10.1016/j.wocn.2006.06.001

Cochrane, T. (2009). Eight dimensions for the emotions. *Social Science Information*, *48*(3), 379–420. https://doi.org/10.1177/0539018409106198

Cole, J., Steffman, J., Shattuck-Hufnagel, S., & Tilsen, S. (2023). Hierarchical distinctions in the production and perception of nuclear tunes in American English. *Laboratory Phonology*, *14*(1). https://doi.org/10.16995/labphon.9437

Dilley, L. C., & Heffner, C. C. (2021). role of f0 alignment in distinguishing intonation categories: Evidence from American english. *Journal of Speech Sciences*, *3*(1), 3–67. https://doi.org/10.20396/joss.v3i1.15039

D'Imperio, M. (2000). *The Role of Perception in Defining Tonal Targets and their Alignment* [PhD Thesis].

Farkas, D. F., & Bruce, K. B. (2010). On Reacting to Assertions and Polar Questions. *Journal of Semantics*, *27*(1), 81–118. https://doi.org/10.1093/jos/ffp010

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The World of Emotions is not Two-Dimensional. *Psychological Science*, *18*(12), 1050–1057. https://doi.org/10.1111/j.1467-9280.2007.02024.x

Genolini, C., & Falissard, B. (2011). Kml: A package to cluster longitudinal data. *Computer Methods and Programs in Biomedicine*, *104*(3), e112–e121. https://doi.org/10.1016/j.cmpb.2011.05.008

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1–2), 103–138. https://doi.org/10.1016/0378-5955(90)90170-T

Goodhue, D., Harrison, L., Wagner, M., & Yuen Tung Clémentine, S. (2016). Toward a bestiary of English intonational contours. In C. Hammerly & B. Prickett (Eds.), *Proceedings of the 46th Conference of the North Eastern Linguistic Society (NELS)* (pp. 311–320).

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (Vol. 3). New York: Academic Press.

Gussenhoven, C. (2002). *Intonation and interpretation: Phonetics and phonology*. Speech Prosody, Aix-en-Provence, France. https://www.isca-speech.org/archive_open/sp2002/sp02_047.html

Heim, J. M. (2019). *Commitment and engagement: The role of intonation in deriving speech acts*. https://doi.org/10.14288/1.0380772

Imai, S., & Garner, W. R. (1965). Discriminability and preference for attributes in free and constrained classification. *Journal of Experimental Psychology*, *69*(6), Article 6. https://doi.org/10.1037/h0021980

Jeong, S. (2018). Intonation and Sentence Type Conventions: Two Types of Rising Declaratives. *Journal of Semantics*, *35*(2), 305–356. https://doi.org/10.1093/semant/ffy001

Jun, S.-A. (2005). *Prosodic Typology I*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199249633.001.0001

Jun, S.-A. (Ed.). (2014). *Prosodic Typology II: The phonology of intonation and phrasing*. Oxford University Press.

Kawahara, H., Cheveigné, A. D., Banno, H., Takahashi, T., & Irino, T. (2005). Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT. *Interspeech 2005*, 537–540. https://doi.org/10.21437/Interspeech.2005-335

Ladd, D. R. (1990). Intonation: Emotion vs. Grammar. *Language*, *66*(4), 806. https://doi.org/10.2307/414730

Ladd, D. R. (2008). *Intonational phonology* (2nd ed). Cambridge University Press.

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G., & Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *The Journal of the Acoustical Society of America*, *78*(2), 435–444. https://doi.org/10.1121/1.392466

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLOS ONE*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368. https://doi.org/10.1037/h0044417

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Proc. Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

Morrison, M., Tang, B., Tan, C., & Pardo, B. (2022, April). *Reproducible Subjective Evaluation*. ICLR Workshop on ML Evaluation Standards. https://github.com/reseval/reseval

Mozziconacci, S. J. L. (1998). *Speech variability and emotion: Production and perception*.

Niebuhr, O. (2007). The Signalling of German Rising-Falling Intonation Categories – The Interplay of Synchronization, Shape, and Height. *Phonetica*, *64*, 174–193. https://doi.org/10.1159/0000107915

Pierrehumbert, J. (1990). Phonological and phonetic representation. *Journal of Phonetics*, *18*(3), 375–394.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation* [PhD Thesis]. Massachusetts Institute of Technology.

Pierrehumbert, J. B., & Steele, S. A. (1989). Categories of Tonal Alignment in English. *Phonetica*, *46*(4), 181–196. https://doi.org/10.1159/000261842

Pierrehumbert, J., & Hirschberg, J. B. (1990). The meaning of intonational contours in the interpretation of discourse. *Intentions in Communication*, 271–311.

Rodd, J., & Chen, A. (2023). Internal structure of intonational categories: The (dis)appearance of a perceptual magnet effect. *Frontiers in Psychology*, *13*, 911349. https://doi.org/10.3389/fpsyg.2022.911349

Roettger, T. B., Mahrt, T., & Cole, J. (2019). Mapping prosody onto meaning – the case of information structure in American English. *Language, Cognition and Neuroscience*, *34*(7), 841–860. https://doi.org/10.1080/23273798.2019.1587482

Roettger, T. B., & Rimland, K. (2020). Listeners' adaptation to unreliable intonation is speaker-sensitive. *Cognition*, *204*, 104372. https://doi.org/10.1016/j.cognition.2020.104372

Scherer, K. R. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech & Language*, *27*(1), 40–58. https://doi.org/10.1016/j.csl.2011.11.003

Shue, Y.-L., Keating, P., & Vicenik, C. (2009). VOICESAUCE: A program for voice analysis. *The Journal of the Acoustical Society of America*, *126*(4_Supplement), 2221–2221. https://doi.org/10.1121/1.3248865

Shuman, V., Sander, D., & Scherer, K. R. (2013a). Levels of Valence. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00261

Shuman, V., Sander, D., & Scherer, K. R. (2013b). Levels of Valence. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00261

Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics*, *84*, 101017. https://doi.org/10.1016/j.wocn.2020.101017

Steffman, J., & Cole, J. (2022). An automated method for detecting F measurement jumps based on sample-to-sample differences. *JASA Express Letters*, *2*(11), 115201. https://doi.org/10.1121/10.0015045

Steffman, J., & Cole, J. (2024). Metrical enhancement in American English nuclear tunes. *Glossa: A Journal of General Linguistics*, *9*(1). https://doi.org/10.16995/glossa.15297

Steffman, J., Cole, J., & Shattuck-Hufnagel, S. (2024a). Intonational categories and continua in American English rising nuclear tunes. *Journal of Phonetics*, *104*, 101310. https://doi.org/10.1016/j.wocn.2024.101310

Steffman, J., Cole, J., & Shattuck-Hufnagel, S. (2024b). Intonational categories and continua in American English rising nuclear tunes. *Journal of Phonetics*, *104*, 101310. https://doi.org/10.1016/j.wocn.2024.101310

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the Number of Clusters in a Data Set Via the Gap Statistic. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293

Tilsen, S., Burgess, D., & Lantz, E. (2013). *Imitation of intonational gestures: A preliminary report*. https://doi.org/10.5281/ZENODO.3726927

van Rijn, P., & Larrouy-Maestri, P. (2023). Modelling individual and cross-cultural variation in the mapping of emotions to speech prosody. *Nature Human Behaviour*, *7*(3), 386–396. https://doi.org/10.1038/s41562-022-01505-5

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2019). *Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC*. https://doi.org/10.48550/ARXIV.1903.08008

Veilleux, N., Shattuck-Hufnagel, S., & Brugos, A. (2006). *Transcribing Prosodic Structure of Spoken Utterances with ToBI*.

Ward, G., & Hirschberg, J. (1985). Implicating Uncertainty: The Pragmatics of Fall-Rise Intonation. *Language*, *61*(4), 747–776. https://doi.org/10.2307/414489

Westera, M., Goodhue, D., & Gussenhoven, C. (2020). Meanings of Tones and Tunes. In C. Gussenhoven & A. Chen (Eds.), *The Oxford Handbook of Language Prosody* (pp. 442–453). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198832232.013.29

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Chapter 8: Appendices

Appendix A (Ch. 2): Written materials

Tune	*Emotion*	*Preceding Dialog*	*Target Sentence*	*Continuation Sentence*
HHH	PRIDE	I was just reading about your daughter's award-winning medical unit.	**My Melanie?**	She's so talented.
HHH	LOVE	I received a phone call from your favorite granddaughter on my birthday.	**From Madelyn?**	She's so thoughtful.
HHH	SHAME	Everyone I met today wanted to tell me about your horrifying trip to the vet.	**About Lavender?**	She's so embarrassing.
HHH	ANGER	According to my accountant, you might have to pay your landlord's property taxes.	**For Gallagher?**	I won't let that happen.
HHL	PRIDE	I heard there will be awards given out at the banquet. There'll be one for Tim, Madelyn…	**For Melanie…**	…and hopefully for her entire team
HHL	LOVE	We have so many talented cooks in the family. We have you, your mother, …	**Also Madelyn…**	…and definitely her husband.
HHL	SHAME	You could travel if it wasn't for the drama with your landlord, your work…	**And Lavender…**	…and my fear of flying.
HHL	ANGER	You asked the tenant's union about the repairs, about the taxes…	**About Gallagher…**	...who is always testing my patience.
HLH	PRIDE	Do you know anyone affected by the lawsuit against local doctors?	**Not Melanie…**	She follows every rule to the letter.
HLH	LOVE	I heard your sister will join you on vacation this year, is that true?	**And Madelyn…**	It's the first time in ages we all have the same week off.
HLH	SHAME	Is there anything I should warn your house-sitter about?	**There's Lavender…**	She can demand a lot of attention.
HLH	ANGER	Who is responsible for all these unpaid power bills?	**That's Gallagher…**	I think that it's the only thing he used to pay for.
HLL	PRIDE	Who do you think will win the service award this year?	**It's Melanie.**	She's one of our town's best citizens.
HLL	LOVE	Who is coming over for dinner later?	**My Madelyn.**	I'm making her favorite food.
HLL	SHAME	Whose dog was making all that noise last night?	**That's Lavender.**	I hope the police don't get called.
HLL	ANGER	Who are these tax documents for?	**For Gallagher.**	I shouldn't have to deal with this.
LHH	PRIDE	I heard that the city is creating a new award just to give it to your daughter.	**For Melanie?**	Such an honor.

LHH	LOVE	A reporter called to ask about your favorite granddaughter.	**About Madelyn?**	She's more famous than I even knew.
LHH	SHAME	According to the neighbors, your dog can be overly aggressive.	**My Lavender?**	I didn't know they were complaining.
LHH	ANGER	Apparently your brother was out drinking with your landlord.	**With Gallagher?**	Is that what he's been doing?
LHL	PRIDE	Do we know anyone who can give Madelyn advice about applying to medical school?	**Our Melanie…**	She's the obvious one to ask, having gone through it herself.
LHL	LOVE	Why do we need to be in town this weekend again?	**For Madelyn…**	We have our weekly date, remember?
LHL	SHAME	Where did all of these scratches on your door come from?	**From Lavender…**	You've seen how she acts when I leave.
LHL	ANGER	Who's been giving you so many problems again?	**That's Gallagher…**	But you know all about those issues.
LLH	PRIDE	You were so happy when Madelyn decided to become a doctor.	**My Melanie?**	I think you mean my amazing daughter.
LLH	LOVE	I cleaned your stovetop so that you can cook with Lavender.	**With Madelyn?**	She's much more help than my dog.
LLH	SHAME	I brought over a special treat for the best dog on the block.	**For Lavender?**	She does not deserve a treat.
LLH	ANGER	My neighbor was gifted an expensive bottle of wine from your landlord.	**From Gallagher?**	He should be fixing this building.
LLL	PRIDE	Who's the latest employee of the month at the hospital?	**My Melanie.**	But it doesn't go to her head.
LLL	LOVE	Did anyone send you flowers on your birthday?	**Just Madelyn.**	She's so thoughtful.
LLL	SHAME	Why do you have so many missed calls from the vet?	**It's Lavender.**	She's in real trouble now
LLL	ANGER	There's a pile of boxes on your porch, are they for you?	**For Gallagher.**	I hate that he uses the porch like a PO box.

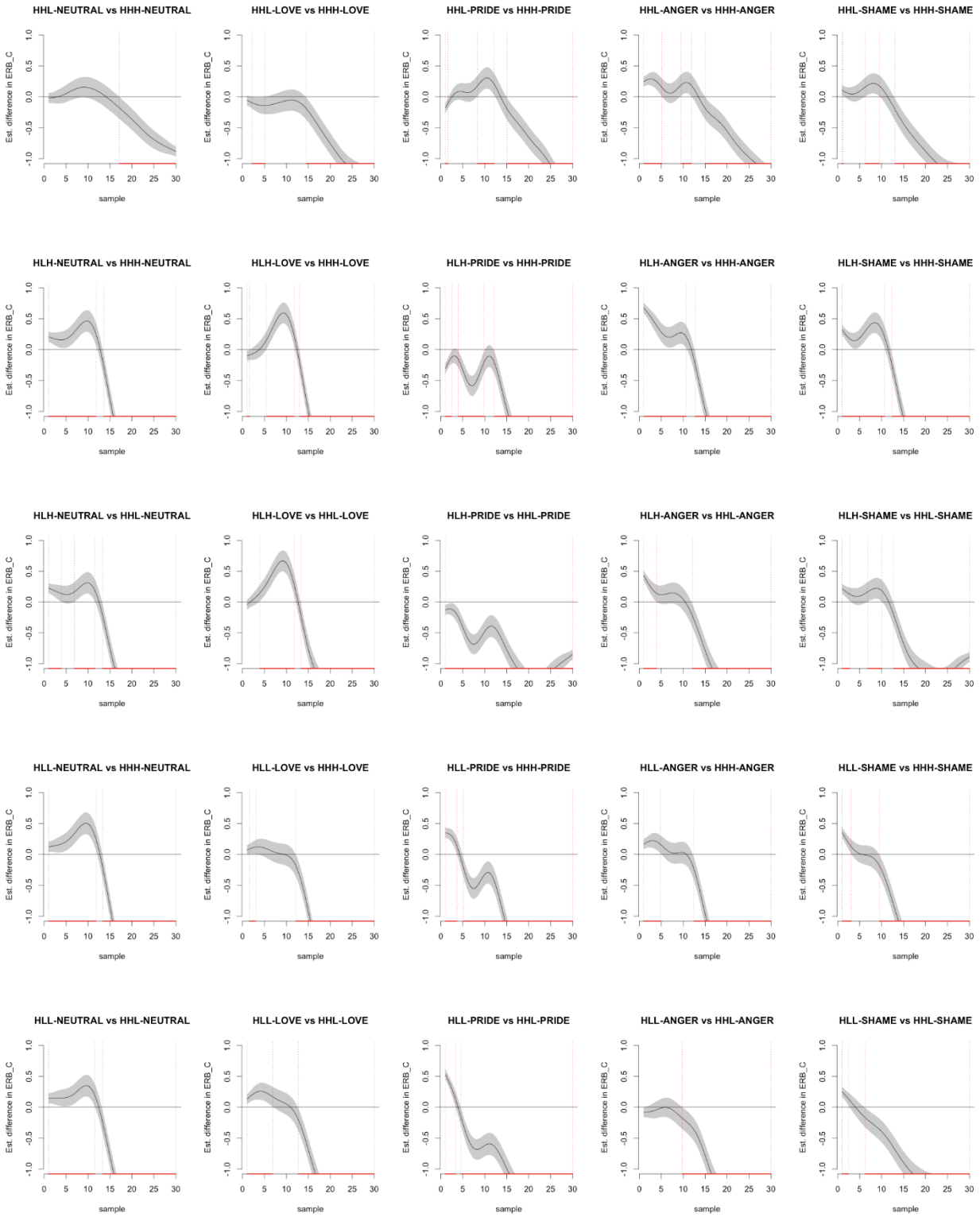
Appendix B (Ch. 2): Primary GAMM output

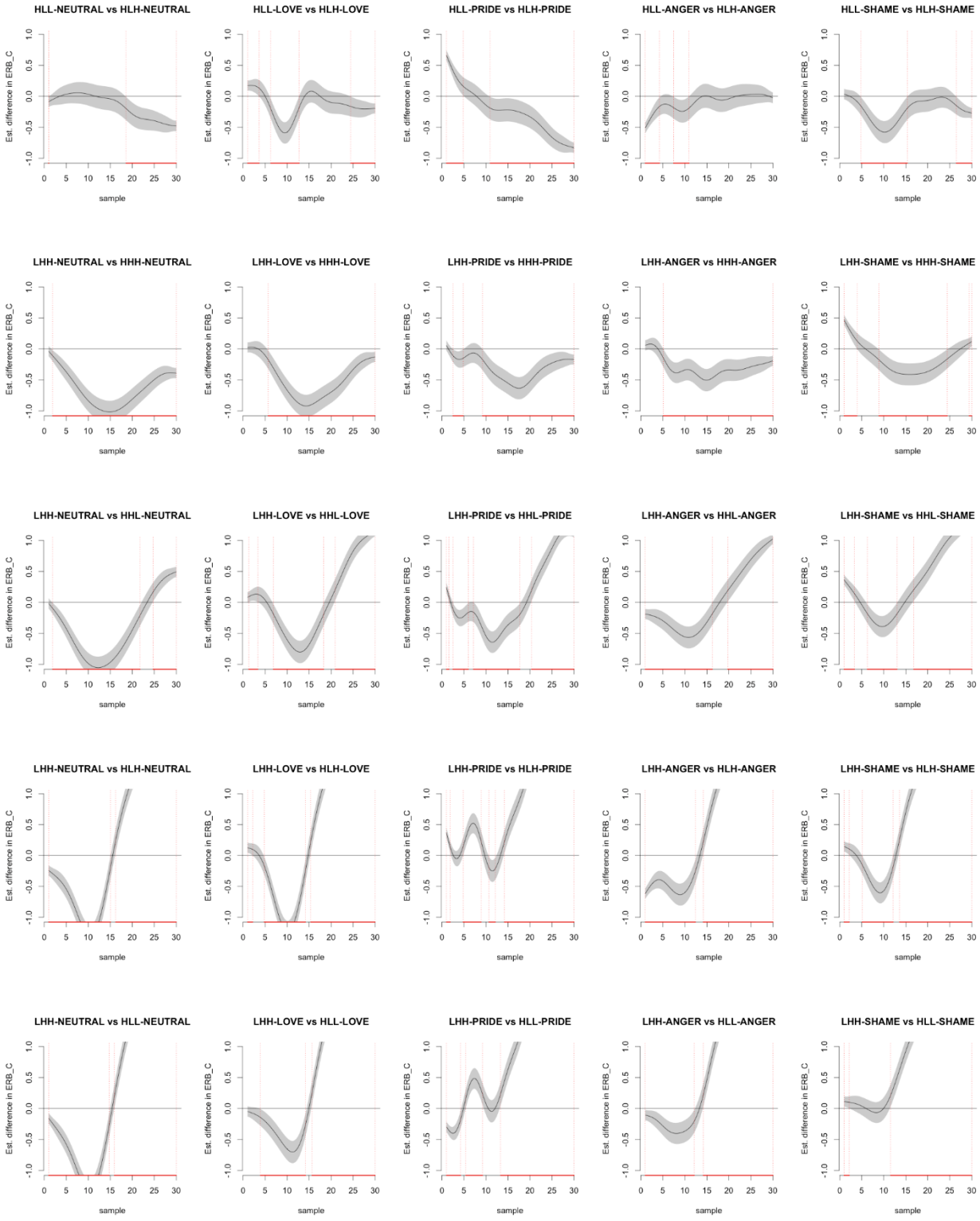
SMOOTH TERMS:					
TERM(S)	**edf**	**Ref. df**	**F**	**p-value**	
LLL	3.404	4.066	5.285	0.000276	***
HHH	7.21	7.681	7.414	< 2e-16	***
HHL	6.425	7.012	5.823	5.73E-06	***
HLH	8.73	8.824	41.223	< 2e-16	***
HLL	7.45	7.607	23.186	< 2e-16	***
LHH	7.652	8.044	11.404	< 2e-16	***
LHL	4.874	5.596	2.991	0.014269	*
LLH	7.398	7.845	5.839	4.66E-06	***
NEUTRAL	7.829	8.335	4.246	2.25E-05	***
ANGER	8.254	8.469	8.051	< 2e-16	***
LOVE	7.471	7.921	7.179	< 2e-16	***
PRIDE	5.9	6.34	1.032	0.399777	
SHAME	7.503	7.964	5.506	5.33E-06	***
PARTICIPANT	194.06	319	3.188	< 2e-16	***
PARAMETRIC COEFFICIENTS:					
TERM(S)	Estimate	Std. Error	t-value	p-value	
(INTERCEPT)	-0.8368	0.0661	-12.66	< 2e-16	***
LLL-NEUTRAL					
HHH	1.76363	0.08845	19.939	< 2e-16	***
HHL	1.54613	0.09102	16.987	< 2e-16	***
HLH	0.67454	0.06073	11.108	< 2e-16	***
HLL	0.40976	0.05496	7.455	9.01E-14	***

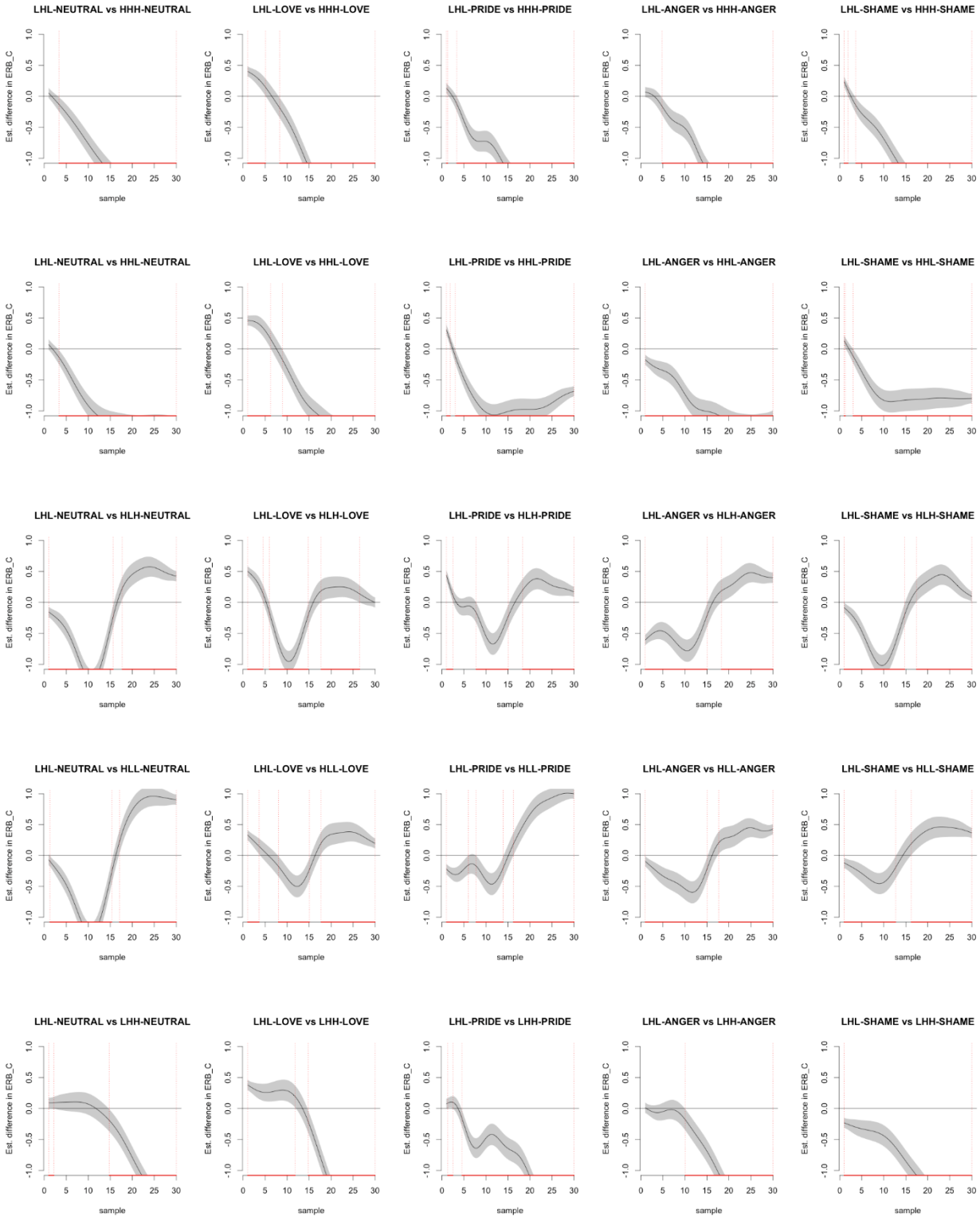
LHH	1.23992	0.06974	17.778	< 2e-16	***
LHL	0.55403	0.0579	9.569	< 2e-16	***
LLH	0.78419	0.06511	12.043	< 2e-16	***
ANGER	-0.20189	0.0587	-3.439	0.000584	***
LOVE	0.31526	0.05999	5.255	1.48E-07	***
PRIDE	0.41268	0.07192	5.738	9.58E-09	***
SHAME	0.12868	0.05396	2.385	0.017083	*
HHH:ANGER	-0.06312	0.04206	-1.501	0.133377	
HHL:ANGER	-0.21424	0.04254	-5.037	4.74E-07	***
HLH:ANGER	0.02643	0.04326	0.611	0.541304	
HLL:ANGER	0.01617	0.04287	0.377	0.706012	
LHH:ANGER	0.09655	0.04226	2.285	0.022344	*
LHL:ANGER	-0.14025	0.04238	-3.309	0.000935	***
LLH:ANGER	0.17198	0.04282	4.016	5.92E-05	***
HHH:LOVE	-0.23399	0.04029	-5.807	6.37E-09	***
HHL:LOVE	-0.54413	0.04082	-13.329	< 2e-16	***
HLH:LOVE	-0.45579	0.04047	-11.263	< 2e-16	***
HLL:LOVE	-0.21731	0.04087	-5.317	1.06E-07	***
LHH:LOVE	-0.11472	0.04078	-2.813	0.004908	**
LHL:LOVE	-0.29505	0.04065	-7.258	3.95E-13	***
LLH:LOVE	0.12185	0.04049	3.009	0.002621	**
HHH:PRIDE	-0.2616	0.04109	-6.367	1.94E-10	***
HHL:PRIDE	-0.56808	0.04132	-13.748	< 2e-16	***
HLH:PRIDE	-0.33259	0.04152	-8.01	1.16E-15	***
HLL:PRIDE	-0.2255	0.04173	-5.404	6.52E-08	***
LHH:PRIDE	-0.06715	0.0407	-1.65	0.098942	.
LHL:PRIDE	-0.16197	0.04148	-3.905	9.43E-05	***
LLH:PRIDE	0.1125	0.04139	2.718	0.006563	**

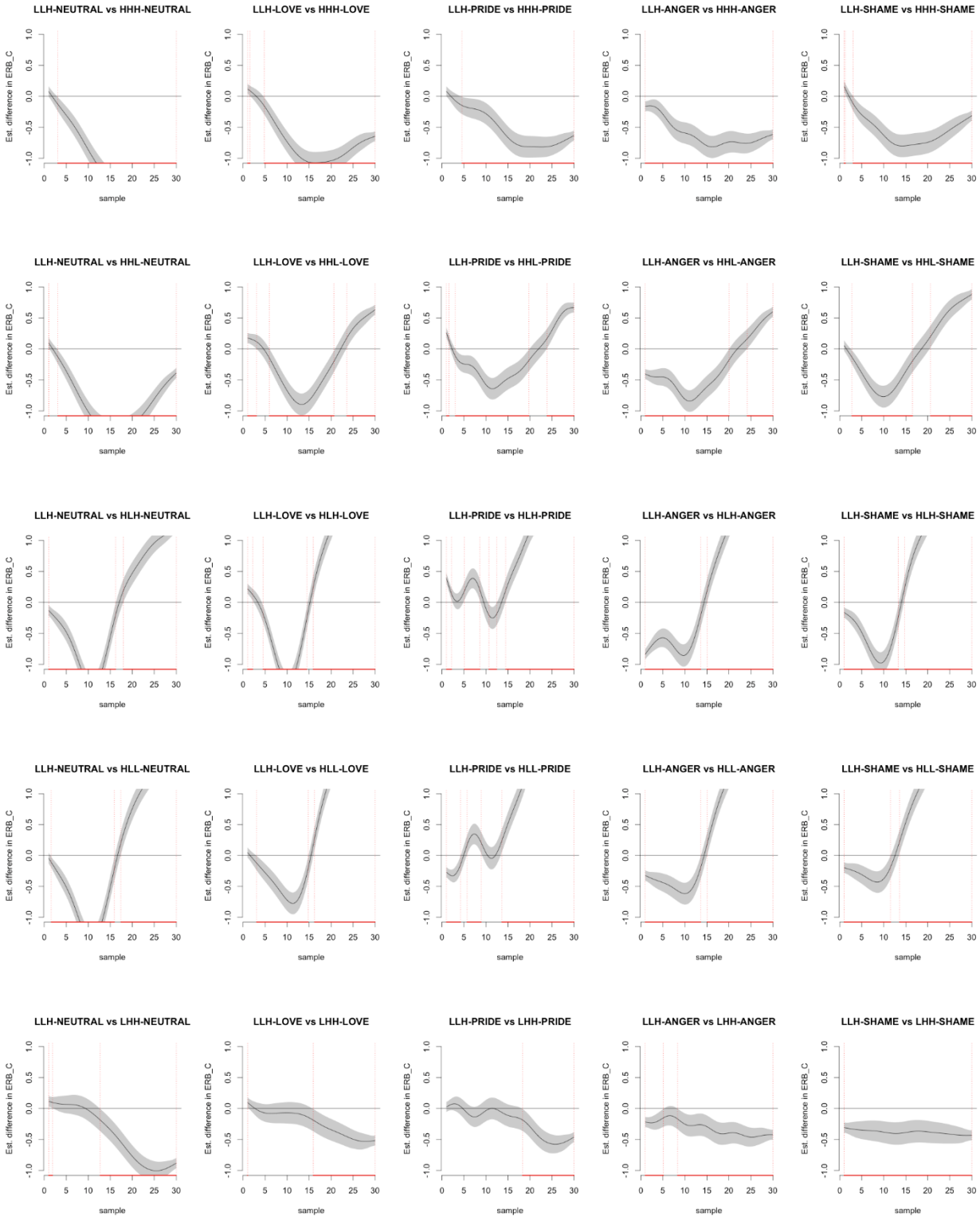
HHH:SHAME	-0.55873	0.04081	-13.692	< 2e-16	***
HHL:SHAME	-0.72036	0.0411	-17.527	< 2e-16	***
HLH:SHAME	-0.29609	0.04073	-7.269	3.64E-13	***
HLL:SHAME	-0.17121	0.04143	-4.133	3.59E-05	***
LHH:SHAME	-0.03806	0.04027	-0.945	0.344635	
LHL:SHAME	-0.43123	0.0407	-10.596	< 2e-16	***

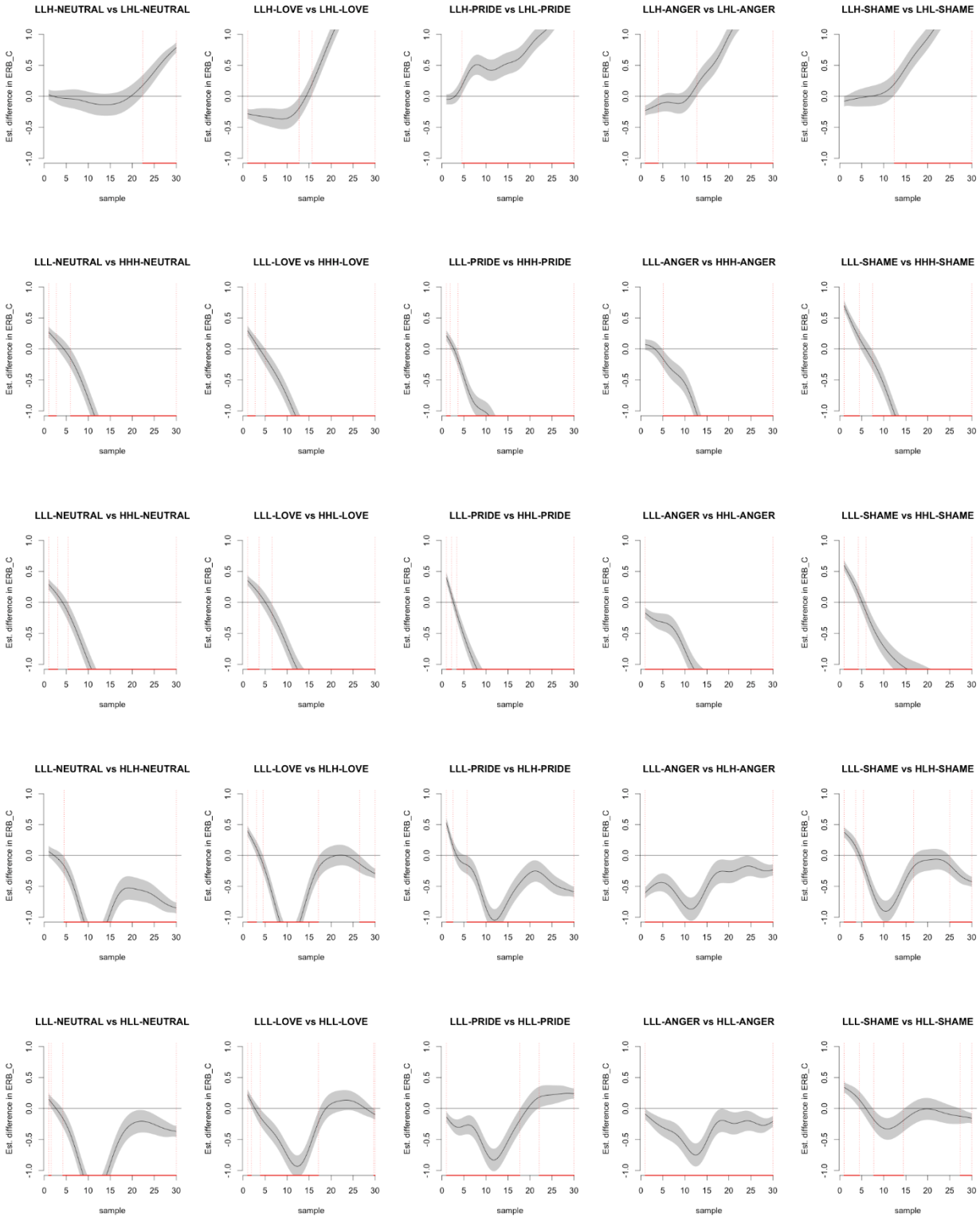
Appendix C (Ch. 2): All between-tune difference GAMMs

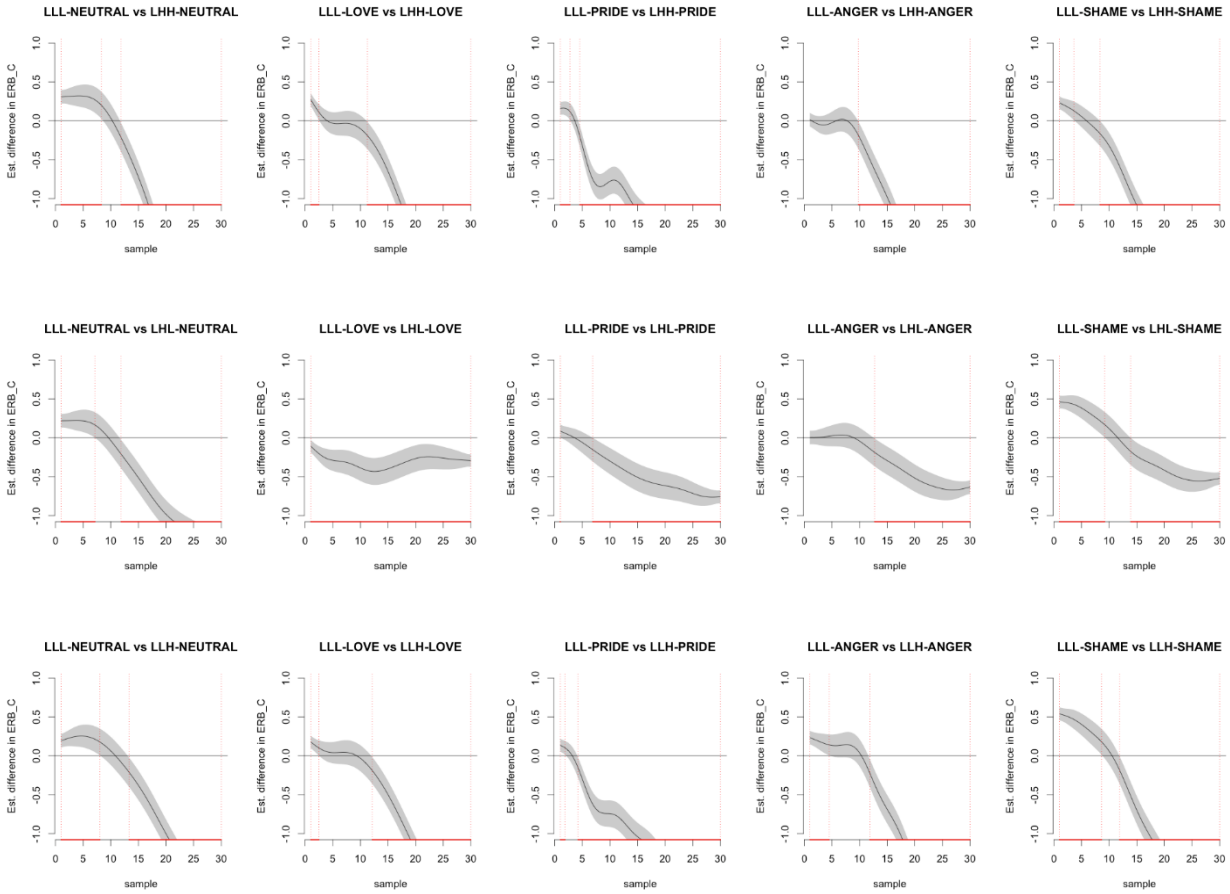



HLL-NEUTRAL vs HLH-NEUTRAL
HLL-LOVE vs HLH-LOVE
HLL-PRIDE vs HLH-PRIDE
HLL-ANGER vs HLH-ANGER
HLL-SHAME vs HLH-SHAME
LHH-NEUTRAL vs HHH-NEUTRAL
LHH-LOVE vs HHH-LOVE
LHH-PRIDE vs HHH-PRIDE
LHH-ANGER vs HHH-ANGER
LHH-SHAME vs HHH-SHAME
LHH-NEUTRAL vs HHL-NEUTRAL
LHH-LOVE vs HHL-LOVE
LHH-PRIDE vs HHL-PRIDE
LHH-ANGER vs HHL-ANGER
LHH-SHAME vs HHL-SHAME
LHH-NEUTRAL vs HLH-NEUTRAL
LHH-LOVE vs HLH-LOVE
LHH-PRIDE vs HLH-PRIDE
LHH-ANGER vs HLH-ANGER
LHH-SHAME vs HLH-SHAME
LHH-NEUTRAL vs HLL-NEUTRAL
LHH-LOVE vs HLL-LOVE
LHH-PRIDE vs HLL-PRIDE
LHH-ANGER vs HLL-ANGER
LHH-SHAME vs HLL-SHAME
Est. difference in ERB_C
sample

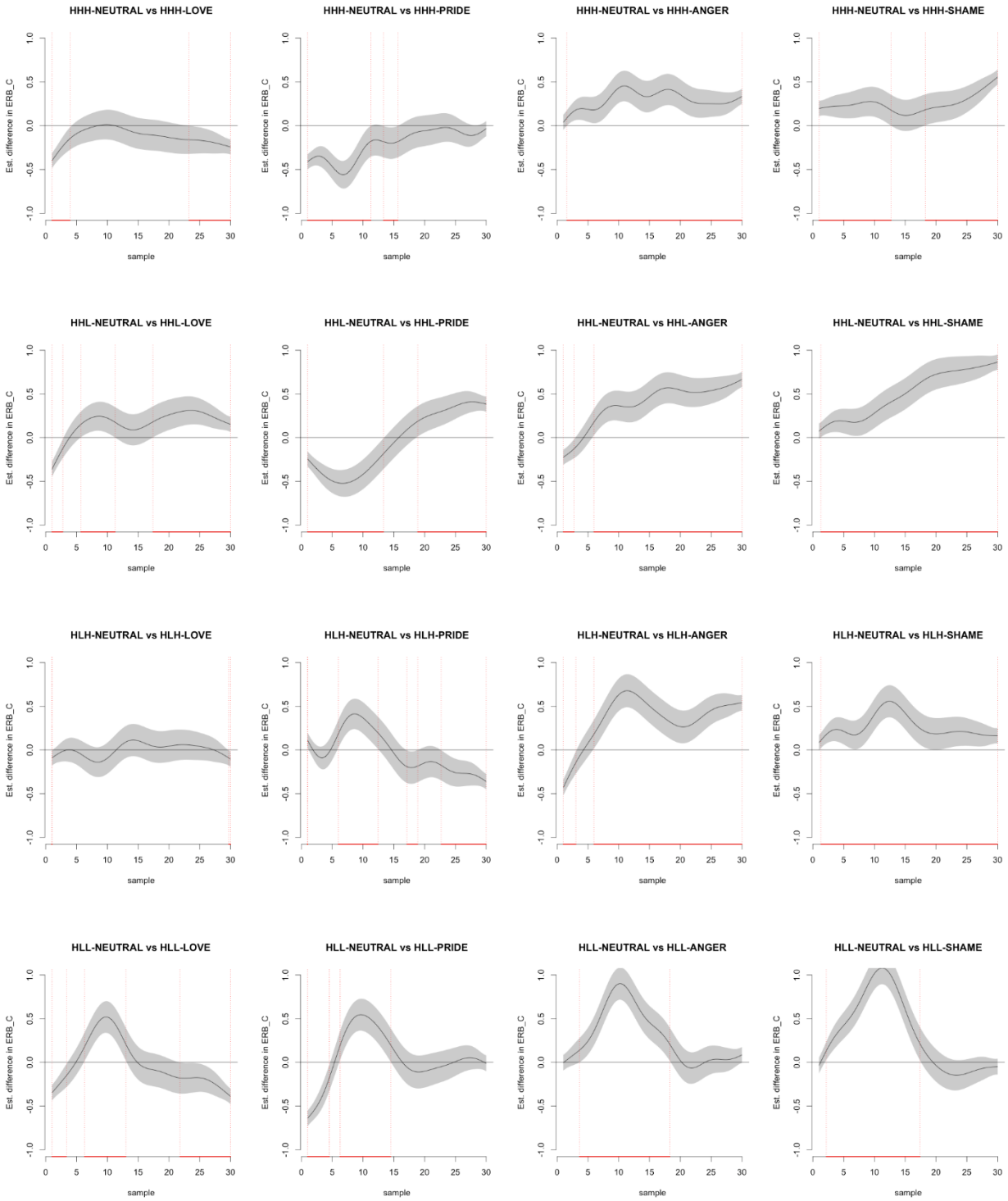

LHL-NEUTRAL vs HHH-NEUTRAL
LHL-LOVE vs HHH-LOVE
LHL-PRIDE vs HHH-PRIDE
LHL-ANGER vs HHH-ANGER
LHL-SHAME vs HHH-SHAME
LHL-NEUTRAL vs HHL-NEUTRAL
LHL-LOVE vs HHL-LOVE
LHL-PRIDE vs HHL-PRIDE
LHL-ANGER vs HHL-ANGER
LHL-SHAME vs HHL-SHAME
LHL-NEUTRAL vs HLH-NEUTRAL
LHL-LOVE vs HLH-LOVE
LHL-PRIDE vs HLH-PRIDE
LHL-ANGER vs HLH-ANGER
LHL-SHAME vs HLH-SHAME
LHL-NEUTRAL vs HLL-NEUTRAL
LHL-LOVE vs HLL-LOVE
LHL-PRIDE vs HLL-PRIDE
LHL-ANGER vs HLL-ANGER
LHL-SHAME vs HLL-SHAME
LHL-NEUTRAL vs LHH-NEUTRAL
LHL-LOVE vs LHH-LOVE
LHL-PRIDE vs LHH-PRIDE
LHL-ANGER vs LHH-ANGER
LHL-SHAME vs LHH-SHAME
Est. difference in ERB_C
sample

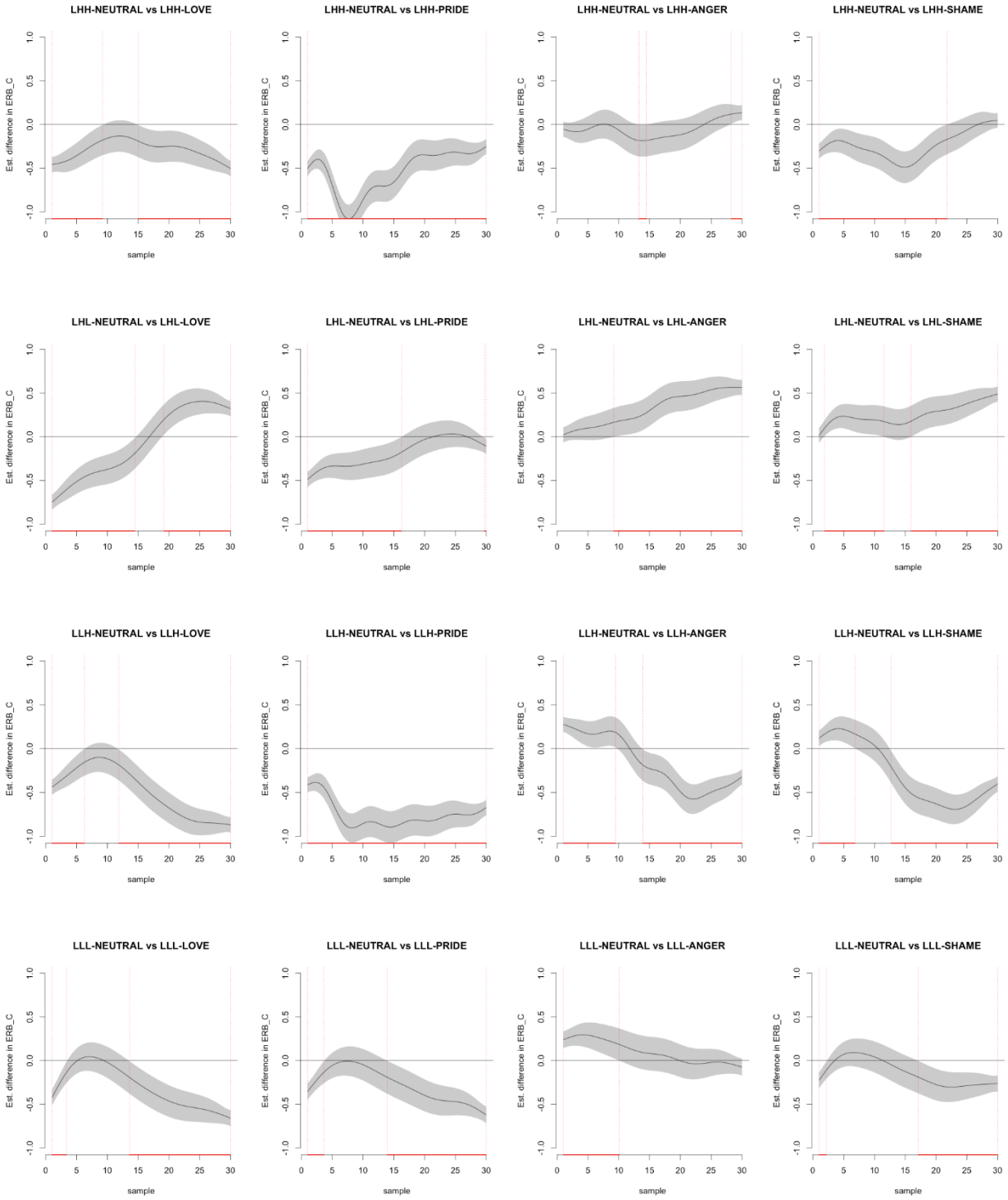

LLH-NEUTRAL vs HHH-NEUTRAL
LLH-LOVE vs HHH-LOVE
LLH-PRIDE vs HHH-PRIDE
LLH-ANGER vs HHH-ANGER
LLH-SHAME vs HHH-SHAME
LLH-NEUTRAL vs HHL-NEUTRAL
LLH-LOVE vs HHL-LOVE
LLH-PRIDE vs HHL-PRIDE
LLH-ANGER vs HHL-ANGER
LLH-SHAME vs HHL-SHAME
LLH-NEUTRAL vs HLH-NEUTRAL
LLH-LOVE vs HLH-LOVE
LLH-PRIDE vs HLH-PRIDE
LLH-ANGER vs HLH-ANGER
LLH-SHAME vs HLH-SHAME
LLH-NEUTRAL vs HLL-NEUTRAL
LLH-LOVE vs HLL-LOVE
LLH-PRIDE vs HLL-PRIDE
LLH-ANGER vs HLL-ANGER
LLH-SHAME vs HLL-SHAME
LLH-NEUTRAL vs LHH-NEUTRAL
LLH-LOVE vs LHH-LOVE
LLH-PRIDE vs LHH-PRIDE
LLH-ANGER vs LHH-ANGER
LLH-SHAME vs LHH-SHAME
Est. difference in ERB_C
sample


LLH-NEUTRAL vs LHL-NEUTRAL
LLH-LOVE vs LHL-LOVE
LLH-PRIDE vs LHL-PRIDE
LLH-ANGER vs LHL-ANGER
LLH-SHAME vs LHL-SHAME
LLL-NEUTRAL vs HHH-NEUTRAL
LLL-LOVE vs HHH-LOVE
LLL-PRIDE vs HHH-PRIDE
LLL-ANGER vs HHH-ANGER
LLL-SHAME vs HHH-SHAME
LLL-NEUTRAL vs HHL-NEUTRAL
LLL-LOVE vs HHL-LOVE
LLL-PRIDE vs HHL-PRIDE
LLL-ANGER vs HHL-ANGER
LLL-SHAME vs HHL-SHAME
LLL-NEUTRAL vs HLH-NEUTRAL
LLL-LOVE vs HLH-LOVE
LLL-PRIDE vs HLH-PRIDE
LLL-ANGER vs HLH-ANGER
LLL-SHAME vs HLH-SHAME
LLL-NEUTRAL vs HLL-NEUTRAL
LLL-LOVE vs HLL-LOVE
LLL-PRIDE vs HLL-PRIDE
LLL-ANGER vs HLL-ANGER
LLL-SHAME vs HLL-SHAME
Est. difference in ERB_C
sample


LLL-NEUTRAL vs LHH-NEUTRAL
LLL-LOVE vs LHH-LOVE
LLL-PRIDE vs LHH-PRIDE
LLL-ANGER vs LHH-ANGER
LLL-SHAME vs LHH-SHAME
LLL-NEUTRAL vs LHL-NEUTRAL
LLL-LOVE vs LHL-LOVE
LLL-PRIDE vs LHL-PRIDE
LLL-ANGER vs LHL-ANGER
LLL-SHAME vs LHL-SHAME
LLL-NEUTRAL vs LLH-NEUTRAL
LLL-LOVE vs LLH-LOVE
LLL-PRIDE vs LLH-PRIDE
LLL-ANGER vs LLH-ANGER
LLL-SHAME vs LLH-SHAME
Est. difference in ERB_C
sample

Appendix D (Ch. 2): All within-tune difference GAMMs

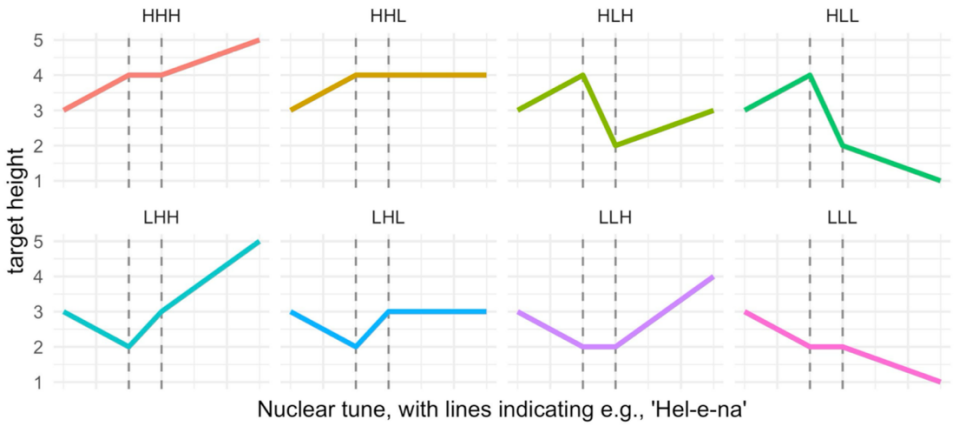



LHH-NEUTRAL vs LHH-LOVE
LHH-NEUTRAL vs LHH-PRIDE
LHH-NEUTRAL vs LHH-ANGER
LHH-NEUTRAL vs LHH-SHAME
LHL-NEUTRAL vs LHL-LOVE
LHL-NEUTRAL vs LHL-PRIDE
LHL-NEUTRAL vs LHL-ANGER
LHL-NEUTRAL vs LHL-SHAME
LLH-NEUTRAL vs LLH-LOVE
LLH-NEUTRAL vs LLH-PRIDE
LLH-NEUTRAL vs LLH-ANGER
LLH-NEUTRAL vs LLH-SHAME
LLL-NEUTRAL vs LLL-LOVE
LLL-NEUTRAL vs LLL-PRIDE
LLL-NEUTRAL vs LLL-ANGER
LLL-NEUTRAL vs LLL-SHAME
Est. difference in ERB_C
sample

Appendix E (Ch. 2): Stimuli F0 targets

From Cole *et al*. (2023:6), target heights for stimuli imitated by participants (top) and corresponding schematized trajectories (bottom).

Relative F0 Level	Male model speaker	Female model speaker
1	80	100
2	105	160
3	130	200
4	225	300
5	265	380



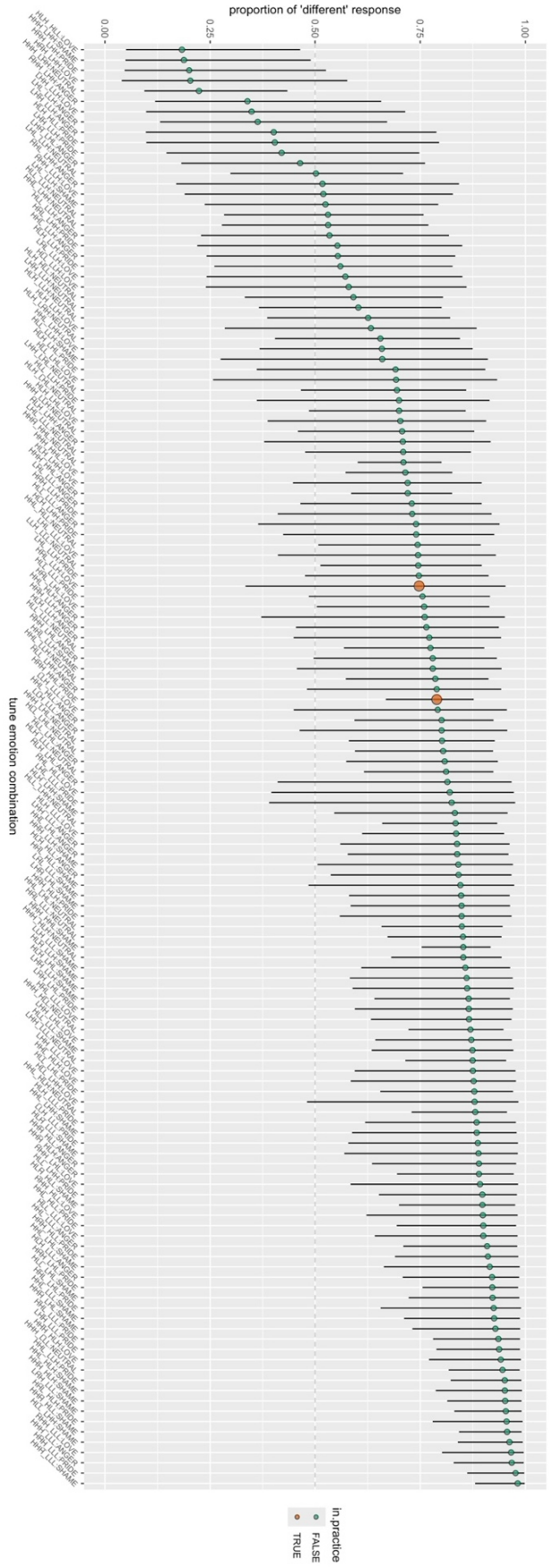
Appendix A (Ch. 3): Summary for the Emotion Matching Model (EMM)

```
fit emotion-match model: 53.605 sec elapsed
 Family: bernoulli
  Links: mu = logit
Formula: different ~ t * e + (1 | s)
   Data: step_1_data (Number of observations: 1214)
  Draws: 2 chains, each with iter = 10000; warmup = 1000; thin = 1;
         total post-warmup draws = 18000

Multilevel Hyperparameters:
~s (Number of levels: 150)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     1.26      0.15     0.99     1.58 1.00     6514    10478

Regression Coefficients:
               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept          0.89      0.25     0.41     1.39 1.00    13289    14509
tHHH_HLH           0.87      0.51    -0.10     1.90 1.00    21661    14916
tHHH_HLL           1.01      0.49     0.07     2.00 1.00    22283    14992
tHHH_LHH          -2.15      0.51    -3.17    -1.16 1.00    20020    15569
tHHH_LHL           0.51      0.55    -0.56     1.61 1.00    19992    15207
tHHH_LLH          -0.01      0.53    -1.05     1.06 1.00    20683    15237
tHHH_LLL           1.97      0.69     0.66     3.34 1.00    30838    14701
tHHL_HLH           1.12      0.53     0.11     2.17 1.00    23717    13587
tHHL_HLL           0.18      0.53    -0.84     1.22 1.00    21307    14513
tHHL_LHH          -0.77      0.54    -1.83     0.29 1.00    20907    15604
tHHL_LHL           0.84      0.55    -0.22     1.92 1.00    19593    14668
tHHL_LLH           0.41      0.52    -0.60     1.43 1.00    20168    15024
tHHL_LLL           0.86      0.53    -0.16     1.89 1.00    21029    14968
tHLH_HLL          -0.53      0.53    -1.58     0.50 1.00    20929    15325
tHLH_LHH          -0.25      0.52    -1.25     0.76 1.00    17602    14842
tHLH_LHL          -0.04      0.48    -0.97     0.91 1.00    19411    14412
tHLH_LLH          -0.38      0.50    -1.37     0.60 1.00    19418    14190
tHLH_LLL           0.57      0.51    -0.41     1.58 1.00    22095    15572
tHLL_LHH           0.73      0.50    -0.23     1.71 1.00    24956    15154
tHLL_LHL           0.52      0.52    -0.48     1.57 1.00    19990    14629
tHLL_LLH          -0.78      0.52    -1.81     0.24 1.00    19511    15239
tHLL_LLL           0.34      0.49    -0.61     1.32 1.00    21120    15541
tLHH_LHL          -0.07      0.49    -1.03     0.90 1.00    20050    15235
tLHH_LLH          -0.48      0.49    -1.44     0.49 1.00    19712    15015
tLHH_LLL           1.06      0.53     0.04     2.12 1.00    20746    14407
tLHL_LLH          -0.89      0.45    -1.76    -0.01 1.00    18129    14787
tLHL_LLL          -0.00      0.51    -0.98     1.01 1.00    19326    15499
tLLH_LLL           0.18      0.53    -0.83     1.25 1.00    23461    14869
eANGER             0.05      0.29    -0.53     0.63 1.00    16238    14646
ePRIDE             0.43      0.30    -0.16     1.02 1.00    17522    14376
eLOVE              0.03      0.31    -0.57     0.63 1.00    17909    14909
eSHAME             0.86      0.31     0.26     1.48 1.00    17371    15086
tHHH_HLH:eANGER    0.30      0.72    -1.09     1.76 1.00    28006    14275
tHHH_HLL:eANGER    0.16      0.80    -1.39     1.76 1.00    32940    13621
tHHH_LHH:eANGER    0.53      0.71    -0.85     1.92 1.00    26514    14844
tHHH_LHL:eANGER   -0.18      0.72    -1.60     1.25 1.00    25199    14178
```

tHHH_LLH:eANGER	0.26	0.74	-1.18	1.73	1.00	27653	14536
tHHH_LLL:eANGER	0.52	0.87	-1.15	2.26	1.00	31939	12791
tHHL_HLH:eANGER	-0.89	0.83	-2.50	0.77	1.00	33088	13239
tHHL_HLL:eANGER	0.28	0.65	-0.99	1.57	1.00	23428	14843
tHHL_LHH:eANGER	-0.11	0.79	-1.65	1.44	1.00	32184	13851
tHHL_LHL:eANGER	-0.11	0.74	-1.56	1.35	1.00	27061	14950
tHHL_LLH:eANGER	-1.13	0.73	-2.57	0.29	1.00	26405	14236
tHHL_LLL:eANGER	0.53	0.77	-0.95	2.06	1.00	29282	13929
tHLH_HLL:eANGER	1.27	0.79	-0.26	2.85	1.00	28852	13364
tHLH_LHH:eANGER	0.21	0.74	-1.23	1.67	1.00	27866	14598
tHLH_LHL:eANGER	0.59	0.87	-1.11	2.33	1.00	36801	13705
tHLH_LLH:eANGER	0.66	0.76	-0.80	2.18	1.00	32397	13594
tHLH_LLL:eANGER	0.96	0.80	-0.59	2.57	1.00	34873	13594
tHLL_LHH:eANGER	-0.35	0.73	-1.78	1.09	1.00	28558	14234
tHLL_LHL:eANGER	-0.02	0.67	-1.35	1.27	1.00	24508	15569
tHLL_LLH:eANGER	-0.03	0.72	-1.43	1.37	1.00	23966	13886
tHLL_LLL:eANGER	-0.23	0.80	-1.78	1.34	1.00	34556	14255
tLHH_LHL:eANGER	-1.03	0.70	-2.41	0.34	1.00	26440	14104
tLHH_LLH:eANGER	-0.89	0.84	-2.56	0.75	1.00	31529	12867
tLHH_LLL:eANGER	-0.34	0.76	-1.80	1.19	1.00	27832	14228
tLHL_LLH:eANGER	-0.63	0.69	-1.99	0.72	1.00	26080	14596
tLHL_LLL:eANGER	0.05	0.67	-1.27	1.37	1.00	22845	14826
tLLH_LLL:eANGER	0.28	0.79	-1.25	1.85	1.00	37459	14200
tHHH_HLH:ePRIDE	-0.45	0.77	-1.92	1.09	1.00	31912	14247
tHHH_HLL:ePRIDE	0.02	0.80	-1.50	1.63	1.00	32282	12782
tHHH_LHH:ePRIDE	-0.59	0.79	-2.15	0.93	1.00	30035	12766
tHHH_LHL:ePRIDE	0.65	0.74	-0.76	2.15	1.00	27867	13973
tHHH_LLH:ePRIDE	-0.31	0.72	-1.74	1.11	1.00	25831	14027
tHHH_LLL:ePRIDE	0.47	0.88	-1.23	2.23	1.00	35965	13068
tHHL_HLH:ePRIDE	0.64	0.87	-1.02	2.36	1.00	32374	13238
tHHL_HLL:ePRIDE	0.73	0.75	-0.71	2.22	1.00	27264	14412
tHHL_LHH:ePRIDE	-0.34	0.76	-1.81	1.17	1.00	32384	15279
tHHL_LHL:ePRIDE	0.37	0.89	-1.35	2.16	1.00	38272	12385
tHHL_LLH:ePRIDE	1.27	0.77	-0.20	2.81	1.00	30667	14023
tHHL_LLL:ePRIDE	0.53	0.76	-0.92	2.05	1.00	31844	14038
tHLH_HLL:ePRIDE	-1.19	0.84	-2.85	0.46	1.00	35957	13165
tHLH_LHH:ePRIDE	-0.01	0.74	-1.44	1.44	1.00	23711	14530
tHLH_LHL:ePRIDE	-0.46	0.74	-1.88	1.00	1.00	30046	14202
tHLH_LLH:ePRIDE	-0.70	0.71	-2.09	0.69	1.00	25330	13704
tHLH_LLL:ePRIDE	0.16	0.80	-1.39	1.76	1.00	32013	14231
tHLL_LHH:ePRIDE	0.15	0.79	-1.36	1.73	1.00	31998	14208
tHLL_LHL:ePRIDE	0.15	0.73	-1.23	1.60	1.00	26008	14562
tHLL_LLH:ePRIDE	0.31	0.76	-1.17	1.80	1.00	30544	14164
tHLL_LLL:ePRIDE	-0.53	0.68	-1.86	0.79	1.00	27075	15144
tLHH_LHL:ePRIDE	0.62	0.77	-0.86	2.14	1.00	32341	13466
tLHH_LLH:ePRIDE	-1.17	0.73	-2.59	0.26	1.00	29246	14235
tLHH_LLL:ePRIDE	0.36	0.76	-1.07	1.87	1.00	29937	14541
tLHL_LLH:ePRIDE	0.66	0.67	-0.62	1.98	1.00	25139	14648
tLHL_LLL:ePRIDE	0.24	0.94	-1.57	2.10	1.00	36865	11395
tLLH_LLL:ePRIDE	0.58	0.86	-1.08	2.29	1.00	35873	13127
tHHH_HLH:eLOVE	1.02	0.80	-0.52	2.62	1.00	31329	12388
tHHH_HLL:eLOVE	0.26	0.84	-1.36	1.92	1.00	33065	13482
tHHH_LHH:eLOVE	-0.16	0.82	-1.79	1.43	1.00	29272	13903
tHHH_LHL:eLOVE	0.69	0.86	-0.95	2.41	1.00	31308	13698
tHHH_LLH:eLOVE	-0.83	0.77	-2.35	0.66	1.00	29489	13919
tHHH_LLL:eLOVE	0.49	0.90	-1.24	2.28	1.00	36802	12704

tHHL_HLH:eLOVE	-0.04	0.82	-1.65	1.60	1.00	35725	13267
tHHL_HLL:eLOVE	0.44	0.91	-1.32	2.24	1.00	34869	13382
tHHL_LHH:eLOVE	0.52	0.71	-0.85	1.89	1.00	22044	15268
tHHL_LHL:eLOVE	-0.60	0.69	-1.96	0.75	1.00	22903	15382
tHHL_LLH:eLOVE	-0.22	0.87	-1.89	1.53	1.00	33476	12383
tHHL_LLL:eLOVE	0.11	0.74	-1.34	1.59	1.00	28909	14239
tHLH_HLL:eLOVE	-1.90	0.74	-3.36	-0.47	1.00	28825	14297
tHLH_LHH:eLOVE	0.28	0.68	-1.04	1.62	1.00	23713	14861
tHLH_LHL:eLOVE	-0.01	0.72	-1.39	1.42	1.00	26616	15045
tHLH_LLH:eLOVE	-0.00	0.75	-1.47	1.47	1.00	29177	14042
tHLH_LLL:eLOVE	0.15	0.69	-1.16	1.49	1.00	25904	15300
tHLL_LHH:eLOVE	0.34	0.94	-1.49	2.19	1.00	34758	12865
tHLL_LHL:eLOVE	-0.64	0.86	-2.33	1.05	1.00	35906	13577
tHLL_LLH:eLOVE	0.18	0.76	-1.30	1.66	1.00	27067	14201
tHLL_LLL:eLOVE	0.95	0.82	-0.61	2.59	1.00	33854	13456
tLHH_LHL:eLOVE	1.08	0.73	-0.32	2.54	1.00	27851	13996
tLHH_LLH:eLOVE	-1.08	0.77	-2.61	0.43	1.00	32825	13316
tLHH_LLL:eLOVE	-0.00	0.81	-1.56	1.58	1.00	32702	13526
tLHL_LLH:eLOVE	0.26	0.74	-1.20	1.70	1.00	28660	13958
tLHL_LLL:eLOVE	0.16	0.75	-1.31	1.65	1.00	27075	14244
tLLH_LLL:eLOVE	0.25	0.80	-1.28	1.84	1.00	31985	13810
tHHH_HLH:eSHAME	0.39	0.79	-1.10	1.95	1.00	35690	13605
tHHH_HLL:eSHAME	0.35	0.75	-1.09	1.86	1.00	30405	13583
tHHH_LHH:eSHAME	-1.09	0.75	-2.56	0.35	1.00	27487	14526
tHHH_LHL:eSHAME	0.34	0.81	-1.20	1.96	1.00	33522	13916
tHHH_LLH:eSHAME	-0.06	0.83	-1.68	1.61	1.00	34411	13261
tHHH_LLL:eSHAME	0.30	0.91	-1.45	2.13	1.00	37416	12913
tHHL_HLH:eSHAME	0.12	0.82	-1.45	1.77	1.00	31480	13293
tHHL_HLL:eSHAME	-0.22	0.87	-1.89	1.47	1.00	32239	12697
tHHL_LHH:eSHAME	1.07	0.82	-0.52	2.70	1.00	36959	14636
tHHL_LHL:eSHAME	-0.19	0.84	-1.80	1.48	1.00	32269	13163
tHHL_LLH:eSHAME	-0.89	0.76	-2.35	0.61	1.00	28357	14130
tHHL_LLL:eSHAME	-0.07	0.81	-1.62	1.59	1.00	34544	13494
tHLH_HLL:eSHAME	0.97	0.74	-0.47	2.45	1.00	26187	14793
tHLH_LHH:eSHAME	0.10	0.74	-1.35	1.58	1.00	28035	14727
tHLH_LHL:eSHAME	0.14	0.79	-1.37	1.73	1.00	33006	13632
tHLH_LLH:eSHAME	0.47	0.78	-1.05	2.02	1.00	30979	13864
tHLH_LLL:eSHAME	-0.24	0.87	-1.91	1.48	1.00	33448	13752
tHLL_LHH:eSHAME	0.78	0.82	-0.78	2.45	1.00	35501	12917
tHLL_LHL:eSHAME	0.22	0.80	-1.30	1.81	1.00	31596	13466
tHLL_LLH:eSHAME	-0.31	0.80	-1.88	1.25	1.00	30622	13579
tHLL_LLL:eSHAME	-0.14	0.75	-1.58	1.35	1.00	28972	14511
tLHH_LHL:eSHAME	0.06	0.75	-1.40	1.54	1.00	29001	13958
tLHH_LLH:eSHAME	0.60	0.71	-0.76	2.03	1.00	25179	14606
tLHH_LLL:eSHAME	0.24	0.77	-1.25	1.80	1.00	27909	13911
tLHL_LLH:eSHAME	-0.77	0.68	-2.10	0.56	1.00	25786	14118
tLHL_LLL:eSHAME	-0.03	0.75	-1.49	1.46	1.00	29313	14200
tLLH_LLL:eSHAME	-0.13	0.74	-1.57	1.32	1.00	29606	13498

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).


proportion of 'different' response
0.00
0.25
0.50
0.75
1.00
tune emotion combination
in.practice
FALSE
TRUE

Appendix B (Ch. 3): Summary for the Tune Match Model (TMM)

```
fit tune-match model: 69.014 sec elapsed
 Family: bernoulli
  Links: mu = logit
Formula: different ~ e + (1 | s)
   Data: step_1_data (Number of observations: 4785)
  Draws: 2 chains, each with iter = 10000; warmup = 1000; thin = 1;
         total post-warmup draws = 18000

Multilevel Hyperparameters:
~s (Number of levels: 150)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.65      0.06     0.55     0.77 1.00     7054    10267

Regression Coefficients:
                          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                     0.65      0.06     0.53     0.78 1.00     6458     9843
eLOVE_NEUTRALMPRIDE_SHAME     0.09      0.09    -0.09     0.28 1.00    27973    13382
eANGER_LOVEMPRIDE_SHAME       0.02      0.10    -0.17     0.22 1.00    28380    13517
eANGER_NEUTRALMPRIDE_SHAME   -0.48      0.10    -0.67    -0.29 1.00    28864    13438
eANGER_PRIDEMPRIDE_SHAME     -0.35      0.09    -0.53    -0.17 1.00    29345    13519
eANGER_SHAMEMPRIDE_SHAME      0.61      0.11     0.41     0.82 1.00    32361    13638
eLOVE_PRIDEMPRIDE_SHAME       0.91      0.11     0.69     1.13 1.00    27185    12957
eLOVE_SHAMEMPRIDE_SHAME      -0.43      0.09    -0.61    -0.25 1.00    29246    14244
eNEUTRAL_PRIDEMPRIDE_SHAME    0.54      0.11     0.32     0.76 1.00    26848    13747
eNEUTRAL_SHAMEMPRIDE_SHAME   -0.11      0.09    -0.29     0.07 1.00    33958    12766

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Appendix C (Ch. 3): Summary for the Tune-Only Model (TOM

```
fit tune only model*: 93.374 sec elapsed
 Family: bernoulli
  Links: mu = logit
Formula: correct ~ t + (1 | s)
   Data: mdata (Number of observations: 4746)
  Draws: 2 chains, each with iter = 10000; warmup = 1000; thin = 1;
         total post-warmup draws = 18000

Multilevel Hyperparameters:
~s (Number of levels: 150)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.58      0.06     0.47     0.70 1.00     6946    10262

Regression Coefficients:
                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept           1.32      0.06     1.20     1.44 1.00    13037    12595
tHHH_HHLMLLH_LLL   -0.64      0.18    -0.99    -0.28 1.00    25629    12934
tHHH_HLHMLLH_LLL    0.09      0.18    -0.26     0.45 1.00    26775    12315
tHHH_HLLMLLH_LLL    0.90      0.25     0.43     1.40 1.00    26342    13261
tHHH_LHHMLLH_LLL   -1.02      0.16    -1.32    -0.71 1.00    24917    13745
tHHH_LHLMLLH_LLL    1.14      0.26     0.65     1.69 1.00    26415    13540
tHHH_LLHMLLH_LLL   -0.47      0.17    -0.80    -0.13 1.00    25954    13618
tHHH_LLLMLLH_LLL    0.95      0.24     0.49     1.43 1.00    27812    13383
tHHL_HLHMLLH_LLL   -0.14      0.17    -0.47     0.21 1.00    27744    12245
tHHL_HLLMLLH_LLL    0.23      0.20    -0.14     0.63 1.00    24932    12376
tHHL_LHHMLLH_LLL   -0.24      0.18    -0.59     0.11 1.00    26869    13798
tHHL_LHLMLLH_LLL    0.37      0.21    -0.02     0.79 1.00    24714    12466
```

tHHL_LLHMLLH_LLL	-0.40	0.18	-0.75	-0.05	1.00	25164	13706
tHHL_LLLMLLH_LLL	0.42	0.20	0.03	0.83	1.00	26939	13369
tHLH_HLLMLLH_LLL	-1.13	0.15	-1.43	-0.83	1.00	27212	13427
tHLH_LHHMLLH_LLL	-0.31	0.17	-0.64	0.03	1.00	27556	14319
tHLH_LHLMLLH_LLL	0.26	0.18	-0.09	0.62	1.00	24896	13426
tHLH_LLHMLLH_LLL	-0.17	0.19	-0.53	0.21	1.00	25947	13739
tHLH_LLLMLLH_LLL	-0.09	0.18	-0.44	0.26	1.00	28443	12660
tHLL_LHHMLLH_LLL	0.19	0.20	-0.18	0.58	1.00	26125	13966
tHLL_LHLMLLH_LLL	0.11	0.20	-0.27	0.51	1.00	27386	12875
tHLL_LLHMLLH_LLL	-0.11	0.18	-0.46	0.24	1.00	25189	13437
tHLL_LLLMLLH_LLL	-0.53	0.16	-0.83	-0.22	1.00	26080	14097
tLHH_LHLMLLH_LLL	0.93	0.25	0.46	1.43	1.00	25919	12490
tLHH_LLHMLLH_LLL	-0.54	0.16	-0.84	-0.23	1.00	26198	12177
tLHH_LLLMLLH_LLL	1.29	0.29	0.76	1.88	1.00	26101	12868
tLHL_LLHMLLH_LLL	-0.33	0.17	-0.66	0.01	1.00	25434	12698
tLHL_LLLMLLH_LLL	-0.65	0.16	-0.96	-0.33	1.00	26029	13228

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).

Appendix D (Ch. 3): Summary for the Tune-Emotion Interaction Model (TIM)

```
fit tune emotion model*: 497.215 sec elapsed
 Family: bernoulli
  Links: mu = logit
Formula: correct ~ t * e + (1 | s)
   Data: mdata (Number of observations: 4746)
  Draws: 2 chains, each with iter = 10000; warmup = 1000; thin = 1;
         total post-warmup draws = 18000

Multilevel Hyperparameters:
~s (Number of levels: 150)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.67      0.06     0.55     0.80 1.00     7887    12311

Regression Coefficients:
                              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                         1.68      0.13     1.42     1.95 1.00    13963    13571
tHHH_HHLMLLH_LLL                 -0.81      0.30    -1.40    -0.22 1.00    18833    14718
tHHH_HLHMLLH_LLL                  0.21      0.32    -0.42     0.86 1.00    19167    15393
tHHH_HLLMLLH_LLL                  0.87      0.37     0.16     1.62 1.00    21813    15150
tHHH_LHHMLLH_LLL                 -0.94      0.29    -1.50    -0.38 1.00    18098    15498
tHHH_LHLMLLH_LLL                  1.18      0.40     0.42     1.98 1.00    21643    14906
tHHH_LLHMLLH_LLL                 -0.56      0.30    -1.14     0.04 1.00    18664    14205
tHHH_LLLMLLH_LLL                  0.93      0.38     0.21     1.70 1.00    22790    15010
tHHL_HLHMLLH_LLL                 -0.22      0.30    -0.80     0.38 1.00    19903    15154
tHHL_HLLMLLH_LLL                  0.28      0.34    -0.39     0.97 1.00    20230    14843
tHHL_LHHMLLH_LLL                 -0.31      0.32    -0.94     0.34 1.00    18826    14274
tHHL_LHLMLLH_LLL                  0.08      0.33    -0.58     0.74 1.00    21553    15651
tHHL_LLHMLLH_LLL                 -0.47      0.32    -1.09     0.17 1.00    19226    14636
tHHL_LLLMLLH_LLL                 -0.00      0.34    -0.66     0.67 1.00    21975    14598
tHLH_HLLMLLH_LLL                 -1.43      0.30    -2.03    -0.84 1.00    16315    14234
tHLH_LHHMLLH_LLL                 -0.40      0.31    -1.00     0.20 1.00    18621    13844
tHLH_LHLMLLH_LLL                  0.02      0.31    -0.56     0.63 1.00    19738    13921
tHLH_LLHMLLH_LLL                 -0.09      0.33    -0.73     0.56 1.00    19308    14771
tHLH_LLLMLLH_LLL                  0.01      0.32    -0.61     0.64 1.00    19097    14656
tHLL_LHHMLLH_LLL                  0.48      0.35    -0.19     1.18 1.00    19708    15517
tHLL_LHLMLLH_LLL                  0.14      0.35    -0.54     0.83 1.00    18525    14129
tHLL_LLHMLLH_LLL                 -0.22      0.30    -0.81     0.37 1.00    19522    14568
tHLL_LLLMLLH_LLL                 -0.46      0.30    -1.05     0.14 1.00    17540    14253
tLHH_LHLMLLH_LLL                  0.91      0.37     0.19     1.65 1.00    22351    14906
tLHH_LLHMLLH_LLL                 -0.92      0.29    -1.49    -0.36 1.00    17718    14627
tLHH_LLLMLLH_LLL                  1.39      0.40     0.62     2.18 1.00    22974    15301
tLHL_LLHMLLH_LLL                 -0.12      0.32    -0.74     0.52 1.00    19409    15642
tLHL_LLLMLLH_LLL                 -0.55      0.30    -1.14     0.04 1.00    18400    14658
eANGER_NEUTRAL                    0.04      0.18    -0.31     0.39 1.00    16917    15463
eANGER_PRIDE                      0.49      0.21     0.09     0.92 1.00    17036    14155
eANGER_SHAME                     -0.93      0.16    -1.25    -0.61 1.00    15985    15580
eLOVE_NEUTRAL                    -0.60      0.17    -0.94    -0.26 1.00    16117    15452
eLOVE_PRIDE                      -0.59      0.17    -0.92    -0.25 1.00    16243    15636
eLOVE_SHAME                       0.16      0.19    -0.20     0.53 1.00    18018    15888
eNEUTRAL_PRIDE                   -0.62      0.18    -0.96    -0.27 1.00    16825    15393
eNEUTRAL_SHAME                    0.15      0.19    -0.21     0.52 1.00    17423    14963
ePRIDE_SHAME                      0.29      0.19    -0.09     0.67 1.00    18604    15508
tHHH_HHLMLLH_LLL:eANGER_NEUTRAL  -0.45      0.56    -1.55     0.66 1.00    29266    13838
tHHH_HLHMLLH_LLL:eANGER_NEUTRAL  -0.56      0.59    -1.69     0.63 1.00    30845    13775
tHHH_HLLMLLH_LLL:eANGER_NEUTRAL   0.41      0.69    -0.90     1.84 1.00    32460    12709
tHHH_LHHMLLH_LLL:eANGER_NEUTRAL   0.11      0.54    -0.92     1.19 1.00    26099    13684
tHHH_LHLMLLH_LLL:eANGER_NEUTRAL   0.63      0.80    -0.88     2.27 1.00    35234    12571
tHHH_LLHMLLH_LLL:eANGER_NEUTRAL   0.80      0.56    -0.26     1.93 1.00    27812    13156
tHHH_LLLMLLH_LLL:eANGER_NEUTRAL   0.80      0.77    -0.64     2.39 1.00    34738    13583
```

tHHL_HLHMLLH_LLL:eANGER_NEUTRAL	-0.40	0.50	-1.37	0.58	1.00	25163	15735
tHHL_HLLMLLH_LLL:eANGER_NEUTRAL	-0.18	0.61	-1.32	1.07	1.00	29769	14211
tHHL_LHHMLLH_LLL:eANGER_NEUTRAL	-0.89	0.51	-1.87	0.12	1.00	23191	14487
tHHL_LHLMLLH_LLL:eANGER_NEUTRAL	0.13	0.59	-1.01	1.34	1.00	31858	14794
tHHL_LLHMLLH_LLL:eANGER_NEUTRAL	-0.36	0.55	-1.42	0.74	1.00	26404	13396
tHHL_LLLMLLH_LLL:eANGER_NEUTRAL	0.88	0.65	-0.36	2.21	1.00	32123	13922
tHLH_HLLMLLH_LLL:eANGER_NEUTRAL	-0.42	0.54	-1.47	0.65	1.00	26227	14550
tHLH_LHHMLLH_LLL:eANGER_NEUTRAL	0.01	0.57	-1.09	1.16	1.00	30022	13177
tHLH_LHLMLLH_LLL:eANGER_NEUTRAL	-0.54	0.53	-1.56	0.53	1.00	27480	14937
tHLH_LLHMLLH_LLL:eANGER_NEUTRAL	0.78	0.78	-0.68	2.38	1.00	38515	12251
tHLH_LLLMLLH_LLL:eANGER_NEUTRAL	-0.38	0.55	-1.42	0.72	1.00	26479	14191
tHLL_LHHMLLH_LLL:eANGER_NEUTRAL	-0.03	0.59	-1.14	1.18	1.00	28485	13331
tHLL_LHLMLLH_LLL:eANGER_NEUTRAL	-0.39	0.59	-1.52	0.79	1.00	28175	14495
tHLL_LLHMLLH_LLL:eANGER_NEUTRAL	-0.18	0.61	-1.34	1.04	1.00	29985	13976
tHLL_LLLMLLH_LLL:eANGER_NEUTRAL	-0.45	0.53	-1.47	0.59	1.00	28023	15431
tLHH_LHLMLLH_LLL:eANGER_NEUTRAL	-0.72	0.65	-1.97	0.61	1.00	30800	12944
tLHH_LLHMLLH_LLL:eANGER_NEUTRAL	1.03	0.55	-0.01	2.15	1.00	27204	12755
tLHH_LLLMLLH_LLL:eANGER_NEUTRAL	-0.16	0.77	-1.60	1.40	1.00	40356	13996
tLHL_LLHMLLH_LLL:eANGER_NEUTRAL	0.07	0.60	-1.08	1.27	1.00	28716	13578
tLHL_LLLMLLH_LLL:eANGER_NEUTRAL	0.06	0.53	-0.95	1.11	1.00	27986	14959
tHHH_HHLMLLH_LLL:eANGER_PRIDE	-0.26	0.64	-1.49	1.03	1.00	33946	12311
tHHH_HLHMLLH_LLL:eANGER_PRIDE	0.28	0.70	-1.06	1.70	1.00	32400	13652
tHHH_HLLMLLH_LLL:eANGER_PRIDE	-0.61	0.70	-1.93	0.83	1.00	31156	13377
tHHH_LHHMLLH_LLL:eANGER_PRIDE	-0.37	0.51	-1.36	0.64	1.00	26431	15371
tHHH_LHLMLLH_LLL:eANGER_PRIDE	-0.32	0.77	-1.79	1.25	1.00	33795	13691
tHHH_LLHMLLH_LLL:eANGER_PRIDE	-0.09	0.56	-1.16	1.04	1.00	25973	13559
tHHH_LLLMLLH_LLL:eANGER_PRIDE	0.49	0.80	-1.02	2.13	1.00	35080	13041
tHHL_HLHMLLH_LLL:eANGER_PRIDE	-0.76	0.61	-1.94	0.47	1.00	28379	14070
tHHL_HLLMLLH_LLL:eANGER_PRIDE	0.89	0.76	-0.53	2.43	1.00	30773	13251
tHHL_LHHMLLH_LLL:eANGER_PRIDE	-0.86	0.57	-1.97	0.28	1.00	27077	14997
tHHL_LHLMLLH_LLL:eANGER_PRIDE	0.38	0.68	-0.90	1.77	1.00	32617	14357
tHHL_LLHMLLH_LLL:eANGER_PRIDE	-0.04	0.60	-1.18	1.17	1.00	29209	15055
tHHL_LLLMLLH_LLL:eANGER_PRIDE	-0.20	0.57	-1.31	0.93	1.00	27913	14152
tHLH_HLLMLLH_LLL:eANGER_PRIDE	-0.36	0.51	-1.35	0.64	1.00	22422	14847
tHLH_LHHMLLH_LLL:eANGER_PRIDE	-1.34	0.58	-2.46	-0.21	1.00	27858	15718
tHLH_LHLMLLH_LLL:eANGER_PRIDE	0.30	0.62	-0.86	1.55	1.00	29016	13367
tHLH_LLHMLLH_LLL:eANGER_PRIDE	-0.72	0.54	-1.75	0.35	1.00	25483	14449
tHLH_LLLMLLH_LLL:eANGER_PRIDE	0.38	0.70	-0.94	1.82	1.00	29364	12680
tHLL_LHHMLLH_LLL:eANGER_PRIDE	-0.72	0.59	-1.85	0.47	1.00	28695	14134
tHLL_LHLMLLH_LLL:eANGER_PRIDE	-0.23	0.67	-1.48	1.13	1.00	31806	14364
tHLL_LLHMLLH_LLL:eANGER_PRIDE	-0.62	0.55	-1.67	0.48	1.00	24453	15170
tHLL_LLLMLLH_LLL:eANGER_PRIDE	0.36	0.54	-0.66	1.44	1.00	25986	14544
tLHH_LHLMLLH_LLL:eANGER_PRIDE	-0.32	0.67	-1.58	1.05	1.00	31614	14174
tLHH_LLHMLLH_LLL:eANGER_PRIDE	0.01	0.57	-1.08	1.15	1.00	29429	12943
tLHH_LLLMLLH_LLL:eANGER_PRIDE	0.22	0.88	-1.44	1.97	1.00	35260	13360
tLHL_LLHMLLH_LLL:eANGER_PRIDE	0.34	0.61	-0.82	1.60	1.00	28349	14519
tLHL_LLLMLLH_LLL:eANGER_PRIDE	-0.10	0.56	-1.18	1.02	1.00	27653	14574
tHHH_HHLMLLH_LLL:eANGER_SHAME	0.22	0.56	-0.87	1.31	1.00	28767	14625
tHHH_HLHMLLH_LLL:eANGER_SHAME	-0.86	0.48	-1.80	0.08	1.00	23779	15550
tHHH_HLLMLLH_LLL:eANGER_SHAME	-0.34	0.59	-1.48	0.85	1.00	29576	14029
tHHH_LHHMLLH_LLL:eANGER_SHAME	0.28	0.49	-0.70	1.25	1.00	27056	15019
tHHH_LHLMLLH_LLL:eANGER_SHAME	-0.47	0.57	-1.57	0.65	1.00	29609	15697
tHHH_LLHMLLH_LLL:eANGER_SHAME	0.15	0.54	-0.89	1.23	1.00	27011	14094
tHHH_LLLMLLH_LLL:eANGER_SHAME	-0.86	0.54	-1.91	0.19	1.00	26881	15447
tHHL_HLHMLLH_LLL:eANGER_SHAME	-0.08	0.56	-1.19	1.05	1.00	28372	14574
tHHL_HLLMLLH_LLL:eANGER_SHAME	-0.61	0.50	-1.58	0.36	1.00	25572	15235
tHHL_LHHMLLH_LLL:eANGER_SHAME	-0.08	0.61	-1.27	1.13	1.00	33408	14077
tHHL_LHLMLLH_LLL:eANGER_SHAME	-0.41	0.55	-1.49	0.65	1.00	29247	14950
tHHL_LLHMLLH_LLL:eANGER_SHAME	-0.17	0.56	-1.27	0.92	1.00	27494	14472
tHHL_LLLMLLH_LLL:eANGER_SHAME	-0.86	0.56	-1.96	0.24	1.00	30126	15346
tHLH_HLLMLLH_LLL:eANGER_SHAME	1.42	0.55	0.35	2.50	1.00	28119	13927
tHLH_LHHMLLH_LLL:eANGER_SHAME	-0.50	0.50	-1.50	0.49	1.00	26174	15564
tHLH_LHLMLLH_LLL:eANGER_SHAME	0.71	0.54	-0.34	1.79	1.00	30279	13876
tHLH_LLHMLLH_LLL:eANGER_SHAME	1.18	0.63	-0.03	2.48	1.00	32696	13847
tHLH_LLLMLLH_LLL:eANGER_SHAME	0.27	0.48	-0.67	1.22	1.00	23859	15518
tHLL_LHHMLLH_LLL:eANGER_SHAME	0.40	0.64	-0.80	1.70	1.00	33222	13531
tHLL_LHLMLLH_LLL:eANGER_SHAME	-0.66	0.57	-1.77	0.48	1.00	26953	14281
tHLL_LLHMLLH_LLL:eANGER_SHAME	0.82	0.51	-0.14	1.84	1.00	26501	14443
tHLL_LLLMLLH_LLL:eANGER_SHAME	-0.56	0.56	-1.67	0.56	1.00	28861	13782
tLHH_LHLMLLH_LLL:eANGER_SHAME	0.99	0.75	-0.44	2.50	1.00	33388	13415

tLHH_LLHMLLH_LLL:eANGER_SHAME	0.77	0.54	-0.27	1.83	1.00	28821	13951
tLHH_LLLMLLH_LLL:eANGER_SHAME	-0.35	0.64	-1.58	0.95	1.00	32814	14942
tLHL_LLHMLLH_LLL:eANGER_SHAME	0.18	0.54	-0.87	1.27	1.00	26657	14682
tLHL_LLLMLLH_LLL:eANGER_SHAME	-0.59	0.52	-1.60	0.43	1.00	25956	15120
tHHH_HHLMLLH_LLL:eLOVE_NEUTRAL	0.81	0.63	-0.41	2.07	1.00	36170	13917
tHHH_HLHMLLH_LLL:eLOVE_NEUTRAL	-0.01	0.58	-1.13	1.14	1.00	30371	14428
tHHH_HLLMLLH_LLL:eLOVE_NEUTRAL	0.57	0.69	-0.75	1.99	1.00	34398	13343
tHHH_LHHMLLH_LLL:eLOVE_NEUTRAL	0.31	0.57	-0.82	1.45	1.00	30167	14294
tHHH_LHLMLLH_LLL:eLOVE_NEUTRAL	-0.72	0.68	-2.03	0.63	1.00	32927	13804
tHHH_LLHMLLH_LLL:eLOVE_NEUTRAL	-0.05	0.59	-1.19	1.12	1.00	30040	13157
tHHH_LLLMLLH_LLL:eLOVE_NEUTRAL	0.72	0.68	-0.54	2.10	1.00	30853	13843
tHHL_HLHMLLH_LLL:eLOVE_NEUTRAL	0.39	0.55	-0.66	1.51	1.00	27216	14542
tHHL_HLLMLLH_LLL:eLOVE_NEUTRAL	-0.38	0.55	-1.47	0.71	1.00	27551	14085
tHHL_LHHMLLH_LLL:eLOVE_NEUTRAL	0.33	0.60	-0.83	1.56	1.00	29264	13692
tHHL_LHLMLLH_LLL:eLOVE_NEUTRAL	0.42	0.64	-0.79	1.72	1.00	32896	13773
tHHL_LLHMLLH_LLL:eLOVE_NEUTRAL	-0.15	0.48	-1.08	0.81	1.00	24508	14982
tHHL_LLLMLLH_LLL:eLOVE_NEUTRAL	0.93	0.60	-0.20	2.13	1.00	29317	13617
tHLH_HLLMLLH_LLL:eLOVE_NEUTRAL	0.40	0.49	-0.55	1.36	1.00	24873	15552
tHLH_LHHMLLH_LLL:eLOVE_NEUTRAL	0.28	0.51	-0.69	1.30	1.00	26088	15139
tHLH_LHLMLLH_LLL:eLOVE_NEUTRAL	-0.17	0.56	-1.25	0.95	1.00	29963	13914
tHLH_LLHMLLH_LLL:eLOVE_NEUTRAL	-0.20	0.53	-1.23	0.86	1.00	25146	15109
tHLH_LLLMLLH_LLL:eLOVE_NEUTRAL	0.09	0.62	-1.10	1.30	1.00	30528	14655
tHLL_LHHMLLH_LLL:eLOVE_NEUTRAL	-0.95	0.55	-2.02	0.13	1.00	25435	15241
tHLL_LHLMLLH_LLL:eLOVE_NEUTRAL	0.21	0.57	-0.87	1.35	1.00	27072	13965
tHLL_LLHMLLH_LLL:eLOVE_NEUTRAL	-0.25	0.57	-1.35	0.88	1.00	28110	13794
tHLL_LLLMLLH_LLL:eLOVE_NEUTRAL	-0.51	0.53	-1.56	0.51	1.00	25449	15089
tLHH_LHLMLLH_LLL:eLOVE_NEUTRAL	0.06	0.66	-1.19	1.38	1.00	31380	15039
tLHH_LLHMLLH_LLL:eLOVE_NEUTRAL	0.08	0.57	-1.06	1.19	1.00	31871	14371
tLHH_LLLMLLH_LLL:eLOVE_NEUTRAL	-0.19	0.62	-1.37	1.06	1.00	29841	15222
tLHL_LLHMLLH_LLL:eLOVE_NEUTRAL	-1.04	0.54	-2.10	0.01	1.00	26504	14153
tLHL_LLLMLLH_LLL:eLOVE_NEUTRAL	0.71	0.57	-0.38	1.85	1.00	28863	14078
tHHH_HHLMLLH_LLL:eLOVE_PRIDE	0.10	0.52	-0.90	1.11	1.00	27437	14897
tHHH_HLHMLLH_LLL:eLOVE_PRIDE	-0.12	0.55	-1.19	0.98	1.00	28592	14481
tHHH_HLLMLLH_LLL:eLOVE_PRIDE	0.49	0.71	-0.83	1.94	1.00	34952	14150
tHHH_LHHMLLH_LLL:eLOVE_PRIDE	-0.07	0.49	-1.01	0.90	1.00	25266	15073
tHHH_LHLMLLH_LLL:eLOVE_PRIDE	-0.67	0.65	-1.92	0.63	1.00	30994	14402
tHHH_LLHMLLH_LLL:eLOVE_PRIDE	-1.06	0.49	-2.04	-0.12	1.00	24345	15079
tHHH_LLLMLLH_LLL:eLOVE_PRIDE	0.15	0.64	-1.09	1.44	1.00	32540	14344
tHHL_HLHMLLH_LLL:eLOVE_PRIDE	0.12	0.53	-0.88	1.16	1.00	27164	14671
tHHL_HLLMLLH_LLL:eLOVE_PRIDE	0.15	0.60	-1.01	1.35	1.00	30149	14633
tHHL_LHHMLLH_LLL:eLOVE_PRIDE	0.30	0.59	-0.83	1.48	1.00	30763	14355
tHHL_LHLMLLH_LLL:eLOVE_PRIDE	0.87	0.60	-0.26	2.06	1.00	29603	15058
tHHL_LLHMLLH_LLL:eLOVE_PRIDE	-0.09	0.53	-1.13	0.96	1.00	28248	14641
tHHL_LLLMLLH_LLL:eLOVE_PRIDE	0.99	0.65	-0.27	2.29	1.00	32374	13960
tHLH_HLLMLLH_LLL:eLOVE_PRIDE	0.03	0.49	-0.93	0.99	1.00	23828	14662
tHLH_LHHMLLH_LLL:eLOVE_PRIDE	-0.30	0.54	-1.35	0.78	1.00	26286	14274
tHLH_LHLMLLH_LLL:eLOVE_PRIDE	-0.64	0.57	-1.74	0.48	1.00	29798	14683
tHLH_LLHMLLH_LLL:eLOVE_PRIDE	-0.50	0.58	-1.63	0.69	1.00	28136	14416
tHLH_LLLMLLH_LLL:eLOVE_PRIDE	0.30	0.56	-0.77	1.43	1.00	27765	13967
tHLL_LHHMLLH_LLL:eLOVE_PRIDE	-0.48	0.65	-1.72	0.84	1.00	32726	14608
tHLL_LHLMLLH_LLL:eLOVE_PRIDE	-0.44	0.55	-1.51	0.64	1.00	27030	15412
tHLL_LLHMLLH_LLL:eLOVE_PRIDE	0.14	0.52	-0.86	1.15	1.00	27341	15319
tHLL_LLLMLLH_LLL:eLOVE_PRIDE	1.25	0.64	0.04	2.54	1.00	30272	12360
tLHH_LHLMLLH_LLL:eLOVE_PRIDE	1.18	0.72	-0.15	2.65	1.00	35834	14209
tLHH_LLHMLLH_LLL:eLOVE_PRIDE	-0.08	0.50	-1.05	0.90	1.00	24996	15730
tLHH_LLLMLLH_LLL:eLOVE_PRIDE	0.18	0.74	-1.21	1.69	1.00	34598	14197
tLHL_LLHMLLH_LLL:eLOVE_PRIDE	-0.43	0.59	-1.58	0.73	1.00	28741	14832
tLHL_LLLMLLH_LLL:eLOVE_PRIDE	-1.10	0.55	-2.18	-0.04	1.00	27643	13636
tHHH_HHLMLLH_LLL:eLOVE_SHAME	-0.75	0.60	-1.92	0.42	1.00	31786	14206
tHHH_HLHMLLH_LLL:eLOVE_SHAME	0.42	0.70	-0.90	1.87	1.00	34074	13459
tHHH_HLLMLLH_LLL:eLOVE_SHAME	0.71	0.80	-0.79	2.32	1.00	37637	12975
tHHH_LHHMLLH_LLL:eLOVE_SHAME	-0.49	0.53	-1.54	0.54	1.00	28703	15066
tHHH_LHLMLLH_LLL:eLOVE_SHAME	0.66	0.80	-0.84	2.29	1.00	36247	13647
tHHH_LLHMLLH_LLL:eLOVE_SHAME	0.09	0.61	-1.08	1.30	1.00	32948	13753
tHHH_LLLMLLH_LLL:eLOVE_SHAME	0.13	0.73	-1.24	1.60	1.00	34128	13676
tHHL_HLHMLLH_LLL:eLOVE_SHAME	0.34	0.64	-0.89	1.63	1.00	31728	13936
tHHL_HLLMLLH_LLL:eLOVE_SHAME	0.13	0.64	-1.06	1.44	1.00	31992	12895
tHHL_LHHMLLH_LLL:eLOVE_SHAME	0.09	0.66	-1.15	1.42	1.00	30968	13999
tHHL_LHLMLLH_LLL:eLOVE_SHAME	-0.13	0.62	-1.30	1.10	1.00	30893	13739
tHHL_LLHMLLH_LLL:eLOVE_SHAME	0.01	0.60	-1.15	1.24	1.00	30673	14466

tHHL_LLLMLLH_LLL:eLOVE_SHAME	-0.22	0.57	-1.32	0.94	1.00	30307	13101
tHLH_HLLMLLH_LLL:eLOVE_SHAME	-0.61	0.46	-1.52	0.28	1.00	23438	14723
tHLH_LHHMLLH_LLL:eLOVE_SHAME	-0.10	0.51	-1.09	0.91	1.00	26580	14061
tHLH_LHLMLLH_LLL:eLOVE_SHAME	0.90	0.65	-0.30	2.24	1.00	33591	13591
tHLH_LLHMLLH_LLL:eLOVE_SHAME	-0.27	0.62	-1.45	0.97	1.00	32644	13249
tHLH_LLLMLLH_LLL:eLOVE_SHAME	-1.61	0.52	-2.63	-0.61	1.00	27214	14866
tHLL_LHHMLLH_LLL:eLOVE_SHAME	1.01	0.75	-0.38	2.54	1.00	36090	12624
tHLL_LHLMLLH_LLL:eLOVE_SHAME	0.21	0.64	-1.00	1.52	1.00	33451	13857
tHLL_LLHMLLH_LLL:eLOVE_SHAME	0.42	0.63	-0.78	1.71	1.00	29549	13440
tHLL_LLLMLLH_LLL:eLOVE_SHAME	-0.94	0.48	-1.87	-0.00	1.00	24093	14855
tLHH_LHLMLLH_LLL:eLOVE_SHAME	0.07	0.74	-1.31	1.56	1.00	35356	13324
tLHH_LLHMLLH_LLL:eLOVE_SHAME	0.09	0.50	-0.88	1.09	1.00	24553	14817
tLHH_LLLMLLH_LLL:eLOVE_SHAME	-0.57	0.70	-1.90	0.84	1.00	34105	13329
tLHL_LLHMLLH_LLL:eLOVE_SHAME	0.32	0.58	-0.78	1.47	1.00	27964	15027
tLHL_LLLMLLH_LLL:eLOVE_SHAME	0.51	0.71	-0.83	1.96	1.00	36549	13509
tHHH_HHLMLLH_LLL:eNEUTRAL_PRIDE	0.48	0.51	-0.49	1.50	1.00	27818	15519
tHHH_HLHMLLH_LLL:eNEUTRAL_PRIDE	0.12	0.57	-0.97	1.24	1.00	26027	14981
tHHH_HLLMLLH_LLL:eNEUTRAL_PRIDE	-0.48	0.61	-1.66	0.72	1.00	29481	14729
tHHH_LHHMLLH_LLL:eNEUTRAL_PRIDE	-0.53	0.48	-1.49	0.39	1.00	25856	15101
tHHH_LHLMLLH_LLL:eNEUTRAL_PRIDE	0.78	0.79	-0.70	2.36	1.00	35859	12952
tHHH_LLHMLLH_LLL:eNEUTRAL_PRIDE	0.17	0.53	-0.84	1.23	1.00	26627	14549
tHHH_LLLMLLH_LLL:eNEUTRAL_PRIDE	0.37	0.72	-1.00	1.82	1.00	35347	14085
tHHL_HLHMLLH_LLL:eNEUTRAL_PRIDE	0.07	0.55	-0.99	1.16	1.00	28949	14688
tHHL_HLLMLLH_LLL:eNEUTRAL_PRIDE	0.10	0.60	-1.03	1.32	1.00	30188	13973
tHHL_LHHMLLH_LLL:eNEUTRAL_PRIDE	0.08	0.52	-0.93	1.10	1.00	25815	15420
tHHL_LHLMLLH_LLL:eNEUTRAL_PRIDE	0.70	0.57	-0.39	1.86	1.00	28596	14318
tHHL_LLHMLLH_LLL:eNEUTRAL_PRIDE	0.07	0.57	-1.03	1.20	1.00	27933	14570
tHHL_LLLMLLH_LLL:eNEUTRAL_PRIDE	1.13	0.64	-0.06	2.43	1.00	31657	12768
tHLH_HLLMLLH_LLL:eNEUTRAL_PRIDE	-0.63	0.71	-2.08	0.69	1.00	34858	13305
tHLH_LHHMLLH_LLL:eNEUTRAL_PRIDE	1.31	0.57	0.24	2.45	1.00	28620	14201
tHLH_LHLMLLH_LLL:eNEUTRAL_PRIDE	-0.05	0.63	-1.25	1.20	1.00	32879	13845
tHLH_LLHMLLH_LLL:eNEUTRAL_PRIDE	-0.19	0.51	-1.17	0.82	1.00	26939	15089
tHLH_LLLMLLH_LLL:eNEUTRAL_PRIDE	0.52	0.63	-0.68	1.81	1.00	34032	14255
tHLL_LHHMLLH_LLL:eNEUTRAL_PRIDE	-1.48	0.51	-2.49	-0.49	1.00	25041	15535
tHLL_LHLMLLH_LLL:eNEUTRAL_PRIDE	0.73	0.70	-0.60	2.18	1.00	33335	13968
tHLL_LLHMLLH_LLL:eNEUTRAL_PRIDE	0.08	0.54	-0.97	1.16	1.00	27517	13968
tHLL_LLLMLLH_LLL:eNEUTRAL_PRIDE	0.11	0.51	-0.88	1.12	1.00	26256	14163
tLHH_LHLMLLH_LLL:eNEUTRAL_PRIDE	-0.78	0.58	-1.91	0.39	1.00	29892	14007
tLHH_LLHMLLH_LLL:eNEUTRAL_PRIDE	-0.65	0.53	-1.71	0.38	1.00	28424	13741
tLHH_LLLMLLH_LLL:eNEUTRAL_PRIDE	0.89	0.76	-0.53	2.43	1.00	35997	12801
tLHL_LLHMLLH_LLL:eNEUTRAL_PRIDE	-1.07	0.53	-2.11	-0.05	1.00	25175	15109
tLHL_LLLMLLH_LLL:eNEUTRAL_PRIDE	-1.02	0.52	-2.04	-0.02	1.00	26578	14349
tHHH_HHLMLLH_LLL:eNEUTRAL_SHAME	0.41	0.65	-0.81	1.71	1.00	31670	12580
tHHH_HLHMLLH_LLL:eNEUTRAL_SHAME	-0.02	0.53	-1.03	1.02	1.00	26175	15121
tHHH_HLLMLLH_LLL:eNEUTRAL_SHAME	0.11	0.75	-1.31	1.64	1.00	34337	14324
tHHH_LHHMLLH_LLL:eNEUTRAL_SHAME	-0.22	0.58	-1.34	0.94	1.00	31408	13703
tHHH_LHLMLLH_LLL:eNEUTRAL_SHAME	0.93	0.79	-0.56	2.56	1.00	36527	12390
tHHH_LLHMLLH_LLL:eNEUTRAL_SHAME	-0.08	0.57	-1.19	1.03	1.00	27928	14612
tHHH_LLLMLLH_LLL:eNEUTRAL_SHAME	0.22	0.74	-1.14	1.73	1.00	32607	13521
tHHL_HLHMLLH_LLL:eNEUTRAL_SHAME	-0.21	0.54	-1.25	0.88	1.00	28070	15311
tHHL_HLLMLLH_LLL:eNEUTRAL_SHAME	1.01	0.77	-0.43	2.60	1.00	38318	12797
tHHL_LHHMLLH_LLL:eNEUTRAL_SHAME	0.23	0.56	-0.84	1.35	1.00	25876	13333
tHHL_LHLMLLH_LLL:eNEUTRAL_SHAME	1.23	0.74	-0.16	2.72	1.00	33116	12142
tHHL_LLHMLLH_LLL:eNEUTRAL_SHAME	0.27	0.66	-1.00	1.62	1.00	33178	14158
tHHL_LLLMLLH_LLL:eNEUTRAL_SHAME	1.30	0.73	-0.04	2.76	1.00	36977	13745
tHLH_HLLMLLH_LLL:eNEUTRAL_SHAME	0.13	0.51	-0.87	1.14	1.00	24888	14546
tHLH_LHHMLLH_LLL:eNEUTRAL_SHAME	0.37	0.64	-0.84	1.65	1.00	33178	13066
tHLH_LHLMLLH_LLL:eNEUTRAL_SHAME	0.02	0.56	-1.05	1.14	1.00	28430	14044
tHLH_LLHMLLH_LLL:eNEUTRAL_SHAME	-0.94	0.56	-2.02	0.17	1.00	26844	15579
tHLH_LLLMLLH_LLL:eNEUTRAL_SHAME	-0.52	0.53	-1.53	0.52	1.00	26582	15850
tHLL_LHHMLLH_LLL:eNEUTRAL_SHAME	0.98	0.78	-0.47	2.57	1.00	36430	12665
tHLL_LHLMLLH_LLL:eNEUTRAL_SHAME	-0.04	0.57	-1.13	1.11	1.00	27246	14715
tHLL_LLHMLLH_LLL:eNEUTRAL_SHAME	-0.20	0.52	-1.20	0.83	1.00	26404	14006
tHLL_LLLMLLH_LLL:eNEUTRAL_SHAME	-0.95	0.55	-2.02	0.12	1.00	26380	14091
tLHH_LHLMLLH_LLL:eNEUTRAL_SHAME	0.72	0.81	-0.79	2.38	1.00	35071	12779
tLHH_LLHMLLH_LLL:eNEUTRAL_SHAME	0.75	0.57	-0.34	1.89	1.00	29905	13689
tLHH_LLLMLLH_LLL:eNEUTRAL_SHAME	0.61	0.84	-1.00	2.33	1.00	37507	12997
tLHL_LLHMLLH_LLL:eNEUTRAL_SHAME	-1.42	0.49	-2.37	-0.45	1.00	24168	15939
tLHL_LLLMLLH_LLL:eNEUTRAL_SHAME	0.21	0.52	-0.80	1.26	1.00	27012	14798
tHHH_HHLMLLH_LLL:ePRIDE_SHAME	-0.14	0.55	-1.21	0.96	1.00	30095	14205

tHHH_HLHMLLH_LLL:ePRIDE_SHAME	0.11	0.65	-1.11	1.43	1.00	33113	14052
tHHH_HLLMLLH_LLL:ePRIDE_SHAME	0.67	0.81	-0.83	2.33	1.00	34173	12471
tHHH_LHHMLLH_LLL:ePRIDE_SHAME	-1.25	0.52	-2.29	-0.23	1.00	29648	15027
tHHH_LHLMLLH_LLL:ePRIDE_SHAME	0.73	0.80	-0.78	2.37	1.00	34359	12843
tHHH_LLHMLLH_LLL:ePRIDE_SHAME	0.05	0.66	-1.18	1.40	1.00	30692	12459
tHHH_LLLMLLH_LLL:ePRIDE_SHAME	0.76	0.80	-0.75	2.41	1.00	37251	12325
tHHL_HLHMLLH_LLL:ePRIDE_SHAME	0.05	0.56	-1.03	1.16	1.00	29608	14525
tHHL_HLLMLLH_LLL:ePRIDE_SHAME	-0.35	0.63	-1.56	0.92	1.00	30707	13178
tHHL_LHHMLLH_LLL:ePRIDE_SHAME	0.01	0.57	-1.09	1.15	1.00	27343	14409
tHHL_LHLMLLH_LLL:ePRIDE_SHAME	-0.22	0.62	-1.40	1.07	1.00	31890	13499
tHHL_LLHMLLH_LLL:ePRIDE_SHAME	0.76	0.67	-0.49	2.15	1.00	32249	12544
tHHL_LLLMLLH_LLL:ePRIDE_SHAME	0.64	0.83	-0.93	2.31	1.00	38134	13131
tHLH_HLLMLLH_LLL:ePRIDE_SHAME	0.61	0.51	-0.39	1.63	1.00	23972	13171
tHLH_LHHMLLH_LLL:ePRIDE_SHAME	0.56	0.70	-0.76	1.98	1.00	36475	13620
tHLH_LHLMLLH_LLL:ePRIDE_SHAME	0.48	0.61	-0.66	1.72	1.00	32122	13344
tHLH_LLHMLLH_LLL:ePRIDE_SHAME	0.20	0.66	-1.03	1.56	1.00	32536	14503
tHLH_LLLMLLH_LLL:ePRIDE_SHAME	0.10	0.77	-1.34	1.66	1.00	34640	12042
tHLL_LHHMLLH_LLL:ePRIDE_SHAME	0.12	0.77	-1.32	1.69	1.00	35424	12583
tHLL_LHLMLLH_LLL:ePRIDE_SHAME	-0.25	0.63	-1.47	1.00	1.00	30185	13873
tHLL_LLHMLLH_LLL:ePRIDE_SHAME	-0.60	0.64	-1.81	0.67	1.00	31035	13484
tHLL_LLLMLLH_LLL:ePRIDE_SHAME	-0.88	0.52	-1.89	0.16	1.00	26001	15485
tLHH_LHLMLLH_LLL:ePRIDE_SHAME	-0.31	0.68	-1.60	1.08	1.00	33691	14029
tLHH_LLHMLLH_LLL:ePRIDE_SHAME	0.08	0.48	-0.85	1.03	1.00	24833	14572
tLHH_LLLMLLH_LLL:ePRIDE_SHAME	0.05	0.74	-1.33	1.55	1.00	37129	13316
tLHL_LLHMLLH_LLL:ePRIDE_SHAME	1.04	0.76	-0.39	2.57	1.00	32776	12781
tLHL_LLLMLLH_LLL:ePRIDE_SHAME	-0.56	0.47	-1.48	0.38	1.00	21660	15604

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).

Appendix A (Ch. 4): Written materials (Rating task)

In the Rating task, before giving a rating based on the Target sentence, participants see a Preceding context, which varies based on tune-emotion combination, as shown in the table below.

Tune	Emtn	Preceding context	Target sentence†
HHH	Anger	According to my accountant, your landlord might charge you extra fees.	For Gallagher?
HHL	Anger	You asked the tenant's union about the repairs, about the taxes…	About Gallagher…
HLH	Anger	Who is responsible for all these unpaid power bills?	That's Gallagher…
HLL	Anger	Who are these tax documents for?	For Gallagher.
LHH	Anger	Apparently your brother was out drinking with your landlord.	With Gallagher?
LHL	Anger	Who's been giving you so many problems again?	That's Gallagher…
LLH	Anger	My neighbor was gifted an expensive bottle of wine from your landlord.	From Gallagher?
LLL	Anger	There's a pile of boxes on your porch, are they for you?	For Gallagher.
HHH	Love	I received a phone call from your favorite granddaughter on my birthday.	From Madelyn?
HHL	Love	We have so many talented cooks in our family. There's your mother, …	Also Madelyn…
HLH	Love	I heard your sister will join you on vacation this year, is that true?	And Madelyn…
HLL	Love	Who is coming over for dinner later?	My Madelyn.
LHH	Love	A reporter called to ask about your favorite granddaughter.	About Madelyn?
LHL	Love	Why do we need to be in town this weekend again?	For Madelyn…
LLH	Love	I cleaned your stovetop so that you can cook with Lavender.	With Madelyn?
LLL	Love	Did anyone send you flowers on your birthday?	Just Madelyn.

†Participants saw the sentences punctuated as shown, which provided additional cues to the attested intonational meanings for the tunes.

Appendix B (Ch. 4): Statistical model summary: TOM (GLMM, Sorting task)

```
TOM: 26247.455 sec elapsed
 Family: bernoulli
  Links: mu = logit
Formula: grouped ~ tune_pair + (1 | id)
   Data: data (Number of observations: 412560)
  Draws: 2 chains, each with iter = 10000; warmup = 1000; thin = 1;
         total post-warmup draws = 18000

Multilevel Hyperparameters:
~id (Number of levels: 105)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     1.50      0.11     1.30     1.73 1.00      570     1010

Regression Coefficients:
                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept          -2.12      0.16    -2.42    -1.81 1.00      249      475
tune_pairLLL_LLL   -0.07      0.05    -0.16     0.03 1.00     1950     5031
tune_pairHLL_LLL    0.23      0.05     0.13     0.32 1.00     1721     4383
tune_pairLLH_LLL   -0.38      0.05    -0.48    -0.28 1.00     1988     4664
tune_pairHLH_LLL   -0.02      0.05    -0.12     0.07 1.00     1906     4369
tune_pairLHL_LLL    0.07      0.05    -0.02     0.16 1.00     1840     4580
tune_pairLHH_LLL   -0.77      0.06    -0.88    -0.66 1.00     2333     5254
tune_pairHHH_LLL   -0.65      0.05    -0.76    -0.54 1.00     2172     5547
tune_pairHHL_LLL   -0.42      0.05    -0.52    -0.32 1.00     2107     4042
tune_pairHLL_HLL   -0.02      0.05    -0.11     0.07 1.00     1930     3911
tune_pairHLL_LLH   -0.26      0.05    -0.35    -0.16 1.00     1958     4644
tune_pairHLH_HLL    0.25      0.05     0.16     0.34 1.00     1768     4438
tune_pairHLL_LHL    0.07      0.05    -0.02     0.17 1.00     1921     4733
tune_pairHLL_LHH   -0.52      0.05    -0.63    -0.42 1.00     2115     4508
tune_pairHHH_HLL   -0.47      0.05    -0.57    -0.37 1.00     2188     4894
tune_pairHHL_HLL   -0.30      0.05    -0.40    -0.20 1.00     2084     5276
tune_pairLLH_LLH    0.02      0.05    -0.07     0.11 1.00     1830     4597
tune_pairHLH_LLH   -0.30      0.05    -0.40    -0.20 1.00     1983     4873
tune_pairLHL_LLH   -0.23      0.05    -0.33    -0.14 1.00     1944     4752
tune_pairLHH_LLH    0.31      0.05     0.22     0.40 1.00     1738     4048
tune_pairHHH_LLH    0.30      0.05     0.21     0.39 1.00     1772     4127
tune_pairHHL_LLH    0.06      0.05    -0.04     0.15 1.00     1847     4495
tune_pairHLH_HLH   -0.20      0.05    -0.30    -0.10 1.00     1981     5194
tune_pairHLH_LHL   -0.11      0.05    -0.21    -0.01 1.00     1960     4626
tune_pairHLH_LHH   -0.66      0.05    -0.77    -0.55 1.00     2264     5382
tune_pairHHH_HLH   -0.54      0.05    -0.64    -0.43 1.00     2107     5101
tune_pairHHL_HLH   -0.34      0.05    -0.44    -0.24 1.00     1929     4540
tune_pairLHL_LHL   -0.06      0.05    -0.16     0.03 1.00     1933     4685
tune_pairLHH_LHL   -0.60      0.05    -0.70    -0.50 1.00     2297     5683
tune_pairHHH_LHL   -0.56      0.05    -0.66    -0.46 1.00     2213     5713
tune_pairHHL_LHL   -0.32      0.05    -0.41    -0.22 1.00     1992     4811
tune_pairLHH_LHH   -0.10      0.05    -0.20    -0.00 1.00     1882     4453
tune_pairHHH_LHH    0.33      0.05     0.24     0.42 1.00     1661     3816
tune_pairHHL_LHH   -0.09      0.05    -0.18     0.01 1.00     1913     4257
tune_pairHHH_HHH    0.02      0.05    -0.07     0.11 1.00     1918     4528
tune_pairHHL_HHL   -0.12      0.05    -0.22    -0.02 1.00     1902     5117

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Appendix C (Ch. 4): Statistical model summary: TEM (GLMM, Sorting task)

```
TEM: 19196.752 sec elapsed
 Family: bernoulli
  Links: mu = logit
Formula: grouped ~ tune_pair + emtn_pair + (1 | id)
   Data: data (Number of observations: 412560)
  Draws: 2 chains, each with iter = 10000; warmup = 1000; thin = 1;
         total post-warmup draws = 18000

Multilevel Hyperparameters:
~id (Number of levels: 105)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     1.50      0.10     1.31     1.72 1.00      375      627

Regression Coefficients:
                    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept              -2.19      0.16    -2.50    -1.85 1.01      101      339
tune_pairLLL_LLL       -0.07      0.05    -0.16     0.03 1.00     1162     3126
tune_pairHLL_LLL        0.23      0.05     0.14     0.32 1.00     1096     2882
tune_pairLLH_LLL       -0.38      0.05    -0.48    -0.27 1.00     1295     2731
tune_pairHLH_LLL       -0.02      0.05    -0.11     0.08 1.00     1194     2775
tune_pairLHL_LLL        0.07      0.05    -0.02     0.17 1.00     1090     2704
tune_pairLHH_LLL       -0.77      0.06    -0.88    -0.66 1.00     1494     3673
tune_pairHHH_LLL       -0.65      0.05    -0.75    -0.54 1.00     1352     3561
tune_pairHHL_LLL       -0.41      0.05    -0.52    -0.31 1.00     1269     3298
tune_pairHLL_HLL       -0.02      0.05    -0.11     0.08 1.00     1149     2981
tune_pairHLL_LLH       -0.25      0.05    -0.35    -0.15 1.00     1315     2929
tune_pairHLH_HLL        0.25      0.05     0.16     0.34 1.00     1073     2529
tune_pairHLL_LHL        0.08      0.05    -0.02     0.17 1.00     1128     2822
tune_pairHLL_LHH       -0.52      0.05    -0.62    -0.41 1.00     1369     3957
tune_pairHHH_HLL       -0.47      0.05    -0.57    -0.36 1.00     1392     3291
tune_pairHHL_HLL       -0.30      0.05    -0.40    -0.19 1.00     1237     2659
tune_pairLLH_LLH        0.02      0.05    -0.07     0.12 1.00     1173     2727
tune_pairHLH_LLH       -0.29      0.05    -0.39    -0.19 1.00     1269     3978
tune_pairLHL_LLH       -0.23      0.05    -0.33    -0.13 1.00     1274     2994
tune_pairLHH_LLH        0.32      0.05     0.23     0.41 1.00     1131     2599
tune_pairHHH_LLH        0.30      0.05     0.21     0.40 1.00     1063     2863
tune_pairHHL_LLH        0.06      0.05    -0.03     0.16 1.00     1083     2792
tune_pairHLH_HLH       -0.20      0.05    -0.30    -0.09 1.00     1308     3269
tune_pairHLH_LHL       -0.11      0.05    -0.20    -0.01 1.00     1234     2839
tune_pairHLH_LHH       -0.66      0.06    -0.76    -0.55 1.00     1408     3522
tune_pairHHH_HLH       -0.53      0.05    -0.64    -0.43 1.00     1359     3342
tune_pairHHL_HLH       -0.34      0.05    -0.44    -0.24 1.00     1286     3122
tune_pairLHL_LHL       -0.06      0.05    -0.16     0.04 1.00     1202     3197
tune_pairLHH_LHL       -0.60      0.05    -0.70    -0.49 1.00     1376     3181
tune_pairHHH_LHL       -0.56      0.05    -0.66    -0.45 1.00     1308     3678
tune_pairHHL_LHL       -0.31      0.05    -0.41    -0.21 1.00     1247     3024
tune_pairLHH_LHH       -0.09      0.05    -0.19     0.00 1.00     1314     3420
tune_pairHHH_LHH        0.33      0.05     0.24     0.43 1.00     1094     3055
tune_pairHHL_LHH       -0.08      0.05    -0.18     0.02 1.00     1118     2829
tune_pairHHH_HHH        0.03      0.05    -0.07     0.12 1.00     1150     2383
```

tune_pairHHL_HHL	-0.12	0.05	-0.21	-0.02	1.00	1220	3771
emtn_pairANGER_ANGER	0.11	0.03	0.06	0.16	1.00	3540	7472
emtn_pairANGER_SHAME	0.07	0.03	0.02	0.13	1.00	3259	8000
emtn_pairANGER_LOVE	-0.08	0.03	-0.13	-0.02	1.00	3568	7111
emtn_pairANGER_PRIDE	-0.07	0.03	-0.12	-0.01	1.00	3531	6707
emtn_pairSHAME_SHAME	0.13	0.03	0.08	0.18	1.00	3381	7371
emtn_pairPRIDE_SHAME	0.00	0.03	-0.05	0.06	1.00	3571	7469
emtn_pairLOVE_LOVE	0.17	0.03	0.11	0.22	1.00	3283	6642
emtn_pairLOVE_PRIDE	0.14	0.03	0.09	0.19	1.00	3350	6996
emtn_pairPRIDE_PRIDE	0.19	0.03	0.14	0.24	1.00	3290	6245

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).

Appendix D (Ch. 4): CLMM ANOVA results: TOM vs TEM vs TIM (Rating task)

Tune	Meaning	Test	Resid. df	-2logLik	LR stat.	Pr(Chi)	p<05	Pr(Chi)_
HHH	commit	TEM vs TIM	575	1750.05972	51.0423978	9.04E-13	TRUE	<.001
HHH	commit	TOM vs TEM	576	1801.10212	0.01887063	0.89073802	FALSE	0.891
HHH	floorhold	TEM vs TIM	575	1789.12607	0.17896265	0.67226624	FALSE	0.672
HHH	floorhold	TOM vs TEM	576	1789.30503	6.13673994	0.01324012	TRUE	0.013
HHH	question	TEM vs TIM	575	1384.31527	0.17550581	0.67526427	FALSE	0.675
HHH	question	TOM vs TEM	576	1384.49078	0.00113684	0.97310282	FALSE	0.973
HHL	commit	TEM vs TIM	574	1756.2	9.96505639	0.00159539	TRUE	0.002
HHL	commit	TOM vs TEM	575	1766.16506	0.00551402	0.94080633	FALSE	0.941
HHL	floorhold	TEM vs TIM	577	1685.82155	13.0254649	3.07E-04	TRUE	<.001
HHL	floorhold	TOM vs TEM	578	1698.84702	2.01205082	0.15605424	FALSE	0.156
HHL	question	TEM vs TIM	577	1683.26505	119.547164	0	TRUE	<.001
HHL	question	TOM vs TEM	578	1802.81222	6.70035016	0.0096394	TRUE	0.01
HLH	commit	TEM vs TIM	577	1457.75478	21.8786319	2.90E-06	TRUE	<.001
HLH	commit	TOM vs TEM	578	1479.63341	0.4528247	0.50099659	FALSE	0.501
HLH	floorhold	TEM vs TIM	577	1751.28729	0.2828225	0.59485719	FALSE	0.595
HLH	floorhold	TOM vs TEM	578	1751.57011	2.08659663	0.14859679	FALSE	0.149
HLH	question	TEM vs TIM	574	1481.88122	1.47402922	0.22471105	FALSE	0.225
HLH	question	TOM vs TEM	575	1483.35525	0.35649964	0.5504567	FALSE	0.55

HLL	commit	TEM vs TIM	576	1215.30571	0.60714493	0.43586482	FALSE	0.436
HLL	commit	TOM vs TEM	577	1215.91285	0.77308807	0.37926343	FALSE	0.379
HLL	floorhold	TEM vs TIM	576	1616.32936	1.78504639	0.18153088	FALSE	0.182
HLL	floorhold	TOM vs TEM	577	1618.1144	4.7205255	0.02980463	TRUE	0.03
HLL	question	TEM vs TIM	573	1224.51019	0.09960414	0.75230521	FALSE	0.752
HLL	question	TOM vs TEM	574	1224.60979	1.25552454	0.26249994	FALSE	0.262
LHH	commit	TEM vs TIM	573	1715.33278	97.8797149	0	TRUE	<.001
LHH	commit	TOM vs TEM	574	1813.21249	0.53332318	0.46521307	FALSE	0.465
LHH	floorhold	TEM vs TIM	578	1764.6684	4.78897315	0.0286425	TRUE	0.029
LHH	floorhold	TOM vs TEM	579	1769.45737	1.92779566	0.16499981	FALSE	0.165
LHH	question	TEM vs TIM	574	1266.9348	8.14729378	0.00431255	TRUE	0.004
LHH	question	TOM vs TEM	575	1275.08209	1.85838172	0.17281169	FALSE	0.173
LHL	commit	TEM vs TIM	575	1503.03209	8.32940969	0.00390083	TRUE	0.004
LHL	commit	TOM vs TEM	576	1511.3615	1.19527234	0.27426864	FALSE	0.274
LHL	floorhold	TEM vs TIM	577	1725.07078	6.49998173	0.01078756	TRUE	0.011
LHL	floorhold	TOM vs TEM	578	1731.57076	0.05292799	0.81804449	FALSE	0.818
LHL	question	TEM vs TIM	573	1364.92259	2.69925168	0.10039536	FALSE	0.1
LHL	question	TOM vs TEM	574	1367.62185	0.07335267	0.78651642	FALSE	0.787
LLH	commit	TEM vs TIM	576	1797.22313	32.9378306	9.52E-09	TRUE	<.001
LLH	commit	TOM vs TEM	577	1830.16096	0.00210493	0.96340629	FALSE	0.963
LLH	floorhold	TEM vs TIM	578	1793.04411	2.25678269	0.13303018	FALSE	0.133

LLH	floorhold	TOM vs TEM	579	1795.30089	1.46018391	0.22690086	FALSE	0.227
LLH	question	TEM vs TIM	571	1461.95452	1.42874513	0.23196934	FALSE	0.232
LLH	question	TOM vs TEM	572	1463.38327	4.28E-07	0.99947789	FALSE	0.999
LLL	commit	TEM vs TIM	573	1349.31997	6.48735844	0.01086443	TRUE	0.011
LLL	commit	TOM vs TEM	574	1355.80733	0.22974966	0.63170951	FALSE	0.632
LLL	floorhold	TEM vs TIM	576	1609.943	4.86199028	0.02745465	TRUE	0.027
LLL	floorhold	TOM vs TEM	577	1614.80499	4.99958295	0.02535343	TRUE	0.025
LLL	question	TEM vs TIM	570	1264.13081	11.2763267	7.85E-04	TRUE	0.001
LLL	question	TOM vs TEM	571	1275.40714	0.42660885	0.51365764	FALSE	0.514

[Left intentionally blank]