Computational and behavioural approaches to understanding perception of speech variability

Bing'er Jiang

Department of Linguistics

McGill University, Montréal

June 2021

A thesis submitted to McGill University in partial fulfillment for the requirements of the degree of Doctor of Philosophy

Abstract

This dissertation focuses on the broad question of how humans are able to make sense of speech and interpret it as meaningful units, despite extensive variation – one instance of humans' remarkable ability to perceive cognitive units (speech sounds) from noisy continuous data. This dissertation addresses this question by examining different levels of human speech processing, from low-level phonetics to higher-level abstract patterning: listeners' variable use of acoustic cues in different linguistic contexts (Project 1), the perceptual representation integrating all acoustic dimensions for a phonological contrast (Project 2), and the linguistic knowledge used for processing phonological changes (Project 3).

Two chapters of this dissertation focus on a similar question – how listeners perceive tone and tonal register contrasts in Chinese languages – using perceptual experiments and computational modelling. While phonological contrasts typically correlate with more than one acoustic cue, questions remains about how listeners weight and integrate multiple cues for making the contrast. The first project investigates how multiple acoustic cues contribute to multi-dimensional phonological contrasts and how dialectal experience shapes listeners' perceptual strategies. The central question is: how do listeners differ in their use of acoustic cues? This project focuses on three cues in the tonal register contrast in two Chinese Wu dialects: pitch height, voice quality, and pitch contour. The findings reveal that listeners differ mainly in their overall cue acuity (e.g. there are listeners with flatter and steeper boundaries between sounds – across all cues). Moreover, for certain contrasts signaled without a dominant cue, individuals further differ in their choice of the primary cue. Finally, listeners' use of cues is affected by their dialect background. For a cue less important in their native dialect, listeners do not make better use of it even when the cue becomes more salient in the same contrast (e.g. in a different dialect).

The second project investigates a similar question to the first project, using computational modelling. The goals are to study the low-dimensional representation of tones in Mandarin Chinese continuous speech, and how different acoustic correlates map onto this representation. Adopting a data-driven method using raw speech, this project explores the representation of tones by examining a low-dimensional layer learnt in a deep-neural-network tone classification model. The model can be

seen as an 'ideal listener' doing the same task as human listeners. Unlike the human brain which can only be indirectly probed through responses, the computational model provides a learnt representation one can directly examine. The analysis of the representation reveals that while the input is high-dimensional (feature vetors encoding raw speech), two dimensions are enough to represent the tonal contrast. The two dimensions largely encode average pitch height and pitch contour, which converges with previous findings from the perception literature, and calls into question the conventional tonal notation which uses onset and offset as tones' pitch targets.

The third project investigates the role of phonological knowledge in speech perception. For predictable changes caused by phonological assimilation (English place assimilation and French voicing assimilation), native listeners are able to detect such changes and recover the original sounds, without taking them as mispronunciations. This project investigates what knowledge is minimally required for the language-specific perceptual effect. Standard automatic speech recognition systems trained on English and French are used to represent 'ideal listeners'. Each language has 13 models with different complexities to represent listeners with different scopes of linguistic knowledge. The models then perform the same task as humans did in a previous study (Darcy et al., 2009). From comparing the model and human results, the successful human-like models employ contextually sensitive acoustic knowledge and phonotactics, but do not require higher-level knowledge of a lexicon or word boundaries.

To summarize, this dissertation investigates different aspects of perception, building on evidence from diverse languages. The combination of perceptual experiments and computational modelling mutually benefit each other: the perceptual experiments examine how listeners vary and provide empirical data from human listeners, while computational modelling of 'ideal listeners' offers potential explanations for human speech perception.

Résumé

Cette thèse vise éclaircir comment les humains interprètent la langue parlée malgré la variation importante qui s'y trouve, démontrant la capacité humaine de cerner les unités cognitives (e.g. phonèmes) à partir de réalisations phonétiques hautement variables. La question est abordée dans trois niveaux de traitement de la parole : la sensibilité au contexte linguistique de l'emploi d'indices acoustiques (étude 1), la représentation en perception des maintes dimensions acoustiques d'un contraste phonologique (étude 2) et les connaissances linguistiques exigées pour traiter les changements phonologiques (étude 3).

Le premier projet s'attaque à la manière que les auditeurs emploient plusieurs indices acoustiques associés à un même contraste phonologique et à l'effet de l'expérience dialectale dans la catégorisation en perception, tenant compte de la variation individuelle. Le contraste du registre tonal dans deux variétés du chinois Wu se manifeste par la hauteur et le contour de la fréquence fondamentale en plus de la qualité vocale. Les résultats révèlent que l'acuité des auditeurs aux indices acoustiques; il existe des auditeurs utilisant fortement et faiblement chaque indice acoustique. De plus, les auditeurs privilégient des indices acoustiques différents lorsqu'un contraste n'a aucun indice dominant. Enfin, les auditeurs n'augmentent pas leur sensibilité à un indice acoustique saillant lorsque l'indice a peu d'importance dans leur variété maternelle.

Le deuxième projet reprend la question des indices acoustiques associés aux tons (cette fois du mandarin), mais par biais de la modélisation informatique. Le projet vise déterminer (a) la représentation des tons en parole continue brute et (b) la relation entre cette représentation et les indices acoustiques. Les réseaux neuronaux profonds sont employés pour générer un modèle de classification des tons, duquel une couche de dimension basse est analysée comme équivalent de ce qu'un humain pourrait créer comme représentation des tons. L'analyse du modèle informatique révèle que deux dimensions de catégorisation suffisent pour classer les tons même si la variation comprend plusieurs dimensions. De plus, le modèle représente la fréquence fondamentale par sa hauteur et son contour, un résultat qui reflète la perception humaine et remet en question la notation tonale conventionnelle qui décrit plutôt les hauteurs en début et en fin de cible tonale.

Le troisième projet étudie le rôle des connaissances phonologiques dans la perception de la parole, se penchant deux car d'assimilation phonologique prévisible (l'assimilation de lieu en anglais et l'assimilation de voisement en français). Les auditeurs natifs peuvent détecter les changements et conséquemment récupérer les sons originaux, sans les percevoir comme des erreurs de prononciation. La question qui en découle est ce que sont les connaissances minimales requises pour obtenir une telle adaptation perceptuelle. Des systèmes de reconnaissance vocale automatique sont formés sur des données provenant soit de l'anglais, soit du français, avec treize modèles différents par langue qui se distinguaient par la complexité des représentations. Les modèles effectuent ensuite la même tâche que les humains dans l'expérience antérieure effectuée par (Darcy et al., 2009). Les résultats démontrent que les humains (a) incorporent des connaissances acoustiques et phonétiques adaptées au contexte, mais (b) n'emploient probablement des connaissances ni du lexique, ni des frontières lexicales.

Bref, cette étude se penche sur différents aspects de la perception. La combinaison d'expériences perceptuelles et de modélisation informatique se complémentent. D'un côté, les expériences perceptuelles explorent la variation entre auditeurs et permettent de guider le développement de modèles informatiques pouvant tenir compte de cette variation. De l'autre côté, la modélisation informatique permet de tester les hypothèses à propos du comportement des auditeurs.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This dissertation would have not been possible without the help and support of many people. I am deeply thankful to my advisors Morgan Sonderegger and Meghan Clayards. They both have long been mentors to me, and have been extremely supportive in whichever research topics I wanted to pursue, and showed great understanding even when I made slow progress. Morgan is very knowledgeable about statistical modelling and has taught me a great deal about analyzing data and interpreting the results. His classes were the very first I took to write code, which sparked my interest and gave me confidence in my subsequent exploration of computational linguistics. He has been a very caring advisor and offered tremendous help when I came across difficulties, from my decision to transition to computational modelling to deciding on the projects in this dissertation. It would not have been possible to conduct the project in Chapter 4 without him introducing me to the Cognitive Machine Learning group at LSCP, École Normale Supérieure.

Meghan is an extremely kind advisor and is very knowledgeable about speech perception. I was very fortunate to have her when I first came to McGill as my first-year advisor, who helped me quickly adapt to life in graduate school and a foreign country, and kindly mentored me throughout my graduate studies. She has taught me a great deal about speech perception, and I would not have been able to conduct studies in chapter 2 without her supervision. She has also co-supervised my subsequent computational modeling projects, and brought great insights to interpreting the results. Her genuine interest in different ways of studying speech perception, even when results were hard to interpret, has been a constant source of inspiration and motivation.

I would also like to thank Emmanuel Dupoux, Ewan Dunbar and Timothy J. O'Donnell who co-supervised me on the projects in this dissertation. I thank Emmanuel Dupoux for kindly accepting me as a visiting student in his group. Emmanuel has great expertise in computational modelling, and came up with interesting ideas in every meeting we had. Ewan Dunbar co-supervised the project in chapter 4, and not only provided great insight in terms of designing the experiments and interpreting the results, but also encouraged me with the enthusiasm and energy he radiated in every discussion we had about the project. Timothy J. O'Donnell

co-supervised the project in chapter 3, and helped greatly with interpreting the results and revising the manuscript—especially writing and organizing my thoughts clearly.

I am also thankful to the Linguistics Department at McGill University, and past and present students in the department. The Linguistics Department is an exceptionally warm community where I feel a sense of belonging. I am deeply thankful for the department's support of my switch to a computational research direction, by supporting all my unconventional requests—from taking a full year of computer science classes to arranging financial support during the reulting delay. I especially thank the instructors of my first-year classes: Jessica Coon, Brendan Gillon, Michael Wagner and Morgan Sonderegger. With their help, I learned a lot and gained confidence, coming from a non-linguistics background. I also thank my cohort and fellow students for making my graduate life enjoyable and intellectually stimulating: Martha Schwarz, September Cowley, Amy Bruno, Kevin Qin, Jurij Bozic, Jeff Lamontagne, Hye-Young Bang, Donghyun Kim, James Tanner, Yeong Woo Park, Jiaer Tao, Henrison Hsieh, and all the others. I especially thank Jeff Lamontagne for translating my thesis abstract into French. I would also like to thank the students and engineers at the Cognitive Machine Learning group for helpful discussion and setting up experiments, especially Juliette Millet, Mathieu Bernard and Julien Karadayi.

The writing of this dissertation happened in a difficult time during the pandemic, and would have not been possible without support from many people. I thank my long-time friend Mengdi Tang for always giving me unconditional support and care throughout my graduate studies. I thank my friends in the *Brilliant, My Brother!* group for all their support, and just being there. I thank my friends in the *Cut the fish-touching hands* (meaning 'stop wandering around') group, Chenyu Tu and Zi Huang, who worked with me to build social pressure for each other to keep working, from the beginnng of the pandemic. I thank Yenan Sun for daily check-ins and encouragement. The McGill Graphos writing workshop has been very helpful for me to get writing done in an efficient manner. I thank my parents for their unconditional love, understanding and support for whatever decisions I make.

Finally, I would like to thank composers and performers of classical music, whose works have given me constant emotional support. The writing of chapter 3 would not have been possible without Schubert's Impromptus Op.90, D.899, and chapter 4 without Mahler's Symphony No.3. I especially thank Mahler, Brahms, Schubert, Franck, Rachmaninoff and Shostakovich for composing various masterpieces. I also thank my neighbors for bearing with and encouraging my piano practice. I thank Orchestre Symphonique de Montréal and Salle Bourgie for reopening March 27, 2021, and hosting many concerts, which have helped me stay sane and motivated during the last and the most difficult period of thesis writing.

Contribution of authors

The studies presented in this thesis have each been prepared for publication in peer-reviewed journals. I am the primary author of each manuscript.

Chapter 2 has been published in *Laboratory Phonology*, and was co-authored with Meghan Clayards and Morgan Sonderegger. I am responsible for conceiving of the study and its research questions, data collection, and the original draft of the manuscript. I designed the experiments in consultation with Meghan Clayards. Statistical analysis was performed by me in consultation with Morgan Sonderegger. The interpretation of the results was performed by me in consultation with Meghan Clayards. Meghan Clayards and Morgan sonderegger provided comments for revision of the manuscript.

Chapter 3 has been prepared for submission, and is co-authored with Meghan Clayards, Mirco Ravanelli and Timothy J. O'Donnell. The research question was conceived by Morgan Sonderegger and me. I am responsible for the design of the experiments, data extraction, building of the model and the first draft of the manuscript. I did the training in consultation with Mirco Ravanelli. I led the interpretation of the results in consultation with Meghan Clayards and Timothy J. O'Donnell. I revised the manuscript with input from Meghan Clayards, Timothy J. O'Donnell and Morgan Sonderegger.

Chapter 4 has been prepared for submission, and is co-authored with Ewan Dunbar, Morgan Sonderegger, Meghan Clayards and Emmanuel Dupoux. Emmanuel Dupoux, Ewan Dunbar and I are responsible for conceiving of the study and design of the experiments. The data was cleaned and provided by Emmanuel Dupoux's Cognitive Machine Learning Group. I am responsible for the original draft of the manuscript. The interpretation of the results was performed by me with feedback from all authors. The evaluation metrics were designed by me in consultation with Morgan Sonderegger. The first four authors provided comments for revision of the manuscript. A preliminary report of this study was published in *Proceedings of CogSci 2020* (Jiang et al., 2020).

Chapter 1

Introduction

During speech communication, listeners need to understand the speech produced by the speaker, a process of extracting and integrating various acoustic phonetic cues from the complex waveform to identify the spoken words. One challenge imposed on speech perception is the high variability in the speech signal. Studies have found that non-canonical variants constitute 27% to 75% of instances for some sounds in conversational speech (e.g. Dilley and Pitt, 2007). Indeed, when presented with the same task of spoken word recognition, while state-of-the-art Automatic Speech Recognition (ASR) systems reach near-perfect performance when given clear read speech, they perform much worse on more noisy and variable speech (Davis and Scharenborg, 2016; Spille et al., 2018). Humans, on the other hand, have no trouble processing speech with all kinds of variations.

Given such discrepancy between humans and machines, the perception and recognition of highly variable speech has been of central interest in the fields of both automatic speech recognition (ASR: the engineering perspective) and human speech perception (the cognitive science perspective). The long-standing problem of the lack of invariance in speech has inspired extensive research in speech perception to understand how humans deal with noisy speech signals. At the same

time, it is also one of the central tasks to be solved in ASR, to get high quality transcription. While the two fields have different focuses—human perception and ASR models—they are intrinsically interdependent. Humans' behavioral results provide the benchmark for evaluating ASR models, and the explicit implementation of the ASR models qualified by the evaluation improves our understanding of human speech perception. The comparison between models and humans also sheds light on how we may improve ASR systems to better account for pronunciation variability.

In this chapter, I first review the sources of variability in speech and potential solutions used by human listeners, then discuss the behavioral-experiment and computational-modeling approaches to studying perception of variability in speech. I conclude the chapter by outlining the studies making up this dissertation.

1.1 Source of variability in speech

1.1.1 Phonetic Context

Coarticulation is a major source of variation from the canonical form. It results from the overlap of gestures of adjacent sounds, which produces context-specific variation in phonetic realization of the target sound. Some types of coarticulation are language-universal, which we touch on in Chapter 3; others are language-specific, such as phonological assimilation, which we examine in Chapter 4.

There are various kinds of language-universal coarticulation, such as consonant-to-vowel coarticulation (e.g. Harrington et al., 2008; Sussman and Shore, 1996), nasal-vowel coarticulation (e.g. Beddor and Krakow, 1999; Beddor et al., 2013), effects of lip rounding on surrounding phones (e.g. Fujisaka and Kunisaki, 1976; Heinz and Stevens, 1961) and so on. Aside from segmental coarticulation, tonal coarticulation

is also found in various tonal languages (Brunelle, 2003; Chang and Hsieh, 2012; Chen et al., 2018; Potisuk et al., 1997; Zhang and Liu, 2011), where the F0 contour of lexical tones are affected by the preceding or the following tone. These kinds of coarticulation are 'universal' in the sense that they exist in qualitatively similar form across languages (e.g. /a/ is realized with lower F1 before /i/); the actual phonetic implementation (e.g. the degree of F1 change) differs by language.

In addition, some types of coarticulation are language-specific; these are usually termed 'assimilation' (Farnetani and Recasens, 1997). Unlike language-universal coarticulation, specific types of phonological assimilation are only found in certain languages. For example, as discussed in Chapter 4, regressive nasal place assimilation is found in English, but not French. Another example is vowel harmony, where vowels in a word show agreement in terms of some phonological property (Van der Hulst, 2016).

1.1.2 Dialectal differences

Dialects also introduce structured variability within the same language (Weinreich, 2012). The realization of various speech sounds have been extensively studied, such as vowels (Williams and Escudero, 2014), consonants (Tanner et al., 2020), and tones (Li and Chen, 2016; Xu, 1994, 1997; Zhang and Liu, 2011). Allophonic variants can also vary among dialects, for example, intervocalic /t/ in post-tonic position is usually realized as glottalized in British English (e.g. Ashby and Przedlacka, 2014; Przedlacka and Ashby, 2011; Stuart-Smith, 1999) but becomes a flap in American English (e.g. De Jong, 1998; Riehl, 2003).

Moreover, dialects further differ in terms of the importance of cues to the same phonological contrast. For example, as the study in Chapter 2 shows, while different varieties of Wu Chinese use voice quality (breathiness) as one cue to distinguish be-

tween 'upper' and 'lower' register tones, the importance of this cue varies between varieties.

In addition, dialects differ from each other in the specific categories they use. For example, among varieties of Wu Chinese, traditional varieties have eight tones (e.g. Jiaxing, Shaoxing: Yu, 1988; Zhang, 2006); while varieties with more contact with Mandarin, which has a four-tone system, have undergone tone mergers (e.g. Shanghai Wu with five tones: Chen and Gussenhoven, 2015). As a segmental example, different dialects of North American English differ in their vowel inventories, such as whether low back vowels (*cot*, *caught*) have merged or remain distinct.

1.1.3 Talker idiosyncrasy

Aside from linguistic effects, individual talker variability is a source of signal variability. One source of systematic differences between talkers is gender (Oh, 2011), resulting from a combination of physiological differences and sociolingusitic factors (Herrmann et al., 2014; Whiteside, 1996). Male speakers typically have larger vocal tracts with longer vocal folds than female speakers, resulting in systematic gender differences such as lower F0 and vowel formants (Stevens and House, 1955; Titze, 1994). However, there are cases where inconsistent gender effects have been found in different studies of the same case—such as VOT in English—which may be caused by non-biological factors, such as speech style differences.

Beyond gender differences, individual talkers vary from each other in the use of acoustic cues. For example, different talkers produce different ranges of VOT to signal stop contrasts (Allen et al., 2003; Chodroff and Wilson, 2017; Scobbie, 2009) and sibilants (Bang and Clayards, 2016). Individuals show structured variation in their use of cues within and across different types of contrasts (Bang and Clayards, 2016; Chodroff et al., 2015).

1.1.4 Summary

To summarize, the variability in speech stems from a number of sources, both linguistic and non-linguistic. Linguistic sources include the phonetic context, while non-linguistic sources include talker idiosyncrasy, speech rate, etc. In this thesis, Chapter 2 addresses dialectal differences, Chapter 3 investigates multiple types of variability for a single contrast, and Chapter 4 focuses on language-specific phonological assimilation.

1.2 Solutions to variability: multidimensionality in perception

The extensive variability in speech, introduced in the previous section, poses a challenge for listeners during speech communication. In order to distinguish between different speech sounds, listeners use meaningful temporal and spectral correlates, termed cues, to different extent (e.g. Clayards, 2018; Francis et al., 2008; Mayo and Turk, 2004). It is hard to find 'invariance' in how any given cue is used to signal a contrast, given the many sources of variability. The distributions of single cues to a contrast often overlap, even for the best-known cases, such as voice onset time (VOT) differences between voiced and voiceless stops in English (e.g. 'peach', 'beach': Lisker and Abramson, 1967). Therefore, listeners are not able to rely only on one cue to distinguish a contrast.

However, as phonologicial contrasts are typically multidimensional, listeners can gather information from a number of cues to jointly identify the speech sounds. In other words, contrasts betwee pairs of speech sounds are usually signaled by more than one acoustic-phonetic cue, and this 'multidimensionality' of contrasts is one solution listeners use to deal with extensive variability in speech (Raphael,

2021; Schertz and Clare, 2020). For example, English stops differ not only in VOT, but also the F0 of the onset of the following vowel. Native English listeners use a combination of both cues to distinguish voiced stops from voiceless in pre-stress syllable-initial position, although they rely primarily on the VOT difference and less on the vowel onset F0 (Abramson and Lisker, 1985; Gordon et al., 1993; Lisker, 1978; Whalen et al., 1993). The set of cues used, and their relative weighting, differs for the same contrast in other postions (e.g. word-final, where VOT is undefined, and preceding vowel duration is most important).

Moreover, variability itself can be informative, where the source of variability can be identified (e.g. surrounding context, talker gender), and used to compensate for the resulting variation. Researchers hold different views in terms of how listeners incorporate information about sources of speech variability into perception. Some hold the view that with sufficient cues, compensation may not be needed (e.g. Nearey, 1990, 1997; Oden and Massaro, 1978; Toscano and McMurray, 2010) while others argue there is evidence that listeners do need to take into account what they know about sources of variation such as talker gender or phonetic context (Cole et al., 2010; Jongman and McMurray, 2017).

Multidimensionality of phonological contrasts also facilitates speech perception in adverse conditions. When the primary cue—the cue listeners rely on the most—becomes less distinguishable, listeners are able to use other non-primary cues. For example, when perceiving tones in Mandarin Chinese, pitch is the primary cue used by listeners. However, a number of studies show that when the pitch cue is no longer available, listeners are found to be able to use other redundant cues (e.g. intensity, spectral information) to distingush tones (Kong and Zeng, 2006; Whalen and Xu, 1992).

To summarize, the multidimensional nature of phonological contrasts allows listeners to gain evidence from multiple acoustic correlates, which listeners use jointly to perceive sounds in running speech.

1.3 Computational Modelling

1.3.1 Rational Analysis

Computational modelling has been extensively used to investigate human cognitive systems. In linguistics and psychology, the dominant computational modeling approach follows the notion of Rational Analysis (Anderson, 2013). The core idea is that given a well-defined task to be solved by a cognitive system, whether human or machine, the system should behave rationally by finding a solution that is somehow optimal. By this logic, a computational model which solves the task in an optimal way should share some similarity with the human cognitive system, and we can gain insight into the human system by implementing and comparing different computational models, differing in their knowledge and/or algorithm. This approach allows us to study higher-level human cognition at the 'computational' level (Marr, 1982), without neeeding to make specific assumptions (e.g. about neural implementation).

The rational analysis approach has been adopted to study various aspects of speech perception, which is treated as statistical inference of the talker's intended pronunciation, or identity (Clayards et al., 2008; Feldman et al., 2009; Kleinschmidt and Jaeger, 2015, 2016; Kleinschmidt et al., 2018; Kronrod et al., 2016; Laurent et al., 2017; Schatz et al., 2021; Sonderegger and Yu, 2010). Feldman et al. (2009) and Kronrod et al. (2016) show that one can model perceivers as rational agents and predict the degree of *categorical* effects in perception. Specifically, they show that both a

strong categorical effect of consonants and a weak categorical effect of vowels can be accounted for in a unified model, which infers the speech sounds using the listener's knowledge of speech categories as well as the speech signal. Kleinschmidt et al. (2018) use an ideal observer model to show that one can use the distribution of cues from the speech data to infer the talker's identity. Laurent et al. (2017) use Bayesian models to examine the different roles of the auditory and motor theories of speech perception. Their results show that under perfect conditions, the two theories produce the same result. However, in other conditions, auditory based recognition is more efficient with learnt stimuli while motor based recognition is better in adverse conditions.

1.3.2 Computational models of human speech perception

A strand of computational modeling research particularly relevant for this thesis, into which Chapter 4 fits, uses ASR models to examine human speech perception (Dupoux, 2018; Scharenborg, 2007; Scharenborg et al., 2005; Schatz et al., 2013). Unlike the human brain which can only be indirectly probed through responses, an ASR model has its components clearly defined and implemented and can be trained to represent a 'listener'. When the model listener's 'perceived sound' (i.e. the model's output) corresponds to humans', the system can be seen as a possible parallel for human speech perception. Moreover, one can use the model to test hypotheses impossible to conduct on human participants: for example, testing the importance of the lexicon by not feeding lexical information to the model, while for humans the acquired knowledge cannot be undone. Using computational models to parameterize and test such conditions, which are theoretically plausible but impossible to examine directly in humans, enables us to gain a deeper understanding of the roles of different kinds of linguistic knowledge in speech perception.

Moreover, computational models can deal with data on a larger scale than human experiments, and thus produce more generalizable results, improving the ecological validity of our theories. Human perceptual expertiments typically study a small set of cues, using a selection of sounds that are deemed 'representative' by the researcher, due to practical constraints. Although the methodology has been widely used in the field of speech perception, the scale and variability of sounds studied is still much smaller compared to those listeners actually deal with. Computational models, however, can process a large amount of input in little time, and can take as input speech corpora larger than what could be presented in a perceptual experiment.

Researchers have proposed various computational models to account for human speech perception, for example, non-deep neural network models (TRACE, McClelland and Elman, 1986), and Bayesian inference models (Shortlist B, Norris and McQueen, 2008). While those models are successful in terms of accounting for some aspects of phoneme recognition and spoken word recognition, a limitation of these earlier models is that they do not take realistic continuous speech as input. For example, the TRACE II model takes a seven-dimension phonological feature as input, and the Shortlist B model takes a sequence of multiple phoneme probabilities over three time slices per segment.

Recent work has begun to use more realistic data for computational models of speech perception (e.g. Dupoux, 2018; Magnuson et al., 2020; Scharenborg, 2007). The use of continuous speech requires that the model is able to deal with variable phonetic realizations in speech, an important characteristics of the data which needs to be accounted for either by the human cognitive system or a computational model. Scharenborg (2007) discusses the similarities and differences between human speech perception and automatic speech recognition, and how the study of the two fields can inform each other. One advantage of using ASR models is that they allow for

modelling human speech recognition in the context of 'noisy' listening conditions. Dupoux (2018) makes a similar point on using computational models to address puzzles in linguistics and cognitive science in general, with a focus on language acquisition by human infants. Computational models allow for simulations of the language learning process, as they are able to take scalable and realistic speech input which parallel what infants receive.

More recently, the advancement of deep learning has allowed for models of speech perception which can process high-dimensional input and longer/variable-length sequences, and hence can deal with real speech. For example, Magnuson et al. (2020) used real speech as input to their 'EARSHOT' model (which uses Long Short-term Memory models), and found that the model shows several similarities with humans. In terms of the timecourse of phone recognition, EARSHOT exhibits the same qualitative pattern as humans for phonological competition. Moreover, the model's internal representation appears to be similar to humans' neural activities, measured using 'representational similarity analysis', which quantifies the similarity of feature and phoneme selectivity in EARSHOT with human electrocorticography data. Magnuson et al. show that the part of EARSHOT mapping speech sounds to sematic representations (vectors representing words) learns similar linguistic features (e.g. sonorants, fricatives) to humans. This work illustrates how the use of real speech is crucial to explore similarity between computational models and humans.

1.3.3 Exploring the learnt knowledge of ASR models

Aside from using computational models to study human speech perception, we can also in turn use what we know about human speech perception to study the specific linguistic knowledge learnt by computational models. While deep neural network models achieve superior performance to earlier methods in many domains, they

suffer from a lack of interpretability, mainly due to their large parameter size and lower modularity (e.g. versus HMM-GMM models for ASR). However, it is important to understand what is learnt by the model, or in other words, what evidence the model uses to make its predictions.

Researchers have started to examine the learnt latent representations in various kinds of neural network models (Belinkov et al., 2019; Belinkov and Glass, 2017; Nagamine et al., 2015; Scharenborg et al., 2018; Ten Bosch and Boves, 2018; Weber et al., 2016). Many have found that those models learn similar linguistic units as humans, for example, phonological features and their groupings. Weber et al. (2016) examined a two-dimensional representation extracted from a phone classification model (a multi-layer perceptron) trained on vowels. They found that the dimensions of the representation corresponded to the first and second formants of the vowels – precisely the cues humans use to distinguish among vowels. Moreover, they found that the learnt representation of vowels are similar to the quadrilateral vowel space in phonetics. Other studies show that the clustering of phones using a learnt low-dimensional representation corresponds roughly to linguistic feature-based segmental categories such as stops and fricatives (Bai et al., 2018; Grósz et al., 2020).

1.4 Overview

This dissertation addresses the issue of perceiving highly variable speech using a combination of behavioural and computational approaches, applied to data from multiple languages. I examine different levels of human speech processing, from low-level phonetics to higher-level abstract patterning: listeners' specific use of acoustic dimensions in various linguistic contexts, the perceptual representation

integrating all acoustic dimensions for a phonological contrast, and the linguistic knowledge used for processing phonological changes.

This dissertation consists of three studies. The first study (Chapter 2) uses perceptual experiments to investigate how listeners differing in dialectal background make use of multiple cues to a contrast. In addition, we investigate whether or not individuals show differences in perceptual strategies in a structured manner. The contrast of interest is the two-way tonal register contrast in two Chinese Wu dialects signaled by multiple cues. We focus on the two well-studied pitch cues (pitch height and pitch contour) as well as an understudied voice quality cue (the degree of breathiness). The findings reveal that listeners differ mainly in their overall cue acuity (e.g. there are listeners with flatter and steeper boundaries between sounds – across all cues). Moreover, for certain contrasts signaled without a dominant cue, individuals further differ in their choice of the primary cue. Finally, listeners' use of cues is affected by their dialect background. The two Wu dialects mainly differ in terms of how multidimensional the tonal contrast is: Jiashan Wu is more multidimensional while Shanghai Wu is dominated by pitch height, with voice quality being a minor cue. We found that Shanghai listeners do not make better use of the voice quality cue more even when listening to the Jiashan stimuli, which are produced with more salient breathiness.

The second study (Chapter 3) answers a similar question of the use of cues for a multi-dimensional tonal contrast from the computational modelling perspective. This study targets the four-way tonal contrast in Mandarin Chinese and has two goals: 1) the relative contribution of cues for the tonal contrast in continuous speech, focusing on pitch, intensity and duration; 2) the perceptual representation of tones in a low-dimensional space. While Mandarin tones have been extensively studied, the majority of phonetic research has focused on isolated words, and it is still unclear how tones are perceived in continuous speech by integrating the rich set of

spectral-temporal cues. The results show that in continuous speech, pitch is again the most important cue, while intensity is the least important. Moreover, we found that the models learn a two-dimensional tone representation compressing the high-dimensional information in the input, without sacrificing accuracy. A closer examination of the perceptual tonal representation reveals that pitch is represented as average pitch height and pitch contour in the two dimensions. This lends converging evidence on the representation of Mandarin tones to behavioral studies, which have argued for the same representation from perceptual experiments. Furthermore, segmental information, encoding the segment- tone correlation, is implicitly learnt and used for tone prediction. The general methodology used in this study, applied to a language (Mandarin) where much is already known about the tonal contrast, opens up a new approach for investigating tonal representation across languages, including languages for which there is little previous work (most tone languages).

The third study (Chapter 4) uses Automatic Speech Recognition models to examine the type and complexity of linguistic knowledge needed for compensation for phonological assimilation. A series of ASR models trained for phone recognition are used, differing in the types and extent of linguistic complexity they 'know', and then compared against human benchmarks performed on the same task. Human listeners display language-specific compensation patterns, where English listeners compensate more for place assimilation than voicing assimilation, while French listeners show the opposite pattern. A first question is simply whether *any* ASR model captures this type of knowledge. The results reveal that some models show language-specific patterns comparable to those shown by human listeners. Models that best predict the human pattern use contextually sensitive acoustic models and language models, which capture allophony and phonotactics, but do not make use of higher-level knowledge of a lexicon or word boundaries. Moreover, the ASR system's model of the pronunciation of phones (the acoustic model) turns out to be

more important than knowledge about sequences of phones (the language model), meaning that successful ASR models encode language-specific phonetic knowledge to realize the language-specific compensation pattern.

Chapter 2

Individual and dialect differences in perceiving multiple cues: A tonal register contrast in two Chinese Wu dialects

2.1 Introduction

Phonological contrasts are usually signaled by multiple acoustic correlates (see Raphael, 2021, for an overview). In perceiving each contrast, listeners rely on each of these cues to a different extent (e.g. Clayards, 2018; Francis et al., 2008; Mayo and Turk, 2004). Such inequivalence in the contribution to contrasts is called cue weighting (Holt and Lotto, 2006). A widely studied example is the stop voicing contrast in English, where voiced stops show shorter Voice Onset Time (VOT) and a lower onset F0 while voiceless stops show longer VOT and higher onset F0. Native English speakers primarily use VOT to distinguish voiced stops from voiceless in pre-stress

syllable-initial position, with F0 playing a smaller role in perceiving the contrast (Abramson and Lisker, 1985; Gordon et al., 1993; Lisker, 1978; Whalen et al., 1993).

Studies show that cue weights vary as a function of the phonological contrast being signaled and the relative importance of cues also varies with linguistic contexts within a language (Oden and Massaro, 1978). For example, Mayo and Turk (2004) examined the role of VOT and formant frequencies for stop voicing contrasts in different vowel contexts in English. They found that while VOT was always the most important cue, listeners used formant transitions more to distinguish between /ta/ and /da/ than between /ti/ and /di/. For the tonal register contrast in Shanghai Wu, Zhang and Yan (2015) found that F0 onset and voice quality (i.e. breathiness) have different relative weights across different syllable onset manners and different utterance positions.

In the current study, we approach how multiple cues signal multi-dimensional contrasts by examining a tonal register contrast in two Chinese Wu dialects – Jiashan Wu and Shanghai Wu, with a focus on the role of secondary/non-primary cues.[1] In both dialects, three cues are used in signaling the contrast: pitch height (F0 onset), voice quality, and pitch contour (pitch slope). Pitch height is considered the primary cue and the role of voice quality is thought to vary across dialects, playing a weaker role in Shanghai Wu in younger generations (e.g. Gao, 2016). No previous studies have examined perception of Jiashan Wu or quantitatively compared the roles of the different cues to this contrast at the same level of detail in either dialect. Furthermore, while previous studies have examined the role of segmental context in Shanghai Wu (Zhang and Yan, 2015), we compare different tones (contexts) across two experiments. Individual differences are examined in both experiments. By examining these sources of variability (dialect, tone context and individual) in this

[1]We use the term 'secondary' to indicate all non-primary cues, not distinguishing between the second most important and any others. Secondary is therefore interchangeable with 'non-primary'.

multidimensional tonal register contrast, this study aims to answer the following questions: 1) Averaging over listeners within each dialect, how important is the secondary cue (i.e. voice quality) in a multidimensional contrast and does it vary by tone contrast, or is it consistent for different tone pairs? 2) As sound change is taking place in Shanghai, does the reduction of saliency of the voice quality cue result in different cue weighting than for traditional Wu (i.e. Jiashan) listeners? 3) Do individuals show structured differences in cue ordering or cue magnitude, or are differences between individuals random variation? Does the status of the cues in the dialect affect the structure of individual variability?

In subsequent sections, we first discuss the role of secondary cues in individual variability and sound change (2.1.1); in section 2.1.2, we discuss the role of voice quality cross-linguistically. We then give a brief introduction to the two dialects in the study (2.1.3) and conclude with a more detailed outline of the current study and its methodological contributions (2.1.4).

2.1.1 Multiple cues: individual variability and sound change

Individual variability in speech perception has been well documented (see Yu and Zellou, 2019, for a review). There is also evidence that individuals may vary in their use of secondary cues in perception (see Schertz and Clare, 2020, for a review of individual variability in cue weights). Some individuals use a secondary cue more than others for F0 in English stop voicing (Kapnoula et al., 2017; Kong and Edwards, 2016; Shultz et al., 2012) and vowel duration for English tense/lax vowels (Kim and Clayards, 2019). A common method to quantify individual cue weights is to use regression coefficients fit to each individual's responses (e.g. Shultz et al., 2012), or to use by-individual deviations from the population coefficient for a single regression model fit to all data ('random slopes', e.g. Clayards, 2018, see also Schertz and Clare,

2020, for discussion of different methods). Using these methods, researchers have tried to determine if individuals differ from each other systematically by looking at whether individuals' cue weights are correlated across dimensions. In other words, they asked whether those with larger than average primary cue coefficients have smaller than average or larger than average secondary cue coefficients.

Results have been mixed, with some studies finding a weak or non-significant relationship between cues in perceiving a contrast (Clayards, 2018; Shultz et al., 2012, for F0 and VOT for English stop voicing) or positive correlations (Clayards, 2018; Kim and Clayards, 2019, for vowel formants and duration for English tense-lax vowels). For the Korean three-way laryngeal contrast, one study found positive, negative or no correlation between cues depending on the contrast pair (Kong and Lee, 2018). Clayards (2018) examined coefficients for individuals across different contrasts in English (e.g. tense/lax vowels and word-final fricative voicing) and found that positive correlations were the most common for both primary and secondary cues (e.g. individuals with larger coefficients for the secondary cue in tense/lax vowels also had larger coefficients for the secondary cue in final fricatives). This suggests that differences between individuals may be systematic and not tied to particular contrasts or dimensions (cf. Hazan and Rosen, 1991). Some researchers have argued that more use of a secondary cue is associated with more gradient sensitivity to primary cues, using a visual analog scaling task rather than a categorical decision task (Kapnoula et al., 2017; Kong and Edwards, 2016). Thus, the positive correlations in Clayards (2018) and the relationship between gradient sensitivity and cue use (Kapnoula et al., 2017; Kong and Edwards, 2016) both point to some listeners' speech perception being more closely tied to the acoustics of the stimulus than others. However, as noted above, not all studies have found this relationship.

Individual variability in secondary cue use, may also play an important role in sound change. Some sound changes involve a non-primary cue taking over the role of a primary cue as occurs in 'tonogenesis' (Kingston, 2011), for example, the case of some younger speakers of Afrikaans shifting from VOT to f0 to signal a voicing contrast (Coetzee et al., 2018). In contrast, in the case of Shanghai Wu, a non-primary cue (i.e. breathiness) to the register contrast is losing importance (Gao, 2016). We thus may expect to observe increased individual variability in the use of non-primary cues in a variety undergoing sound change (though see Coetzee et al., 2018, perception data). Conversely, since Shanghai Wu is undergoing a loss of a non-primary cue, rather than an increase in non-primary cue importance, we may see more individual variability in a contrast in a variety that is *not* undergoing the loss of a non-primary cue (Jiashan Wu), and is therefore more dependent on multiple dimensions (e.g. Mayo and Turk, 2005, find larger differences between individuals, in this case adults and children, on contrasts with a larger role for the non-primary cue).

2.1.2 The perception of voice quality cues

Voice quality, one of the cues of interest in this study, plays different roles in speech perception in different languages. Furthermore, whether voice quality is used to signal a contrast and whether listeners are perceptually sensitive to voice quality are only partially related. This section summarizes the perception of voice quality in three types of languages.

First, in some tone languages, voice quality can be either a phonemic dimension that is independent of pitch, or it can be the main cue to a contrast that also has pitch differences. For example, in Yi, tones with the same pitch can be associated with different phonation types, and listeners rely on phonation cues to distinguish the tones

(Kuang, 2011). Jalapa Mazatec (Garellek and Keating, 2011) provides a similar case, where phonation contrasts are independent of pitch contrast, with each tone (low, mid, high) associating with different phonations (laryngealized, modal, breathy). In addition, phonation cues also appear to be perceptually important when different voice qualities are associated with two tones of similar pitch (although not identical). For example, White Hmong has two tones with similar pitch (21 vs. 22 in Chao numbers) that also contrast in modal vs. breathy voice. Here listeners were found to attend to voice quality information but ignore changes in pitch height or contour (Garellek et al., 2013). Similarly, in Sgaw Karen, pitch information is limited and instead, voice quality is crucial for listeners to distinguish the tonal contrast (Brunelle and Finkeldey, 2011). Thus, voice quality can be an independent contrast, or it can be the primary cue in cases where the contrast also includes is some difference in pitch.

Second, some languages use phonation as a redundant or non-primary cue associated with certain tones. For example, in Mandarin Chinese, creaky voice facilitates the perception of tone 3, a dipping tone (Kuang, 2013). Similarly, creaky voice in Cantonese tone 4 syllables increases tonal identification accuracy (Yu and Lam, 2014). In Black Miao, tones contrasting in both phonation and pitch are better distinguished than tones contrasting only in pitch (Kuang, 2013). In Northern Vietnamese, creakiness turns out to be as important as pitch in tonal categorization for one pair of tones (Brunelle, 2009).

Third, some languages like English do not use phonation to mark contrasts. Nonetheless, English listeners' perception of talker pitch is influenced by spectral slope such that sounds with tenser/flatter spectral slope are heard as having a higher pitch for both synthetic and resynthesized speech (Kuang and Liberman, 2015; **?**). This may allow listeners to normalize for pitch range (Honorof and Whalen, 2005) by making use of changes in voice quality that occur in the higher parts of

speakers' pitch range (Hollien, 1974). However, listeners were not sensitive to the steepness of the spectral slope (as long as it was not entirely flat), i.e. the degree of breathiness/tenseness.

To summarize, voice quality plays a variety of roles cross-linguistically, and is used by listeners for different purposes. The next section gives a brief introduction to the target language of this study, Wu Chinese, in which voice quality facilitates tonal register contrasts.

2.1.3 Wu Chinese

Wu Chinese is spoken in Shanghai, Zhejiang province and southern Jiangsu province of China. The two target dialects, Shanghai and Jiashan, are both sub-dialects of Wu. According to Yip (2002, 1980), a register feature [+/- Upper] divides the tonal space according to pitch: [+Upper] indicates a higher pitch range (corresponding to the historical *yin* tones) and [-Upper] indicates a lower pitch range (corresponding to the historical *yang* tones) creating a tonal register contrast. Historically, the register was also related to initial consonant voicing, that is, voiceless consonants only occurred in upper registers (*yin* tones) and voiced consonants only occurred in lower registers (*yang* tones). While the restriction on the distribution of consonant voicing and register still exists in non-initial position, the voicing contrast has been lost in initial position (e.g. Chen and Gussenhoven, 2015).

The tone inventory of Shanghai and Jiashan dialect are represented using the register features in Table 2.1 and 2.2 together with Chao numbers (Chao, 1930) indicating the pitch contour (numbers 1–5 stand for lower–higher pitch). Tone notations of Shanghai Wu are from Xu and Tang (1988); as there is no previous study on Jiashan Wu, the transcriptions are based on the first author's experience as a native speaker and Yu (1988) on Jiaxing Wu, a highly similar dialect spoken nearby.

Dialect	Register	Falling	Level	Rising	Checked
Shanghai	*Yin*/ +Upper (modal)	53		34	55
	Yang/ -Upper (breathy)			23	12
Jiashan	*Yin*/ +Upper (modal)	53	44	35	55
	Yang/ -Upper (breathy)	31		13	12

Table 2.1: Tone inventory of Shanghai Wu and Jiashan Wu.

There are four types of tones: falling, level, rising, and checked. Checked tones are relatively short in vowel duration and end in a glottal stop. Both dialects have fewer than eight tones due to historical mergers. Shanghai dialect merged the level and the rising tones in the upper register, while falling, level, and rising tones were merged into one rising tone in the lower register (Qian, 1992), leaving a five-tone system. In Jiashan, only the lower-register level and rising tones were merged, resulting in a rising tone (Yu, 1988) and a seven-tone system.

Note that the tone notations in Table 2.1 do not accurately match the F0 trajectories, but are rather an abstract representation. Various other notations are proposed by different researchers (as discussed in Chen and Gussenhoven, 2015, for Shanghai Wu), as the pronunciations are varied across individuals and generations. However, no one notation can accurately reflects the exact F0 trajectories given the large variation. Therefore, similar to differences between formant values and distinct IPA symbols, the tones represented in Chao numbers are phonemic and do not necessarily correspond to the actual phonetic realization. For example, related to the current study, the Jiashan falling tone is realized with a steeper contour in the upper register than the lower register while the Shanghai rising tone is realized with a steeper contour in the lower register than the upper register, although the tone notations do not reflect such differences. For the checked tone, while it is always produced with shorter vowel duration, speakers vary in how audible the glottal stop is.

While pitch range signals the register contrast, it is not the only cue. Others involve voice quality, pitch contour, and duration. Crucially, the upper tonal register is produced with modal voice whereas the lower tonal register is produced with breathy voice (Cao and Maddieson, 1992; Chen, 2010; Gao et al., 2011; Jiang and Kuang, 2016; Zhang and Yan, 2015). In addition, in many Wu dialects, the steepness of the contour in contour tones differs in the two registers (e.g. Chen and Gussenhoven, 2015). [2]

While both Shanghai and Jiashan dialects share the same characteristics of register contrast, the Shanghai dialect is argued to be going through a loss of breathiness in the lower register, at least in production (Gao, 2016; Gao et al., 2011). Based on their acoustic and electroglottographic data, Gao and colleagues found that younger speakers used less breathy voice in the lower register compared to older speakers, possibly due to contact with Mandarin Chinese which does not employ breathy voice in tonal contrasts. However, Shanghai Wu listeners do seem to use voice quality in perception despite the change in production. An early study by Ren (1987) found that breathy voice was a perceptual cue for the lower register. A later perception study by Gao et al. (2020) found that voice quality did influence perception for both natural (produced by a trained phonetician) and synthesized stimuli. While Zhang and Yan (2015) found that Shanghai listeners heavily relied on F0 information for the register distinction, they also used breathiness as one of the non-primary cues. It is not known how much Shanghai listeners use breathiness relative to other Wu listeners, for whom this cue is not being lost. Direct comparison between listener groups is required to determine this.

To summarize, Shanghai Wu and Jiashan Wu both incorporate a breathy-modal distinction into the tonal register contrast as a non-primary cue. The two dialects

[2]Researchers vary in whether they consider the contour difference to be phonetic or phonological. Some treat it as an underlying difference while others consider it different phonetic realizations of the same phonological contrast.

have similar phonological systems, although Jiashan has a slightly richer tone inventory and stronger voice quality distinction. Moreover, as Shanghai is in the process of losing breathiness, its listeners may also exhibit some characteristics of listeners of non-phonatorily contrastive languages like English, that is, relying less on breathiness when compared with Jiashan listeners.

2.1.4 Current study

This paper examines the perceptual difference between dialects and across individuals on the tonal register contrast in Chinese Wu dialects, by manipulating three cues: pitch height (F0 onset), voice quality, and contour (pitch slope). We compare the falling tone pair in Jiashan Wu and the rising tone pair in Shanghai Wu (Experiment 1) and the checked tone pair in both languages (Experiment 2). Because the checked pair is the same in both dialects we can compare both groups of listeners on the same stimuli in Experiment 2. Unlike previous work on perception of voice quality, we explore breathiness as a perceptual cue in fine detail by creating a continuum from natural endpoints that changes in all aspects of the acoustic space for breathiness. Previous studies either only have two levels (breathy versus modal, e.g. Gao et al., 2020; Zhang and Yan, 2015), or manually modified only certain parameters (e.g. Garellek et al., 2013, who modulated H1–H2, H2–H4, H4–2kHz, and 2kHz–5kHz). We also investigate how different dialectal experience caused by sound change is manifested in the perception of multiple cues at both the group level and the individual level by comparing Shanghai and Jiashan Wu listeners on stimuli from both dialects. This paper addresses the three questions raised earlier.

First: averaging over listeners within each dialect, how important is a non-primary cue (i.e. voice quality) in a multidimensional contrast and does it vary by tone contrast, or is it consistent for different tone pairs? The current study manipulates the

three cues independently and includes a five-step breathy-modal continuum. This extends previous studies on cue weighting in Chinese Wu dialects by allowing different combinations of ambiguous cues. We predict that voice quality will significantly influence listeners' perception, at least for Jiashan listeners, but we expect it to be less important than pitch height. The relative importance of voice quality and pitch contour may change depending on the tone pair.

Second: as sound change is taking place in Shanghai, does the decreasing use of the voice quality in production result in different cue weighting than for traditional Wu (i.e. Jiashan) listeners? This study investigates how a dialectal background difference in degree of a non-primary cue production (i.e. voice quality) affects listeners' cue weights. To understand whether listeners from the two dialect groups show different perceptual strategies, and whether the potential difference is caused by the acoustic cues in the stimuli or differences in the listeners, this study examines listeners' cue weightings when they are exposed only to their native dialect (Experiment 1) and when they listen to both dialects (checked tones in Experiment 2). Because the Shanghai dialect is thought to be less breathy, we expect voice quality to have only a small effect on listeners' responses when they listen to the Shanghai stimuli. If both sets of listeners respond to voice quality in the same way, we expect voice quality to have a bigger effect when they listen to Jiashan stimuli, which have a larger voice quality difference. However, if Shanghai listeners are not as sensitive to voice quality as Jiashan listeners we expect their responses to be less affected by voice quality than Jiashan listeners when listening to Jiashan stimuli.

Relevant to the comparison in Experiment 2 is the question of to what extent listeners have experience with the other dialect. It should be noted that both dialects are mutually intelligible, and there are TV shows broadcast in both dialects. Since Shanghai is a large city it may be the case that Jiashan speakers have more opportunities to be exposed to Shanghai Wu than vice-versa. However, it is also

the case that Shanghai listeners have opportunities to be exposed to dialects with larger breathy-modal contrasts, although not necessarily Jiashan Wu. Within the city of Shanghai, there are other dialects of Shanghai Wu that differ from the more common 'downtown Shanghai Wu' used in this study in that they are more typical Wu dialects with larger breathy-modal contrast. Moreover, older Shanghai speakers produce breathier lower register words than younger speakers (Gao, 2016) and residents of Shanghai could also be exposed to speakers of other, mutually-intelligible Wu dialects with a greater breathy-modal contrast. Given these factors it is difficult to say whether one dialect group is more familiar with the other dialect or not.

Third: do individuals show structured differences in cue ordering or cue magnitude, or are differences between individuals random variation? Does the structure of individual variability differ by dialect? To address these issues, this study examines individual variability in relative cue weighting. Previous work on individual variability has examined only two cues to a contrast. By examining three cues we are able to better observe how cue weights relate to each other. Group-level results are sometimes insufficient to understand how a multidimensional contrast is perceived, as the results are averaged over all participants, giving only one pattern of cue weighting. How individual cue weights diverge from the average pattern is important for fully understanding the phonological contrast of the target language as well as the perceptual system of individuals. Individual variability in cue weights may be completely random, or structured (correlated). Individual differences could be small, with individuals only differing in the magnitude of the cues, while sharing the same ordering of cue weights, or larger, with individuals differing in the ordering of cue importance. Because so little is known about individual variability in cue use in perception, especially with more than two cues, it is difficult to make a priori predictions. To examine individual differences, we use the correlation matrix of the random effects of a mixed-effects model fit to the data. A methodological contribu-

tion of this paper is showing how this component of mixed-effects models, which is typically ignored in phonetic studies, can be used to understand the structure of individual differences.

2.2 Methods

The current study consists of two experiments. The first experiment examined Jiashan and Shanghai listeners' use of cues in stimuli from their own dialects; the second experiment again evaluated the three acoustic cues but exposed listeners to stimuli from both dialects and examined whether listeners have different cue weightings for the two sets of stimuli.

2.2.1 Participants

Two groups of listeners participated in the study. 34 native Jiashan Wu speakers (5 males, 29 females, aged 19-62 with mean of 34) were recruited in Jiashan and 35 native Shanghai Wu speakers (11 males, 24 females, aged 18-56 with mean of 22) were recruited in Shanghai (34) and Montreal (1). No participant reported hearing loss.

2.2.2 Stimuli

Experiment 1

In the first experiment, the stimuli varied by five steps in both pitch height and voice quality within the natural range between the upper and lower register, as determined by natural productions of native speakers. Pitch contour, however, only had two levels: the upper register contour and the lower register contour (as in Kirby, 2014, on Khmer). The number of stimuli were limited in this way so that

the experiment could be conducted in a reasonable amount of time, while allowing the cue of primary interest (i.e. breathiness) to vary on a continuum. Moreover, we did not expect the pitch contour to play a large role in listeners' perception. This is because various Chinese Wu dialects do not contrast degree of steepness (Chao, 1928; Qian, 1992) and while some dialects do exhibit different degrees of steepness of the two contours, many researchers treat this as different phonetic realizations of the same underlying form due to articulatory constraints.

The endpoint sounds were selected from recordings of a previous production study on several Wu dialects (Jiang and Kuang, 2016). Various acoustic parameters of breathy voice (*H1, *H1-*H2, *H1-*A1, *H1-*A2, *H1-*A3, CPP) were measured in all 12 speakers. The speaker that produced the largest breathy-modal contrast in each dialect was selected, based on a combination of auditory saliency of breathiness perceived by native speakers and extremity of measures (which largely coincided). From the productions of these two speakers the minimal pairs that had the greatest breathy-modal distinction were chosen. The original productions (i.e. two endpoints) of the Jiashan stimuli were a pair of upper and lower tonal register words ([ka 53] 街 'street' and [ka 31] 扛 'carry on the shoulder') with falling tonal contour produced by a female native speaker aged 43. The two monosyllabic words varied in three dimensions: voice quality (the upper register word [ka 53] is modal while the lower register word [ka 31] is breathy), pitch height ([ka 53] has higher pitch), and pitch contour ([ka 53] is steeper). The two endpoint productions for the Shanghai stimuli were produced by a female native speaker aged 24. The syllables were also ([ka]) while the tones were rising ([ka 23] 茄 'eggplant' and [ka 34] 价 'price'). Rising tones were avoided in Jiashan because they are realized more like dipping tones, sometimes including creakiness at low pitch in a low register word, which undermines the breathy voice quality examined in this study. Falling tones were avoided in Shanghai Wu because there is no low register falling tone (see Table 2.1).

The natural recordings were then modified using a combination of TANDEM STRAIGHT (Kawahara et al., 2008) and the PSOLA method in Praat (Boersma and Weenink, 2019) to create a series of stimuli. TANDEM STRAIGHT provides high quality source-filter resynthesis and allows the user to interpolate between two natural recordings. It morphs a pair of sounds in all dimensions (e.g. F0, duration, spectrum) simultaneously and holistically without reference to specific dimensions. The resulting continuum has the same recording quality and naturalness as the originals and varies in all dimensions of the original two sounds. TANDEM STRAIGHT has been successfully used in studies of many different phonological contrasts cross-linguistically (e.g. fricative place in German and English, nasal place in Japanese, vowel contrasts and sung melodies in Japanese: Bukmaier et al., 2014; McAuliffe and Babel, 2016; Sadakata and Sekiyama, 2011; Yonezawa et al., 2005). We believe that this method is also appropriate for creating a natural breathiness continuum that varies in all acoustic dimensions of breathiness, providing a more thorough modification than previous studies that only focus on a few dimensions.

To create the stimuli, we first normalized vowel duration using the PSOLA method between the two registers within each dialect (duration of voiced portion: JS 219ms; SH 283ms). For the Jiashan stimuli, the VOT of the initial consonant /k/ was similar for the two sounds, so there was no need to normalize. For the Shanghai stimuli, we normalized VOT, although both were very close to 19 ms. Second, we created a second copy of each word that had the pitch contour (but not pitch height) of its pair. To do this, we extracted the pitch contour from both sounds. The upper contour was lowered by 100Hz (the difference between the two sounds) by PSOLA, and was then superimposed on the lower register word. The lower contour was shifted up by 100Hz and was superimposed on the upper register word. Third, we used the two pairs of upper and lower register words matching in contour shape (one natural, one manipulated) to create two five-step continua in TANDEM STRAIGHT. Having

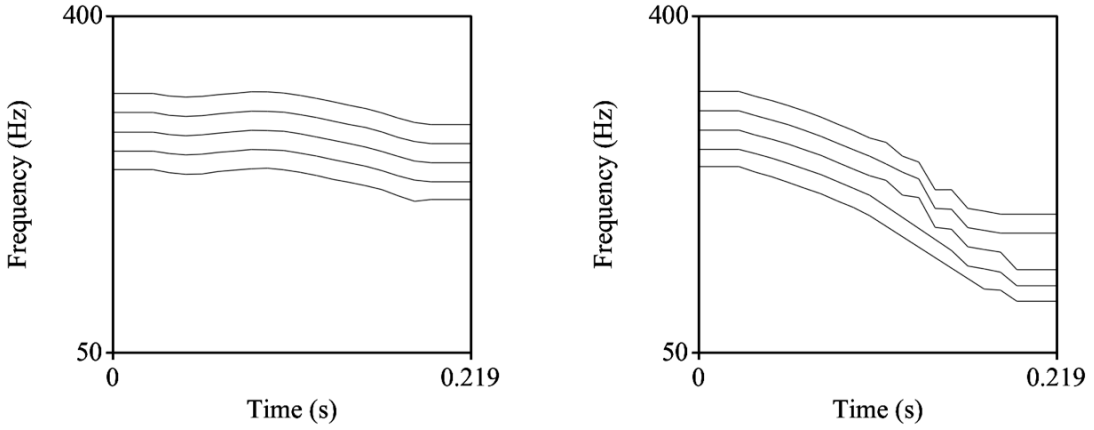
controlled the contour, the two continua varied in both pitch height and voice quality, one with a flatter contour and the other with a steeper contour. Since STRAIGHT does not generate an equal-stepped continuum, we set the program to generate 15 stimuli, from which we picked five that varied in pitch height in 25Hz steps. Since the manipulation changes all aspects of the source, the breathiness also varied in these five steps.[3] Fourth, we created four new versions of each stimulus varying in pitch height. We manipulated the pitch height of each stimulus by shifting it upwards or downwards in 25 Hz steps to create five steps total (Jiashan stimuli: F0 onset from 240Hz to 340Hz; Shanghai stimuli: from 190Hz to 290Hz). Together this process resulted in two 5×5 continua for each dialect varying in breathiness and pitch height, one with the flatter contour and one with the steeper contour for a total of 50 stimuli ($5 \times 5 \times 2$) per dialect. Figure 2.1 shows the 10 different pitch contours used for each of the two dialects.

Experiment 2

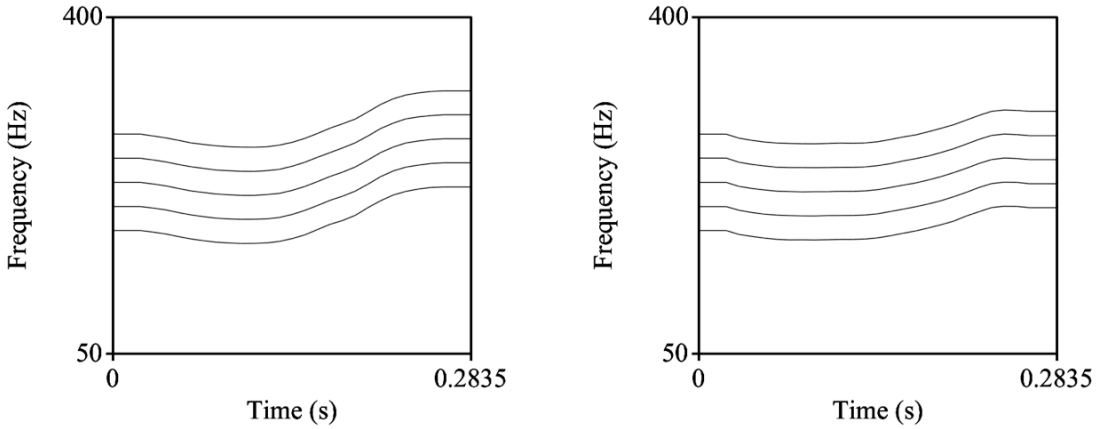
In the second experiment, the stimuli included three five-step continua, one for each cue, and for each continuum the other two cues were held at the acoustic mid-point between the two registers. This aimed to examine each cue independently, and by holding the other cues at what is expected to be the most ambiguous point, it minimized the reliance on other cues to best show listeners' use of the target cue.

The second experiment used a different pair of tones (i.e. checked tone) where the two dialects share the same phonological representation of the words used (i.e. [ka 55] 夹 'clip' and [ka 12] 挤 'jostle') and are mutually intelligible. Listeners heard both sets of stimuli from the two dialects, in order to examine whether the two groups of listeners differ when exposed to the same stimuli.

[3]Note that the generated stimuli are not precisely equal-stepped due to the noise introduced in STRAIGHT continuum generation, although this is unavoidable and we made sure that they are reasonably equal-stepped, given the large range of the register contrast.



(a) Left: /ka 31/, Jiashan, ΔF0 = 65Hz; Right: /ka 53/, Jiashan, ΔF0 = 160Hz.



(b) Left: /ka 23/, Shanghai, ΔF0 = 135Hz; Right: /ka 34/, Shanghai, ΔF0 = 37Hz.

Figure 2.1: Pitch contours (left panels lower register, and right panels upper register) and pitch height (five steps) used in the two dialects. ΔF0 refers to the F0 change between the onset and the offset of the syllable. The breathiness continua also have five steps, but are not shown schematically.

The original productions were a pair of /ka/ syllables with the upper and lower register checked tones. Recordings from the same speakers as in Experiment 1 were used to create three new continua respectively.

As in Experiment 1, stimulus construction was the same for both dialects and used a combination of TANDEM STRAIGHT and PSOLA in Praat. Prior to creating the continua, the two endpoint productions were normalized in duration (JS: 122ms; SH: 120ms) and VOT within dialect. We did not normalize the F0 range between the two dialects, because shifting the F0 could result in a change in breathiness, especially when the change was large.[4]

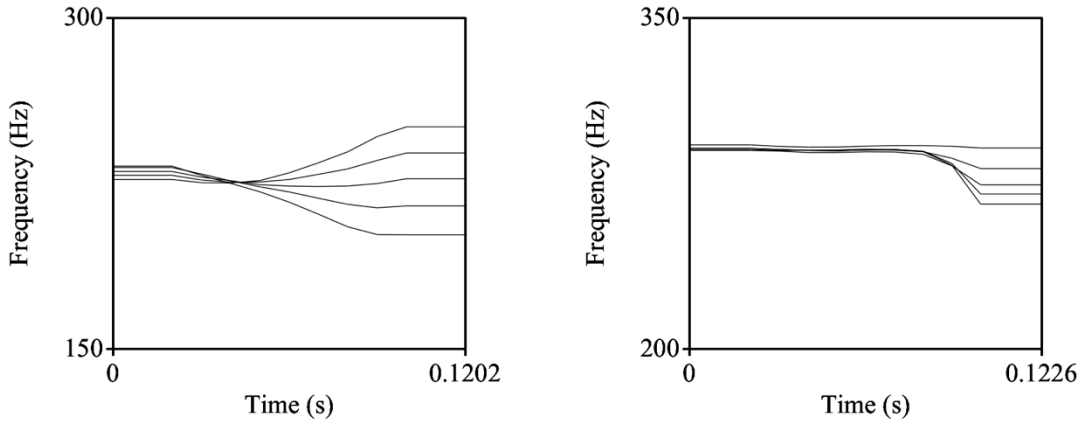


Figure 2.2: Contour continua of Shanghai (left) and Jiashan (right) used in Experiment 2.

BREATHINESS CONTINUUM: As before, we used TANDEM STRAIGHT to create a 15-step continuum from the normalized endpoints. From this we picked five steps that were equally spaced in F0 onset (25 Hz steps, Shanghai 190 Hz to 290 Hz; Jiashan 240 Hz to 340 Hz). We used Step 3 from this five-step continuum as the acoustically ambiguous step in terms of breathiness for the other two con-

[4]Test stimuli showed that manipulating F0 affects the amplitude of the harmonics to some extent, which affects the degree of breathiness. Since the continua already have a large F0 range (100Hz), we decide not to further introduce more confounding factors.

Contour	Shanghai talker		Jiashan talker	
Step	Start f0 (Hz)	End f0 (Hz)	Start f0 (Hz)	End f0 (Hz)
1	230	253	291	291
2	232	245	291	279
3	234	232	291	268
4	236	220	290	262
5	237	210	290	253

Table 2.2: Onset and offset f0 (Hz) for each step of the contour continuua of Experiment 2.

tinua. CONTOUR CONTINUUM: Using Praat, we extracted the endpoint F0 contours from the normalized endpoints and shifted them such that the midpoint F0 of both was midway between the originals. This created two F0 contours that only differ in the contour shape, which were used as the two endpoints of the contour continuum (see Table 2.2 for all values of contour stimuli). We then used a Praat script to linearly interpolate between the two endpoints to create a five-step continuum between these endpoints. We used Step 3 of the contour continuum as the acoustically most ambiguous step to be used for the pitch height and breathiness continua. PITCH HEIGHT CONTINUUM: We chose five F0 onset values to match those chosen for the breathiness stimuli (25 Hz steps, Shanghai 190 Hz to 290 Hz; Jiashan 240 Hz to 340 Hz).

The final contour continuum had the five contours applied to Step 3 of the breathiness continuum with the pitch onset value (i.e. pitch height) from Step 3 of the pitch continuum. The final breathiness continuum had Step 3 of the contour continuum and the pitch onset value from Step 3 of the pitch continuum applied to each of the five steps of the breathiness continuum. The final pitch continuum used Step 3 of the breathiness continuum and the contour from Step 3 of the contour continuum and varied in pitch onset according to the five steps.

The above manipulation created three five-step continua of pitch height, breathiness and contour while holding the other two factors constant and acoustically ambiguous. Each stimulus was repeated five times, resulting in 150 trials in total (5 steps$\times$ 3 continua$\times$ 5 repetitions$\times$ 2 dialects = 150).

All stimuli used in the two experiments can be found in Supplemental Materials at `https://osf.io/u7er5/`, together with corresponding acoustic measures for voice quality (e.g. H1-H2, H1-A1) and plots showing how these measures change along the five-step voice quality continuum. By all measures, the Jiashan stimuli show a larger breathy-modal contrast than the Shanghai stimuli for both experiments, supporting our claim that the TANDEM STRAIGHT method created breathiness continuua which vary simultaneously in all acoustic measures.

2.2.3 Procedures

The experiments were conducted using Matlab in a quiet room. All participants took part in both experiments with a five-to-ten minute break in between. They first listened to 250 sounds of their own dialect (i.e. Experiment 1) and then 150 sounds, 75 from each dialect (i.e. Experiment 2). All stimuli were played in random order. On each trial, participants heard a single syllable and the display presented two Chinese characters corresponding to the upper and lower register words together with the numbers 1 and 2 corresponding to the associated keys. The association between the number and the character varied randomly across the whole experiment. Listeners pressed the key associated to the word they perceived. After selecting the answer, they pressed the space key to proceed to the next trial. In Experiment 2, participants were told that they may hear more than one speaker, but no information about speaker or dialect was given.

2.3 Results

In this section, we first provide group-level results of the two experiments, showing how each of the two listener groups makes use of the three cues on average. We then present results for individual variability and show how listeners differ in a structured manner.

2.3.1 Experiment 1: Group results

Experiment 1 examines listeners' perception of tonal register contrast (upper vs. lower) of their native dialect when exposed to stimuli with different degrees of breathiness and pitch and with two contours. In order to investigate the relative importance of the three cues, two mixed-effects logistic model were fit for the two dialect groups respectively to model participants' response as a function of the three variables.

The data were fitted with a Bayesian mixed-effects logistic regression (MELR) model in R using version 1.0-4 of the blme package (Chung et al., 2013) using the default settings for priors. One advantage of a Bayesian MELR over the more widely used non-Bayesian MELR (e.g. using lme4 in R) is the ability to fit a maximal random-effect structure (all possible slopes and correlation terms; Barr et al., 2013) without convergence issues which frequently arise for non-Bayesian methods (including when applied to our data), usually related to "impossible" 1/-1 correlation values (Nicenboim & Vasishth, 2016; see Vasishth et al., 2018 for a tutorial). The dependent variable was the register response (upper = 1, lower = 0). All predictors were centered and standardized by subtracting the mean and dividing by two standard deviations. This makes the coefficients comparable (e.g. in Table 2.3, the coefficient estimates of pitch are the highest among the three cues, indicating that it is the primary cue) and minimizes collinearity between main effects and interac-

tions (Gelman & Hill, 2006). More importantly, centering means the interpretation of a certain main effect is averaged over other variables this specific variable interacts with. The fixed-effect predictors were voice quality step *breathiness* (coding for five steps are scaled in range [-0.7, 0.7]), pitch height step *pitch* (coding for five steps were scaled in range [-0.7, 0.7]) and pitch contour step *contour* (two levels: upper = 0.5 lower = -0.5). Two-way and three-way interactions were also included because changes in one cue may have an effect on participants' perception of the other cues. The random effects included by-participant random intercepts and random slopes for breathiness, pitch, and contour and their interactions, to account for individual differences in cue weights, as well as all possible correlation terms (the *maximal* structure), to measure potential correlations in cue usage among individuals.[5]

Cue weights in Shanghai Wu

Figure 2.3 shows the mean responses from the Shanghai participants. Table 2.3 shows the model results for fixed-effects terms. All three cues had a significant effect, and only the pitch-breathiness term was significant among all interactions. Note that since all variables are centered, the main-effect coefficient estimates indicate how much the log-odds ratio of responding with the upper register changes as the variable shifts by one unit, averaging over other factors (e.g. the β coefficient of contour means the effect of contour at the average pitch and breathiness values).

According to the coefficient estimates, pitch was the most important cue (β = 6.67, SE = 0.36, $p < 0.001$), much stronger than contour (β = 1.26, SE = 0.18, p ¡ 0.001), while breathiness was the least important (β = 0.55, SE = 0.11, p ¡ 0.001).[6]

[5]Model syntax: `bglmer (response ~ BreathDegree.std * PitchDegree.std * Contour.std + (1 + BreathDegree.std * PitchDegree.std * Contour.std| participant), data=df, family="binomial", control=glmerControl(optimizer = "bobyqa", optCtrl = list (maxfun = 100000)))`

[6]Likelihood ratio tests of the contribution of each cue, removing all pitch/breathiness/contour terms, give the same ordering: Pitch (χ^2 = 6462.3, $p < 0.001$), Contour (χ^2= 304.65, $p < 0.001$), Breathiness (χ^2 = 62.31, $p < 0.001$). Note that the model without pitch terms did not converge, presumably

Moreover, the two-way interaction between breathiness and pitch (β = 0.74, SE = 0.32, p = 0.024) shows that the higher the pitch, the more important voice quality is. There was a trend for a three-way interaction between the cues (β = 1.43, SE = 0.72, p = 0.052), indicating that how much the effect of breathiness is affected by pitch is itself modulated by the specific contour. Together these interactions suggest that the higher pitch onsets and the upper register contour were slightly more ambiguous than the lower pitch onsets and lower register contour.

	Estimate	Std. Error	z value	Pr($>\|z\|$)
Intercept	0.59	0.16	3.54	$<$ **0.001**
breath	0.55	0.11	4.83	$<$ **0.001**
pitch	6.67	0.36	18.76	$<$ **0.001**
contour	1.26	0.18	6.86	$<$ **0.001**
breath×pitch	0.74	0.32	2.25	**0.024**
breath×contour	0.23	0.19	1.16	0.244
pitch×contour	0.33	0.39	0.86	0.386
breath×pitch×contour	1.43	0.73	1.94	0.052

Table 2.3: Summary of the fixed effects for the model of Shanghai participant data from Experiment 1.

Figure 2.3 shows the corresponding heat map of the participants' categorization. The left panel represents the results for the lower register tone contour and the right panel represents the results for the upper register contour. The lighter color indicates higher percentage of upper register responses. In terms of cue weighting, compatible with the model predictions, there was indeed a large effect of pitch, as indicated by the change of colors along the y axis in both panels. Effects of breathiness and contour are much weaker, and are most visible in regions with ambiguous values of pitch (step 3). The effect of breathiness can be seen in a color change along

because the most explanatory variable for the data was omitted, so the Likelihood ratio test result for pitch is approximate.

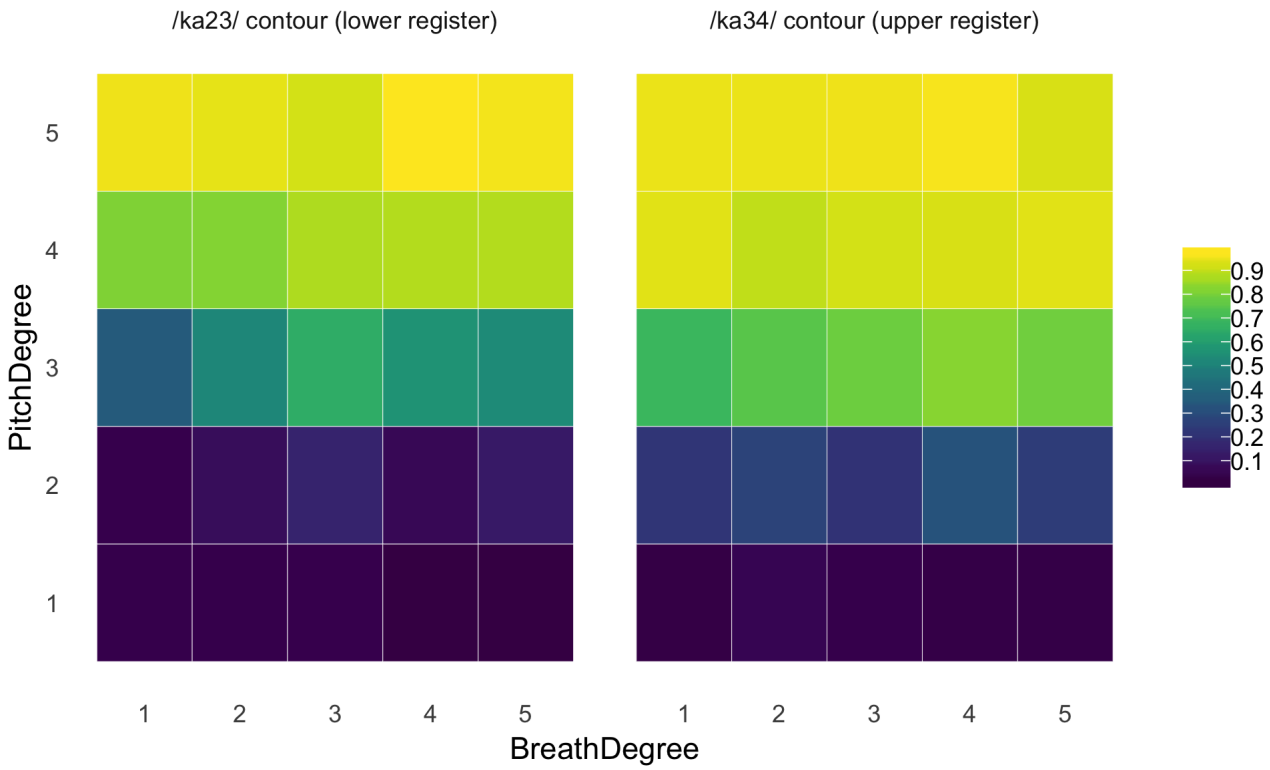


Figure 2.3: Percentage of upper register response from Shanghai participants. The x axis is the breathiness continuum (1 = breathy, 5 = modal). The y axis is the pitch continuum (1 = low pitch, 5 = high pitch).

the x-axis and the effect of contour can be seen in lighter colors in the right panel, consistent with model predictions.

Cue weights in Jiashan Wu

The Jiashan participants, however, show a different pattern. As displayed in Table 2.4, the Jiashan model found that contour had the biggest effect (β = 3.08, SE = 0.46, $p < 0.001$) followed by pitch (β = 1.63, SE = 0.30, $p < 0.001$) and breathiness (β = 0.86, SE = 0.16, $p < 0.001$).[7].

[7]Likelihood ratio tests of the contribution of each cue, removing all pitch/breathiness/contour terms, give the same ordering : Contour (χ^2= 3342.2, $p < 0.001$), pitch (χ^2= 1035.8, $p < 0.001$), breathiness (χ^2= 356.7, $p < 0.001$).

	Estimate	Std. Error	z value	Pr(> |z|)
Intercept	0.73	0.11	6.51	< **0.001**
breath	0.86	0.16	5.14	< **0.001**
pitch	1.63	0.30	5.45	< **0.001**
contour	3.08	0.46	6.62	< **0.001**
breath×pitch	-0.69	0.20	-3.37	< **0.001**
breath×contour	-0.52	0.23	-2.22	**0.02**
pitch×contour	-0.74	0.24	-3.06	**0.002**
breath×pitch×contour	-0.30	0.39	-0.76	0.44

Table 2.4: Summary of the fixed effects for the model of Jiashan participant data from Experiment 1.

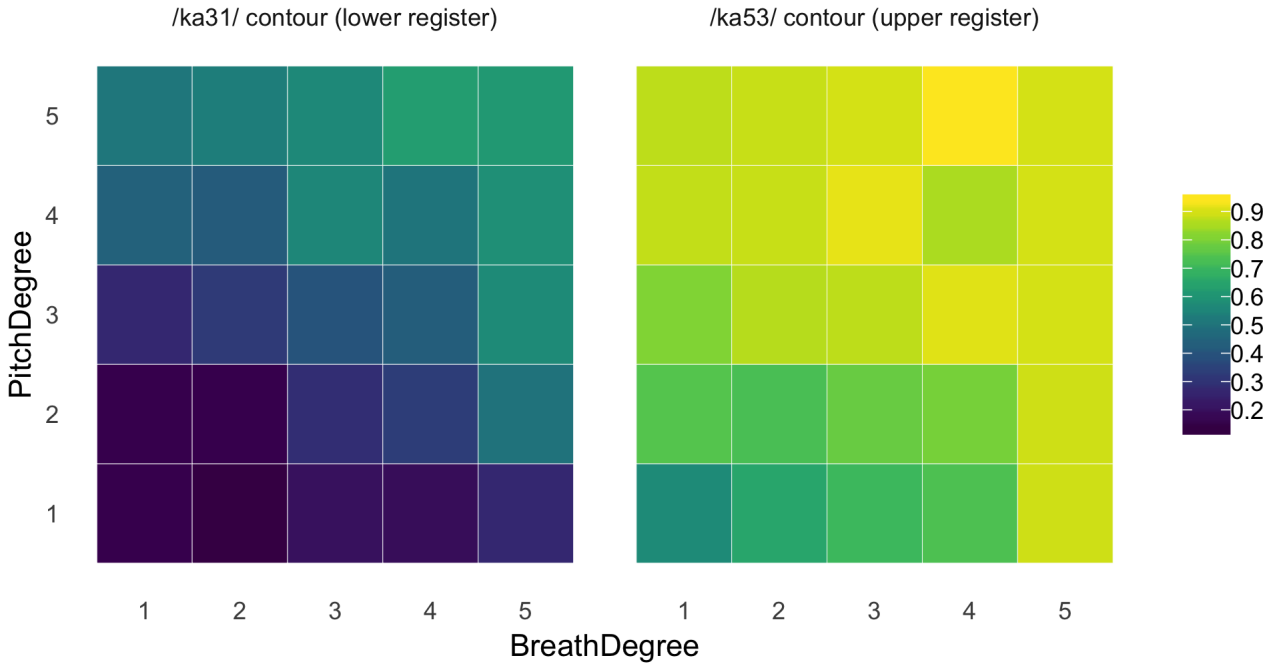


Figure 2.4: Percentage of upper register response from Jiashan participants (axes as in Figure 2.3).

Heat maps of the empirical data also reflect this pattern (Figure 2.4). It is obvious that contour has a large effect on the perception of tonal register for Jiashan listeners. When listening to the upper register contour (right panel), listeners consistently hear it as an upper register word. The heat map also shows that pitch and breathi-

ness affect categorization, more so for the lower register contour, as captured by the significant interaction terms between breathiness and contour (β = -0.35, SE=0.18, p = 0.04) and between pitch and contour (β = -0.62, SE = 0.19, $p < 0.001$), which predict lower breathiness and pitch effects for the upper register contour.

The significant breathiness by pitch interaction (β = -0.69, SE = 0.2, $p < 0.001$) suggests that the effect of breathiness is smaller at higher pitch than at lower pitch. The right panel reflects this prediction in that there is a larger change in color at step 1 of the pitch continuum (the bottom row) than at step 5. As with the Shanghai responses, these interactions likely reflect the fact that the continua were not symmetric in how ambiguous the steps were. In particular, the lower register contour seems to have been more ambiguous than the upper register contour.

Summary of cue weights in both dialects

In summary, both groups of participants use all three cues, although both treat breathiness to be least important. Pitch appears to be more important than contour for Shanghai listeners, while contour is the primary cue for Jiashan listeners. In addition, for Jiashan listeners, breathiness and pitch have a smaller effect with the upper register contour; the effect of breathiness is also smaller at higher pitches for Jiashan listeners but larger at higher pitches for Shanghai listeners.

It is worth noting that although the cue weighting only reflects listeners' perception of the stimuli used in this task, the stimuli for the two dialects differ in ways that are representative of the communities speaking the dialect. Moreover, in terms of the manipulation of the three cues, while pitch and contour share the same range between the two endpoints for the two dialects, the range of breathy-modal distinction is smaller in the Shanghai stimuli (see visualizations in the supplemental materials). As a result, there also exists a difference in the degree that the breathiness cues were manipulated between the two dialects. It should also be noted that the two pairs

of tones used in Experiment 1 are confounded in contour (falling for Jiashan, rising for Shanghai), which further limits the interpretation of difference in cue weighting. We next present Experiment 2, which investigated the dialectal difference by having all participants listen to the same stimuli, in contrast to Experiment 1.

2.3.2 Experiment 2: Group results

This section extends the previous experiment by directly comparing the two dialect groups. In this experiment, listeners listen to the same stimuli (i.e. checked tone) produced by talkers from both dialects. Note that checked tones are marked with much shorter syllable duration and overall flatter contour (as shown in Figure 2.2 and Table 2.2), which may reduce the use of the contour cue. Checked tones are used for Experiment 2, however, because they are the only pair of tones sharing the same phonological representation between the two dialects (see Table 2.1). When exposed to stimuli differing in the degree of phonation contrast, we expect listeners who are sensitive to breathiness to show smaller breathiness cue weights for Shanghai stimuli with a weak breathy-modal distinction, than for Jiashan stimuli which differ more in breathiness.

Experiment 2 also clarifies one remaining problem in Experiment 1. Specifically, some participants relied almost solely on contour (see section 3.3 on individual differences below). One possibility is that the effects of pitch and breathiness were masked by a dominant contour cue and weren't observed because of the lack of an intermediate/ambiguous contour value. By breaking down the contour cue into a five-step continuum (while holding the other two cues at mid-point), Experiment 2 better examines how these listeners use all three cues.

The data were again fitted with a Bayesian mixed-effects logistic regression model in R using the blme package (Chung et al., 2013). As with Experiment 1, the depen-

dent variable was the register response (upper = 1, lower = 0). Independent variables were breathiness, pitch, and contour (continuum step as a numerical variable from 1 to 5) and their interactions with talker dialect and participant dialect (for both talker and participant dialect, Jiashan was coded as 0 and Shanghai was coded as 1) as the effect of the cues may vary by talker and participant dialect. Random effects include by-participant random intercepts, and random slopes for the talker dialect and the three cues. Correlations among all random slopes were included. All predictors were centered and standardized by subtracting the mean and dividing by two standard deviations, as in Model 1 (making Jiashan = -0.5 and Shanghai = 0.5 for the talker and participant dialect variables).[8] Table 2.5 shows the fixed effects of the model.

	Estimate	Std. Error	z value	Pr($>\|z\|$)
Intercept	0.48	0.11	4.28	$<$ **0.001**
Breath	0.43	0.00	4.74	$<$ **0.001**
Contour	0.35	0.08	4.19	$<$ **0.001**
Pitch	3.14	0.19	15.92	$<$ **0.001**
talkerDialect	–4.34	0.27	–15.73	$<$ **0.001**
listenerDialect	1.15	0.22	5.26	$<$ **0.001**
Breath×talkerDialect	0.24	0.17	1.38	0.165
Contour×talkerDialect	0.63	0.18	3.37	$<$ **0.001**
Pitch×talkerDialect	0.86	0.22	3.82	$<$ **0.001**
Breath×listenerDialect	–0.43	0.16	–2.62	**0.008**
Contour×listenerDialect	–0.25	0.14	–1.67	0.091
Pitch×listenerDialect	–0.87	0.38	–2.26	**0.023**
talkerDialect×listenerDialect	–0.18	0.54	–0.34	0.720
Breath×talkerDialect×listenerDialect	0.97	0.31	3.05	**0.002**
Contour×talkerDialect×listenerDialect	–0.13	0.34	–0.39	0.69
Pitch×talkerDialect×listenerDialect	0.42	0.41	1.02	0.30

Table 2.5: Summary of the fixed effects of the model.

[8]Model syntax: `bglmer (response ~ (Breath.std+Contour.std+Pitch.std) * dialect.std * part_dia.std+ (1 + (Breath.std+ Contour.std+ Pitch.std) * dialect.std|participant), data=df.rs, family="binomial", control= glmerControl( optimizer = "bobyqa", optCtrl=list(maxfun=100000)))`

The results show that all the main effects are significant. This means that all three cues, although redundant, are important in perceiving the tonal register contrast for this tone pair. Secondly, responses depended on listener dialect and talker dialect. Specifically, listeners tended to perceive more lower register words when they listened to a Shanghai talker (indicated by the negative talkerDialect coefficient) probably because the talker had an overall lower pitch range. Furthermore, Shanghai listeners tended to hear more upper register words (indicated by the positive listenerDialeect coefficient). However, of main interest were the interactions between the cues and the talker and listener variables.

We found that talker dialect influenced use of contour (β = 0.63, SE = 0.18, $p < 0.001$) and pitch (β = 0.86, SE = 0.22, $p < 0.001$), indicating that listeners were more influenced by both contour and pitch when listening to the Shanghai talker. Listener dialect interacted with breathiness (β = -0.43, SE = 0.16, p = 0.008) and pitch (β = -0.87, SE = 0.38, p = 0.023), which reveals that Shanghai listeners were less influenced by breathiness and pitch. Moreover, there was a significant three-way interaction between talker dialect, listener dialect, and breathiness (β = 0.97, SE = 0.31 p = 0.002) indicating that listeners were more influenced by breathiness when listening to their own dialect than when listening to the non-native dialect, at least for Shanghai listeners (see Figure 2.5).

Given that the model contains many two-way and three-way interactions, and the cue weighting of specific listener-talker combinations require taking these interactions into consideration, it is useful to calculate the coefficient estimates for usage of each cue as a function of talker and listener dialects. Table 2.6 below (also visualized in Figure 2.5) shows the coefficient estimates (with standard errors) of the three cues for different talker-listener combinations, calculated using the *emmeans* package (Lenth, 2018) in R.

Talker	Listener	Pitch	SE	Breath	SE	Contour	SE	Cue weight
JS	JS	3.26	0.26	0.77	0.16	0.13	0.14	P>B>C
JS	SH	2.16	0.25	–0.14	0.21	-0.04	0.19	P>B≥C
SH	JS	3.91	0.37	0.52	0.15	0.84	0.18	P>C>B
SH	SH	3.24	0.35	0.57	0.13	0.52	0.15	P>B≥C

Table 2.6: Cue weights (coefficient estimates) of the three cues for different talker-listener combinations, Experiment 2. SE: standard error of coefficient.

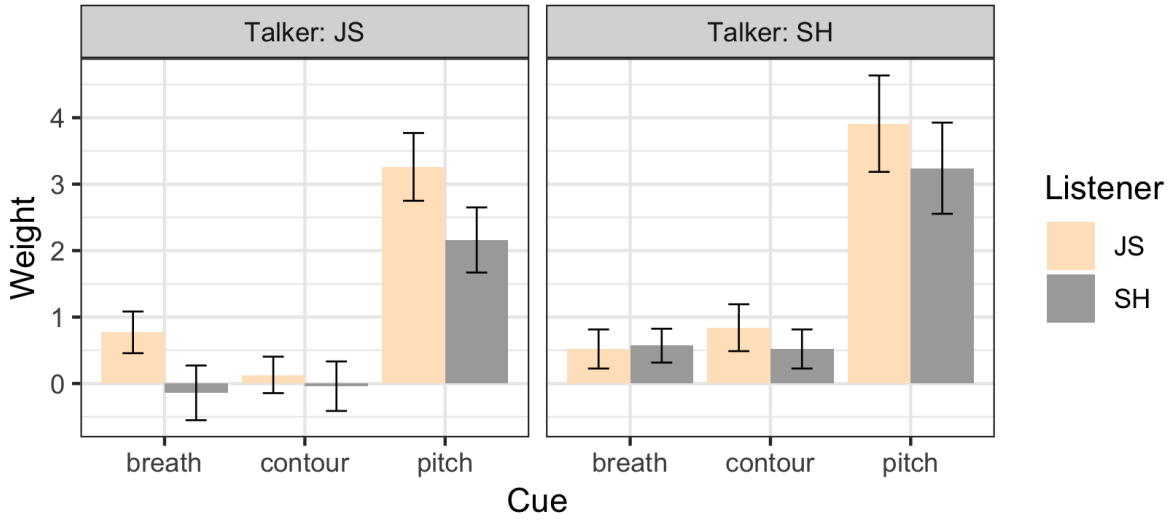


Figure 2.5: Cue weights for Jiashan (JS) and Shanghai (SH) talker by listener with error bars showing 95% speaker variability intervals, capturing individual variation in cue weights. Experiment 2.

Pitch is the primary cue for all talker/listener dialect pairs, being much more important than the secondary cues (breath, contour). This suggests that when listening to checked tones, both Shanghai and Jiashan listeners mainly use pitch to distinguish the two registers. This is in contrast to Experiment 1 in which only Shanghai listeners used pitch as the primary cue. Secondly, when listening to the Jiashan talker, the Jiashan listeners had on average a positive cue weight for breathiness while the Shanghai listeners had on average a coefficient close to zero. Thus even when listening to the same stimuli, Jiashan listeners' responses are more affected by

breathiness. Finally, when listening to the Shanghai talker, Jiashan listeners had a larger coefficient for contour (their secondary cue for that contrast) than the Shanghai listeners. Thus, regardless of the stimuli, Jiashan listeners seem to have larger secondary cue coefficients. In fact, comparing within stimuli and dimensions (e.g. Jiashan talker pitch) in almost every case the Jiashan listeners have larger coefficients than the Shanghai listeners for the same stimuli and dimension.

Furthermore, Jiashan listeners appear to rely more on breathiness than Shanghai listeners as indicated by the change of cue weighting for different talker dialects. When exposed to a larger breathy-modal contrast (Jiashan Wu talker), Jiashan listeners indeed had a larger coefficient for breathiness than they did when listening to Shanghai Wu. Shanghai listeners on the other hand, had a smaller coefficient for breathiness when listening to Jiashan Wu talker, the opposite tendency to what one would expect given the greater role of breathiness in Jiashan Wu in production. Thus, Shanghai listeners did not use breathiness when they were actually exposed to a larger breathy-modal contrast, while they managed to better perceive such a contrast in their own dialect with a smaller difference.

To summarize, Jiashan and Shanghai listeners share a similar order of cue preference for both talkers, with pitch always being the primary cue. However, Jiashan listeners' coefficients more closely relate to the amount of breathy-modal contrast in the stimuli than Shanghai listeners. Thirdly, both listeners show greater use of voice quality cue when listening to their own dialect.

2.3.3 Individual variability

Experiment 1

Group-level results found in Experiment 1 indicate that while all three cues significantly affect both groups of listeners' tonal register categorization, Shanghai listen-

ers' perception was more restricted to a single cue (pitch) while Jiashan listeners on average had a more balanced cue weighting. What is not clear from the group results is whether these average patterns were an equally good description of individuals in the two groups. In the Introduction, we hypothesized that the ongoing loss of breathiness in Shanghai Wu might lead to inter-listener variability. On the other hand, having a contrast that depends on more cues might lead to greater listener variability in the Jiashan listeners. In order to test the two hyposthesis, we compare variability in individual cue weights between the two listener groups. In addition, we wanted to test if there are patterns across cues in how individuals use them. In other words, does having a large primary cue weight predict having a large secondary cue weight, or does it predict having a small secondary cue weight? Examining the structure of the variability between cues allows us to test these competing hypotheses.

The random effects fit in the models in the previous section, summarized in Tables 2.7 and 2.8, allow us to examine individual variability in both the ways detailed above. Specifically, random effects are the part of the statistical model that captures the variability within groups (in this case, the variability of cue weights among different participants). We can examine the standard deviation in estimates for each cue ("SD" column), which capture the degree of variability in cue weights across participants, as well as the correlations between cues (across participants).

Figure 2.6 summarizes the variability across individuals in the weight of each of the three cues, for each listener group. For each cue, the fixed-effect estimate from the model plus or minus twice the random effect standard deviation gives the range of cue values predicted for 95% of speakers, which we call the 'speaker variability interval'. As seen before for the group-level pattern, the importance of cues within dialect varies, and furthermore, Jiashan listeners consistently show larger variability than Shanghai listeners for the primary (pitch for Shanghai, contour for Jiashan), the

Name	Variance	SD	Correlation		
Intercept	0.91	0.95			
Breathiness	0.17	0.41	0.50		
Pitch	3.91	1.97	0.61	0.83	
Contour	0.86	0.93	0.54	0.72	0.63
Breath×Pitch	0.61	0.78			
Breath×Contour	0.21	0.45			
Pitch×Contour	1.89	1.37			
Breath×Pitch×Contour	6.42	2.53			

Table 2.7: Summary of random effects terms Shanghai participants, Experiment 1. Correlations between interaction terms are not shown.

Name	Variance	SD	Correlation		
Intercept	0.34	0.59			
Breathiness	0.68	0.82	0.25		
Pitch	2.71	1.64	0.29	0.93	
Contour	6.86	2.61	-0.23	-0.12	-0.15
Breath×Pitch	0.46	0.67			
Breath×Contour	0.87	0.93			
Pitch×Contour	0.89	0.94			
Breath×Pitch×Contour	1.28	1.13			

Table 2.8: Summary of random effects and correlations of Jiashan participants, Experiment 1. Correlations between interaction terms are not shown.

second important (contour for Shanghai, pitch for Jiashan) and the least important cue (both are breathiness), with variance scaling with the mean.

For Shanghai listeners, pitch is always the most important cue for all participants, and higher pitch always leads to more upper register percepts. The findings are also reflected in the individual cue weighting plots (Figure A3 in Appendix) where pitch always has the highest weighting while the reliance on contour and breathiness varies by individual. For Jiashan listeners, plots of individual cue weights (Figure A3 in Appendix) further support the larger individual variability for these listeners. Unlike Shanghai listeners, Jiashan listeners show different preferences on the primary cue as well as the use of other cues. Moreover, plots of individual

responses also display larger variation for Jiashan listeners (Figures A1 and A2 in Appendix). It is worth noting that while there seem to be large differences in cue weighting, all participants (except for one participant, 62107, who displayed the opposite categorization) were successful in marking the contrast between the two natural tokens. This suggests that different individuals may have different perceptual strategies, an idea we explore further by inspecting the correlation terms.

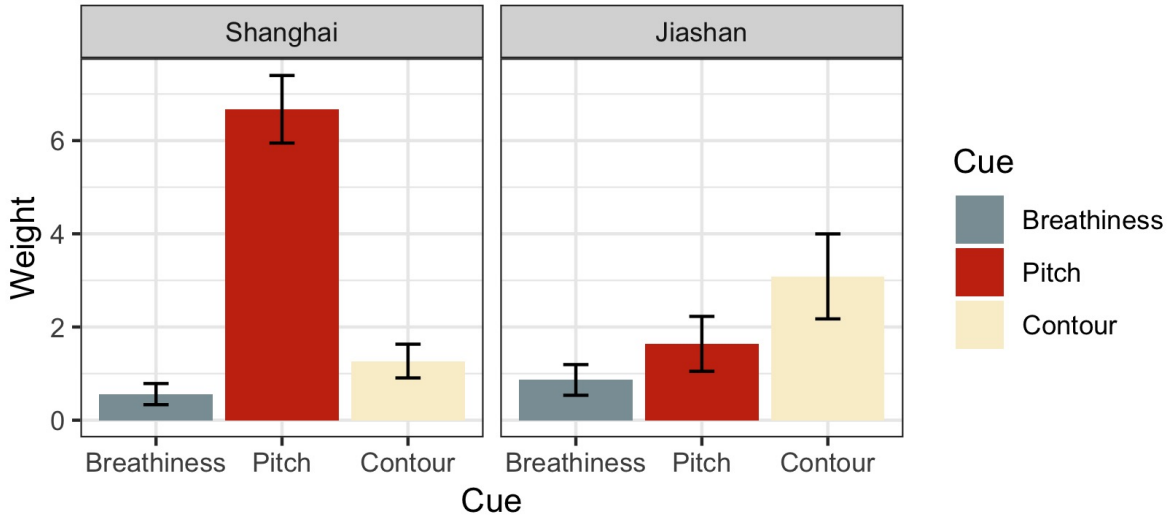


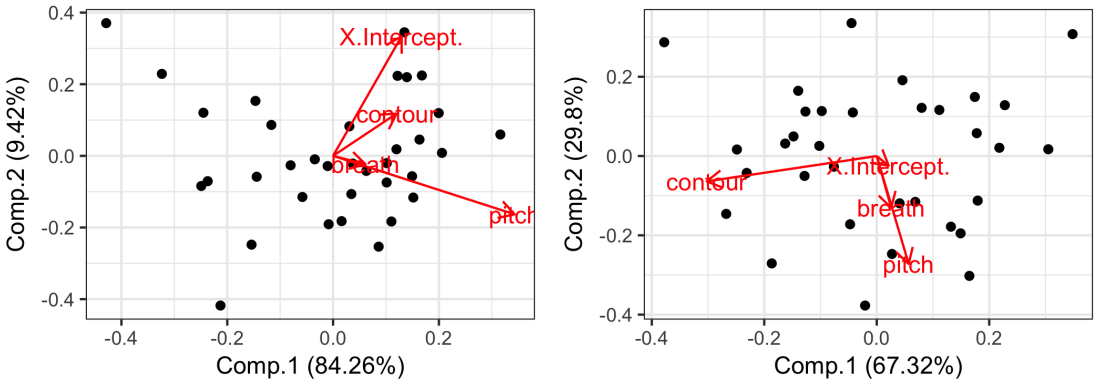
Figure 2.6: Group-level cue weights for Shanghai (left) and Jiashan (right) listeners with error bars showing 95% speaker variability intervals, capturing individual variation in cue weights. Experiment 1.

Tables and 2.7 and 2.8 indicate that the models fit some large positive correlations between the cues across individuals. In order to evaluate the significance of the correlations, we extracted the coefficient estimates from the main effects (no interactions) of the fixed effects and the random effects to produce the cue weights from each individual predicted by the model. This yields four estimates per individual, three slopes for three cues (breathiness, pitch, contour) and one intercept. We used non-parametric correlation tests using the Spearman method to test all correlations, since non-parametric models are robust to outliers. The results yielded

significant correlations among all three cues for the Shanghai listeners (breathiness and pitch: r=0.87, $p < 0.001$; breathiness and contour: r=0.87, $p < 0.001$; pitch and contour: r=0.67, $p < 0.001$) and only between pitch and breathiness for Jiashan listeners (r=0.94, $p < 0.001$). The relationship between the three cues and the intercepts for each language group is also visually presented in Figure 2.7. Figure 2.7 is the result of a Principal Component Analysis (PCA) on the weights calculated for each language separately. PCA is a technique to reduce the dimensionality of the data by transforming possibly correlated variables into a smaller number of uncorrelated variables called *Principal Components* (Jolliffe, 2003). The first principal component accounts for as much of the variability in the data as possible, and each subsequent component accounts for as much of the remaining variability as possible. Figure 2.7 shows that most of the variability in weights for the three dimensions and the intercept can be captured in two dimensions, i.e. the first two principle components (93.68% for Shanghai and 97.12% for Jiashan) and that the two language groups differ in how closely contour patterns with the other two cues and the magnitude of variation in the intercept.

The results show that despite the fact that pitch and breathiness go together (are positively correlated across listeners) for both groups, there are different patterns for the two groups of listeners. Jiashan listeners have a competing primary cue (e.g. participant 62310 has pitch as primary cue and participant 73002 in the same figure has contour; see Figure A2 and A4 in the Appendix), where they may prefer contour or pitch, while for Shanghai listeners, both secondary cues are correlated, and are further correlated with the use of the primary cue (which is always pitch).

In summary, for Jiashan listeners, while contour is the primary cue in aggregate results, this pattern is not representative of all individual listeners. Whether or not a listener relies on contour is independent of their use of pitch and breathiness, while the importance of the other two cues is always positively correlated. For Shanghai



(a) **(b)**

Figure 2.7: Principal component analysis (PCA) for Shanghai listeners (left) and Jiashan listeners (right) cue weights in Experiment 1. Each dot represents the cue weights of one individual projected onto the first two principal components. The length of the arrows reflects the amount of variation of the term, and the angle between two arrows reflects how much the two terms are correlated (more acute means higher positive correlation; more obtuse means higher negative correlation; the closer to right angle, the more the terms are independent).

listeners, on the other hand, all three cues are correlated, with pitch and breathiness having a stronger correlation.

Experiment 2

Similar to Experiment 1, this section explores individual variability and the pattern of cue usage using standard deviations reported in the random effects of the mixed-effects models and per-individual coefficient estimates predicted from the main effects (no interactions) of the fixed effects and the random effects. For the purpose of examining the two groups of participants separately, two new models were fit, one for Shanghai participants and one for Jiashan participants, each using the cor-

responding subset of the data.[9] Tables 2.9 and 2.10 show the summary of random effects for Shanghai and Jiashan participants respectively.

Name	Variance	SD	Correlation			
Intercept	0.87	0.93				
Breathiness	0.13	0.37	0.57			
Contour	0.12	0.35	0.63	0.46		
Pitch	2.13	1.46	0.31	0.482	0.23	
talkerDialect	5.45	2.33	0.13	-0.18	0.49	-0.55
Breath×talkerDialect	0.35	0.59				
Contour×talkerDialectr	0.88	0.94				
Pitch×talkerDialect	1.34	1.15				

Table 2.9: Summary of random effects and correlations of Shanghai participants, Experiment 2. Correlations between interaction terms are not shown.

Name	Variance	SD	Correlation			
Intercept	0.62	0.79				
Breathiness	0.31	0.56	-0.71			
Contour	0.16	0.40	0.52	-0.02		
Pitch	2.19	1.48	0.13	0.40	0.59	
talkerDialect	4.45	2.11	-0.12	-0.06	0.49	-0.25
Breath×talkerDialect	0.98	0.99				
Contour×talkerDialectr	0.99	0.99				
Pitch×talkerDialect	1.56	1.25				

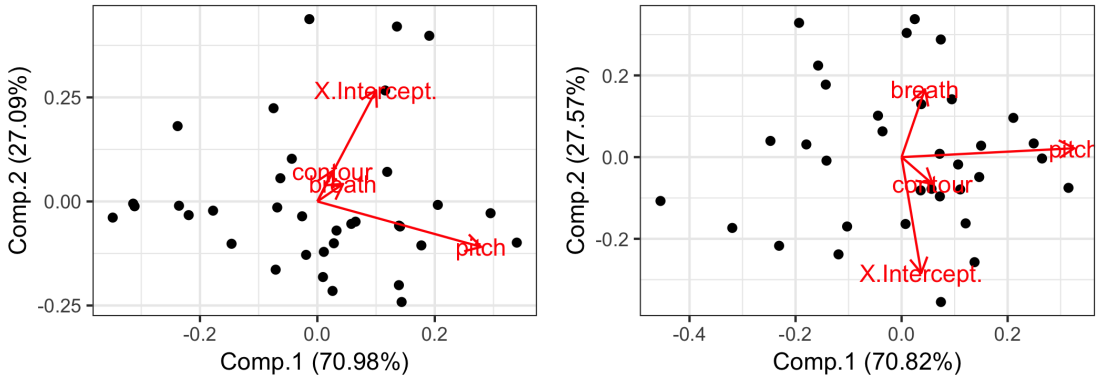
Table 2.10: Summary of random effects and correlations of Jiashan participants, Experiment 2. Correlations between interaction terms are not shown.

As shown by the standard deviations reported in the two tables, the magnitude of individual variability is similar between the two listener groups. Note also that there is large variation in the effect of talker dialect: listeners differ more in the effect of listening to different talkers than in the size of any cue.

[9]Using the random effects of the original model does not serve this purpose, since there is no random-slope term for listener dialect (there cannot be, as this variable is between-participant). Furthermore, it is impossible (for the class of model fit by blme) to let the random effects structure differ between subsets of the data according to listener dialect in this model.

In terms of correlations among cues, the positive correlation between breathiness and pitch is consistently found in both dialect groups and both experiments (Shanghai: r = 0.61, $p < 0.001$; Jiashan: r = 0.4, $p = 0.01$, correlation and significance are from the same Spearman test as Experiment 1), suggesting that listeners' reliance on breathiness is proportional to their reliance on pitch. However, there are some differences between the two experiments. While Shanghai listeners showed a positive correlation between all three cues in Experiment 1, they only show significant correlations between contour and breathiness (r = 0.58, $p < 0.001$) and between pitch and breathiness in Experiment 2. There is also a trend that all correlations are smaller relative to Experiment 1, with the smallest correlation (i.e. contour and pitch) in Experiment 1 being not significant in Experiment 2. Jiashan listeners, on the other hand, display a positive correlation between pitch and contour (r = 0.69, $p < 0.001$), which is not found in Experiment 1. Visualization of the Experiment 2 individual variation using PCA, shown in Figure 2.8, further confirms the results of correlations between individuals' cue weights. The small degree of variability in contour weight shown in Figure 2.8 probably arises because the checked tones in Jiashan Wu have even shorter duration than those in Shanghai Wu, which makes contour less salient.

In summary, pitch and breathiness are consistently correlated in a positive manner for both groups (although less strongly for Experiment 2). Different from Experiment 1, contour is not correlated with Shanghai listeners' primary cue (i.e. pitch), but instead with the other secondary cue, breathiness. Jiashan listeners, however, display positive correlations between the primary cue and the other two secondary cues.



(a) **(b)**

Figure 2.8: Principle component analysis (PCA) for Shanghai listeners (left) and Jiashan listeners (right), Experiment 2.

2.4 Discussion

This study investigates the role of less important cues (mainly focusing on voice quality) as a way to understand multi-dimensional contrast in Chinese Wu dialects. We compared two genetically-related dialects spoken in close proximity, at the group and individual levels. We considered three research questions, the first on the importance of the secondary cue and its (in)consistency cross tone pairs, the second on dialectal differences, and the third on individual differences and their implication for the relationship between multiple perceptual cues. We address the first question in 2.4.1 and 2.4.2 by discussing the performance of Jiashan listeners (2.4.1) and Shanghai listeners (2.4.2), and answer the second question in 2.4.2 by comparing their different perceptual strategies. We address the third question in 2.4.3, and address some limitations in interpreting the results in 2.4.4.

2.4.1 Jiashan listeners rely on multiple dimensions

The two experiments show that breathiness, while being a secondary cue to the multidimensional tonal register contrast, is nevertheless used in both Shanghai Wu and Jiashan Wu listeners. The difference between the two dialects is that Jiashan listeners' perception is overall more multidimensional with relatively more similar cue weights, while Shanghai listeners' perception is mostly dominated by pitch.

Jiashan listeners as a group have a multidimensional percept of register for the contrasts we studied. This is observed in their relatively large cue weights for the non-primary cues across both experiments, especially when listening to their own dialect. The smaller difference between the cue weights suggests that secondary cues play a bigger role in their perception than for Shanghai listeners.

Jiashan listeners also seem to be more context-dependent in how different cues are used, in this case depending on the particular tone. When comparing the cue weights for falling tones (Experiment 1) and checked tones (Experiment 2), Jiashan listeners use contour as the primary cue for falling tones, but contour is the least important cue for checked tones (likely because of the short syllable duration); meanwhile, the importance of pitch and breathiness is greater in checked tones relative to falling tones. Jiashan listeners also show context dependence in their use of secondary cues when listening to stimuli with different degrees of breathy-modal contrast (Jiashan vs. Shanghai talker). Specifically, Jiashan listeners rely more on breathiness when they listened to Jiashan Wu talker— with a larger phonation contrast— than when they listened to Shanghai Wu talker— with a less distinctive contrast in this dimension. It should be noted that these two sets of stimuli vary in more than just the voice quality dimension, as they are different talkers. Nonetheless, what is relevant is that the two listener groups treat the stimuli differently. Moreover, such variation of cue weights under different contexts and speakers/dialects further

shows that the secondary cues are indeed learnt from language-specific experience. Shanghai listeners, on the other hand, do not show such variation according to the degree of breathiness in the signal. In fact, their perception strategy appears to be more similar to listeners of languages with no phonation contrasts, a point which we turn to in the next section.

2.4.2 Shanghai listeners' perception is dominated by pitch height

Shanghai listeners' perception of the tonal register contrast differs from Jiashan listeners in that their perception is dominated by pitch, consistent with previous observations (e.g. Zhang and Yan, 2015). In other words, while Shanghai listeners rely on all three cues including breathiness, they exhibit an invariance in their cue weightings for all three sets of stimuli (i.e. Shanghai contour tones, Shanghai checked tones and Jiashan checked tones): pitch is always the primary cue and breathiness and contour are always the secondary cues. Furthermore, the secondary cues have quite small cue weights across experiments, especially when listening to the other dialect.

In terms of perception of breathiness, Shanghai listeners may be more like listeners of a language where voice quality is not a cue to lexical contrasts (the third type mentioned in section 1.2). They seem to rely on voice quality only for the purpose of facilitating pitch location (see Kuang and Liberman, 2015; **?**, for the influence of voice quality on pitch perception), while Jiashan Wu listeners behave like listeners of a typical language with a phonation contrast. Since there is limited use of breathiness in Shanghai Wu production, it is perhaps not surprising that Shanghai listeners are not very influenced by the breathiness dimension for the Shanghai talker; those stimuli likely do not have a very large differences in breathiness, much like stimuli used in past studies (Zhang and Yan, 2015). What is perhaps surprising is

that they also are not very influenced by the breathiness dimension for the Jiashan talker in Experiment 2. This provides evidence that these listeners may simply be less sensitive to breathiness, probably using it to facilitate pitch location, indirectly contributing to the tonal register contrast.

Another important piece of evidence supporting the different roles that voice quality plays in Shanghai and Jiashan Wu is the different influence of pitch on breathiness (the breath$\times$ pitch interaction terms in Tables 2.3 and 2.4) in the dialects. At lower pitch, breathiness has a smaller cue weight for Shanghai listeners (compared to higher pitch), but a higher weight for Jiashan listeners. This difference in direction may reflect how voice quality is treated differently in the two dialects: breathiness, which is always associated with lower pitch, contributes less to the register categorization in Shanghai, as low pitch is already indicative for lower register; on the other hand, high pitch is likely to be perceived lower when produced with breathy voice. In this way breathiness affects the perception of tonal register. The opposite pattern in Jiashan listeners, however, reveals that listeners are more likely to perceive lower register when lower pitch is accompanied by breathiness, suggesting that breathiness is not merely facilitative for pitch location, but directly influences register categorization. Nevertheless, we also acknowledge an alternative explanation, which may require further study: the Jiashan talker has an overall higher pitch range, so breathiness may be more important to identify the lower register, as pitch itself may be ambiguous, while breathiness is less important when the pitch is already high enough to identify a higher register. It is also possible that both these factors contribute to the different pitch-breathiness relationship observed for the two groups of listeners.

We found that the Shanghai listeners were not very influenced by the contour dimension either. However, unlike breathiness which in Experiment 2 was enhanced for the Jiashan talker relative to the Shanghai talker, Experiment 2 did not provide

very compelling contour cues in either set of stimuli, limiting our ability to assess use of contour by Shanghai listeners. It is also possible that there is an intrinsic difference between rising tones (heard by the Jiashan listeners of Experiment 1) and falling tones (heard by the Shanghai listeners of Experiment 1), as studies have found that Chinese listeners have smaller Just Noticeable Differences for falling tones than rising tones in terms of both pitch height and pitch contour (Jongman et al., 2016). Comparison of different tone pairs would be necessary to understand the contribution of contour to the perception of the register contrast in Shanghai Wu.

The above evidence indicates that Shanghai Wu listeners in general rely less on breathiness than Jiashan Wu listeners (and possibly contour as well), regardless of the acoustic signals they heard. Thus, their register contrast seems to be primarily based on pitch height. The decreased cue weight of the voice quality contrast in perception reflects the on-going loss of breathiness in Shanghai speakers' production of the tonal register contrast (Gao, 2016). One possibility is that the difference between the dialects is due to Shanghai listeners' relatively greater contact with Mandarin which does not employ a voice quality contrast (Gao, 2016). Another possibility is that Jiashan listeners' use of more cues in a context-dependent way compared to Shanghai listeners may be because Jiashan Wu has a relatively more complex tone inventory than Shanghai Wu, which may require multiple acoustic dimensions to distinguish between them.

2.4.3 Individual variability and processing of acoustic cues

We have two major findings on variability between individuals: the variability is highly structured, and there is more variability when the contrast is more multidimensional. First, we generally found positive correlations between coefficients of the different cues, and in many cases, these correlations were quite high (more

discussion below). Second, in three of our four contrasts (Shanghai rising tones in Experiment 1 and checked tones for all listeners in Experiment 2), the coefficient of the primary cue was much larger than the other two cues in the group data. For these contrasts, the individual data (Appendices A3, A5 and A6) showed that almost all listeners used the same primary cue (with only a few exceptions). In the fourth contrast (Jiashan falling tones in Experiment 1), the group data found that the coefficients for the three cues were more similar and in the individual data, listeners differed in *which cue* they used as primary. Inspection of the data in Appendix A4 shows that 20/33 participants had contour as the primary cue while 11/33 had pitch as the primary cue (the last participant had an ambiguous pattern). This may be due to the falling tone contrast having a more multidimensional nature, with each of the cues providing some, and likely redundant, information. Thus, listeners may come to different solutions to the problem.

We now turn to correlations between individuals' cue coefficients. As has been assumed in previous studies, we think it is correct to assume that group coefficient estimates indicate the degree to which the group uses particular acoustic dimensions. However, we argue that an individual's coefficient estimates are influenced by two factors: the relative importance of a cue with respect to other cues (i.e. the relative magnitude of cues) and the individual's ability to categorize speech stimuli consistently (i.e. overall magnitude, averaged across cues) (see also Kapnoula et al., 2017, for similar arguments). We argue that our analysis shows both of these components and that they have not been considered together before in analyses that report only group-level regression coefficients.

First, when we consider the individual-level coefficients, for both groups of listeners, whenever there is a correlation between two cues, it is always positive. We argue that positive correlations between individuals' coefficients is the default when perception is influenced by more than one cue. This is consistent with previous

findings for English showing correlations between regression coefficients for pairs of cues (Clayards, 2018; Hazan and Rosen, 1991; Kim and Clayards, 2019). Furthermore, both Hazan and Rosen (1991) and Clayards (2018) found correlations between the primary cue weights across different contrasts. Together these studies support the idea that some listeners are better able to extract and attend to acoustic phonetic details in a categorization task, regardless of the acoustic dimension. This leads to a more consistent stimulus-response relationship (i.e. sharper categorization functions) and thus larger correlation coefficients.

There are, however, several cases of no significant correlation between cues in the two experiments. Some of these are due to relatively small coefficients of the cues involved, which leads to reduced variability and thus weak correlations (e.g. correlations in Experiment 2 involving contour where contour played only a small role). The more interesting exception is the case of Jiashan listeners' use of pitch and contour in Experiment 1. As discussed above, in this experiment, both cues were important to the contrast and there was very large individual variability in individuals' cue weights. We think the reason for this apparent contradiction comes from the fact that listeners differed in which cue was primary, pitch or contour. Note that previous studies only examine two cues, in which case individuals rarely differ in cue ordering, but in absolute magnitude. Our study, however, focuses on three cues, which provides a venue to examine listeners' cue weights when both *cue ordering* and *weight magnitude* are different across individuals. For the Jiashan listeners in Experiment 1, by definition, those who had contour as a primary cue had smaller cue weights for pitch than contour while those who had pitch as a primary cue had smaller cue coefficients for contour than pitch. Thus, there must be some negative relationship between the cue coefficients for these two cues. We think that this negative relationship in the relative cue coefficients may have been offset by an overall positive relationship between cue coefficients due to individual differences

in response consistency (coefficient magnitude) that we observed in all of the other conditions.

In summary, we find that listeners vary in how consistently they respond to cues. Specifically, when listeners agree in cue ordering but only differ in weight magnitude, there is a clear positive correlation; when listeners differ in both cue ordering and weight magnitude, a positive correlation may be masked by their different choice of the primary cue, in which case the correlation pattern may not surface.

2.4.4 Limitations

One may worry whether or not the talkers chosen in the study are representative of the dialect, in other words, whether listeners' responses are specific to the talker, or generalizable for the dialect. We acknowledge that any choice of talker will introduce talker-specific acoustic characteristics that may not be representative for the dialect, and it is hard to verify which specific acoustic dimensions are idiosyncratic. The ideal case would be to use a number of talkers that could jointly be 'representative' of the dialect, although this would be impractical in a single study. In order to eliminate the possibility of talker effects, a follow-up study is needed to explicitly test how listeners react to different talkers. However, even though a single talker cannot stand for the dialect, the acoustics indeed reflect the difference we expect from the two dialects: the Jiashan stimuli have a larger breathy-modal contrast and the Shanghai stimuli have a smaller contrast.

2.5 Conclusion

This study has examined listeners' perception of a multidimensional tonal register contrast in two Chinese Wu dialects signaled by three cues: pitch height, voice quality, and pitch contour. We found that cue weights are context-specific, i.e. vary by

tone. While the cues are present across all cases, some contrasts are more multi-dimensional than others, as evidenced by cue weights of non-primary cues being bigger for certain tones. In terms of dialectal difference, while both groups of listeners rely on all three cues, Shanghai listeners have smaller cue weights of breathiness than Jiashan listeners. Shanghai listeners rely less on voice quality information even when listening to stimuli with a clear breathy-modal distinction, and their cue weights reflect their dialect-specific experience. Finally, we found structured individual variability: in most cases individual's cue coefficents were positively correlated and furthermore, there is more variability across individuals when the contrast is signaled by more than one salient cue, in which case individuals have different options for choosing the primary cue.

2.6 Appendix

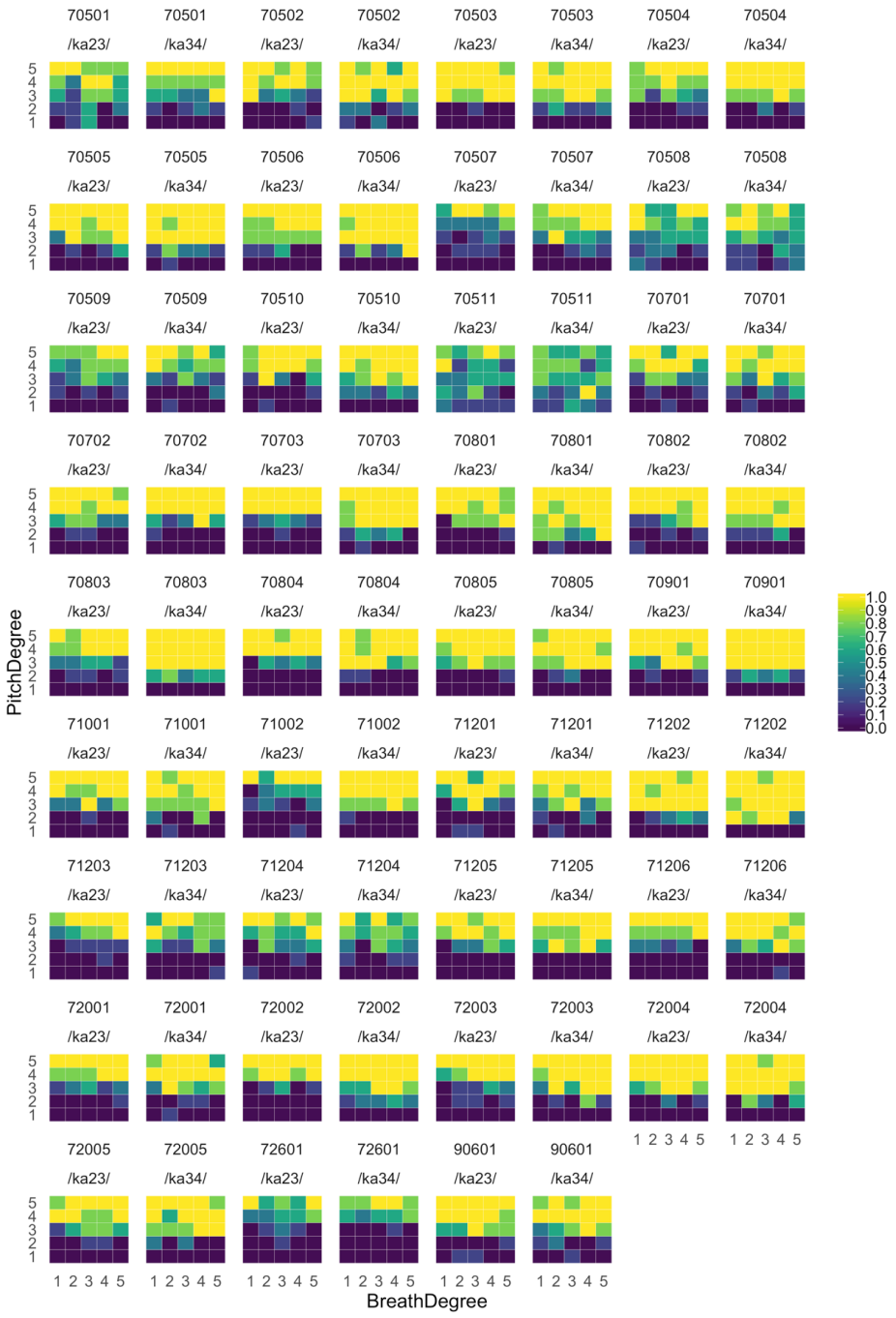


Figure A1: Heat maps for individual responses from all Shanghai listeners in Experiment 1. For each participant, the left panel is the /ka 23/ contour (the lower register contour) and the right panel is the /ka 34/ contour (the upper register contour).

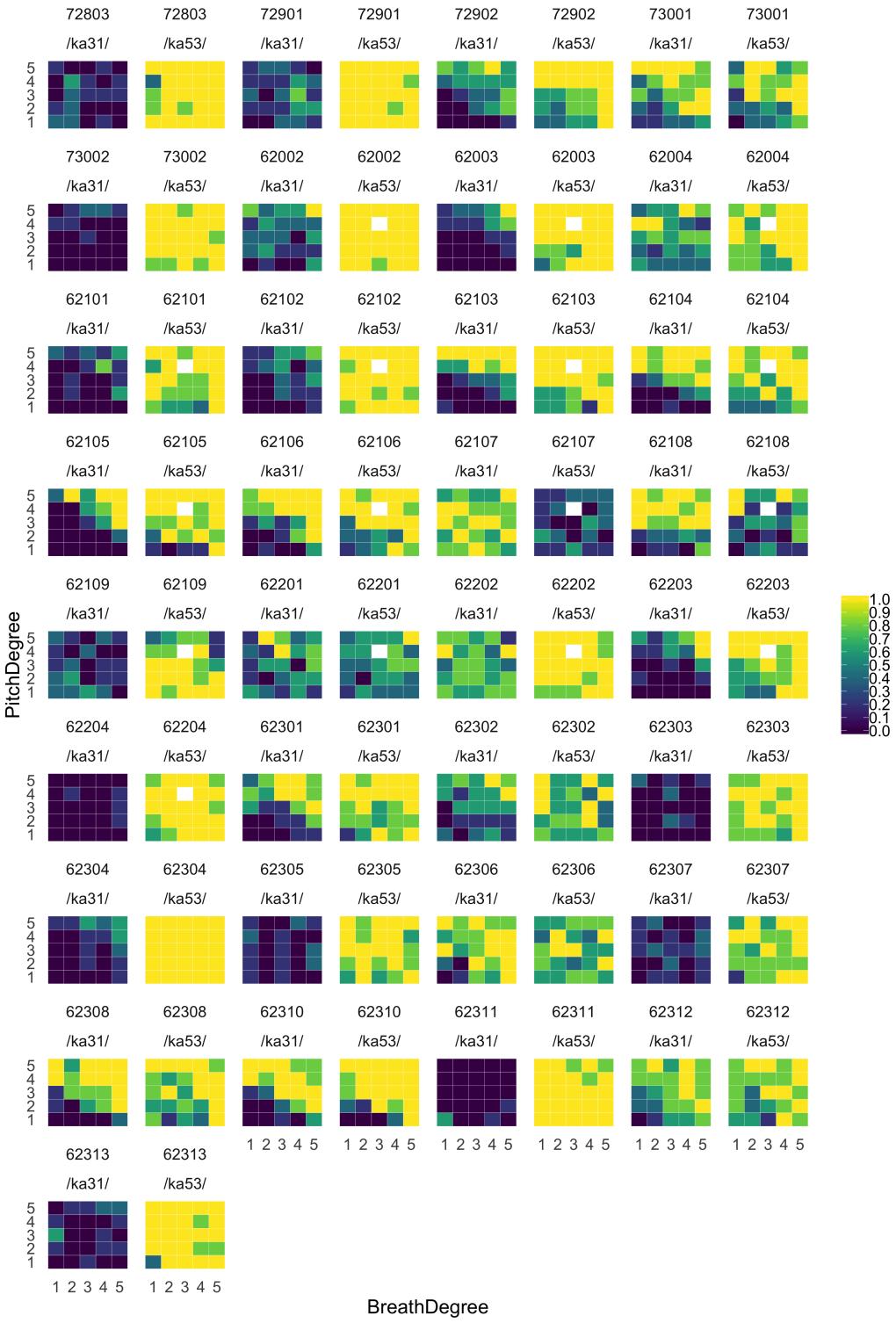


Figure A2: Heat maps for individual responses from all Jiashan listeners in Experiment 1. For each participant, the left panel is the /ka 13/ contour (the lower register contour) and the right panel is the /ka 35/ contour (the upper register contour).

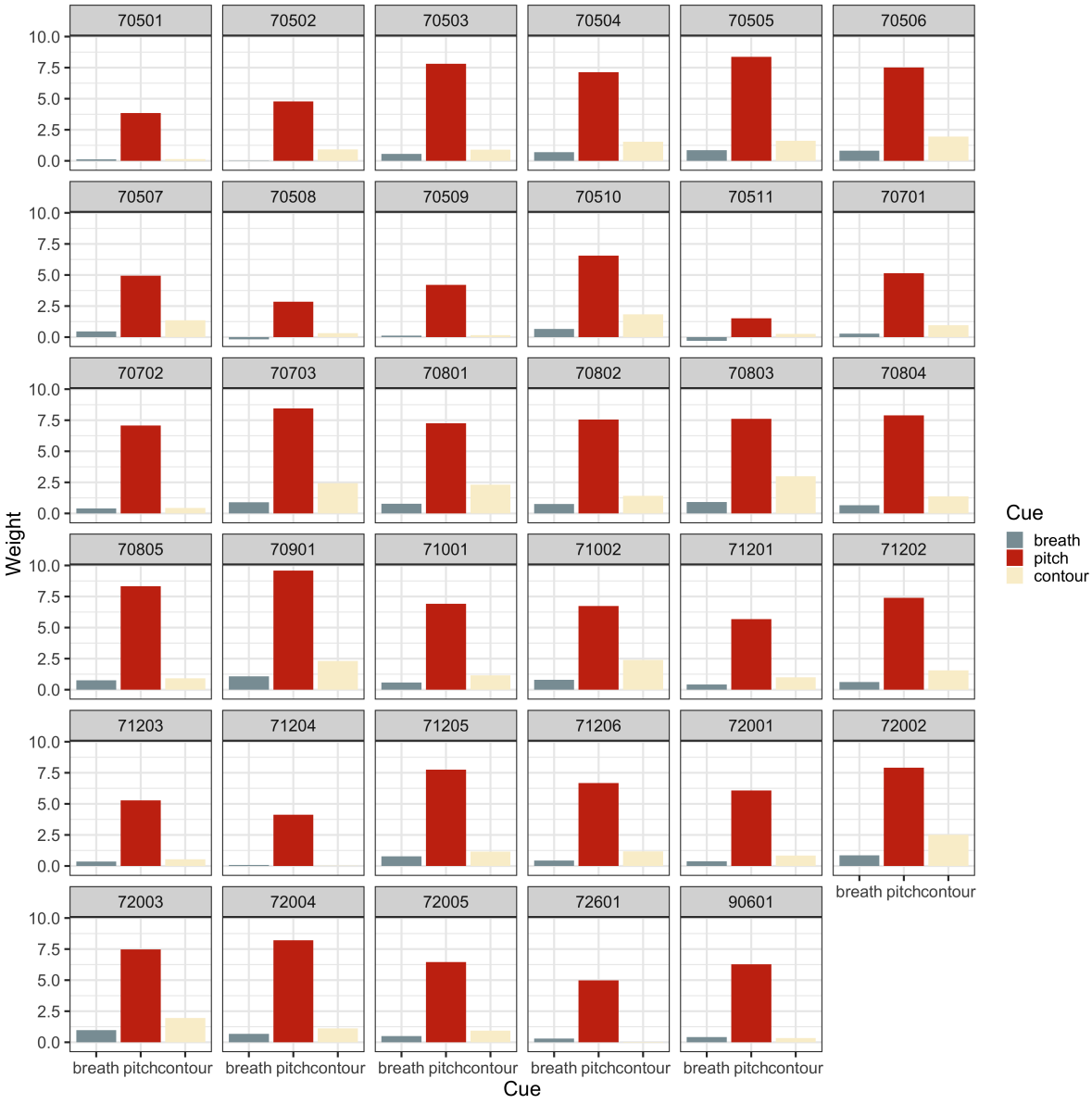


Figure A3: Individual cue weighting of Shanghai participants (Experiment 1).

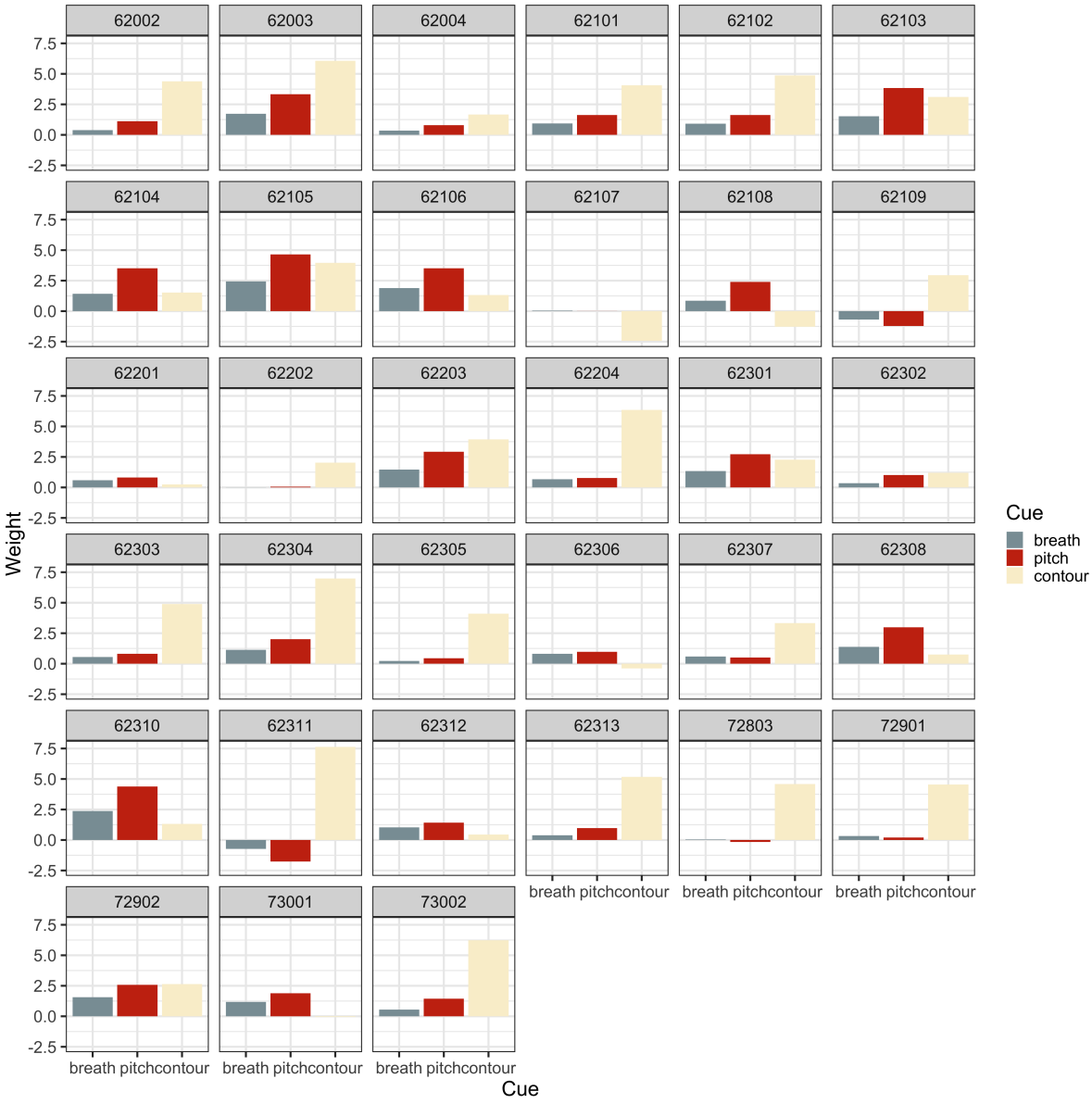


Figure A4: Individual cue weighting of Jiashan participants (Experiment 1).

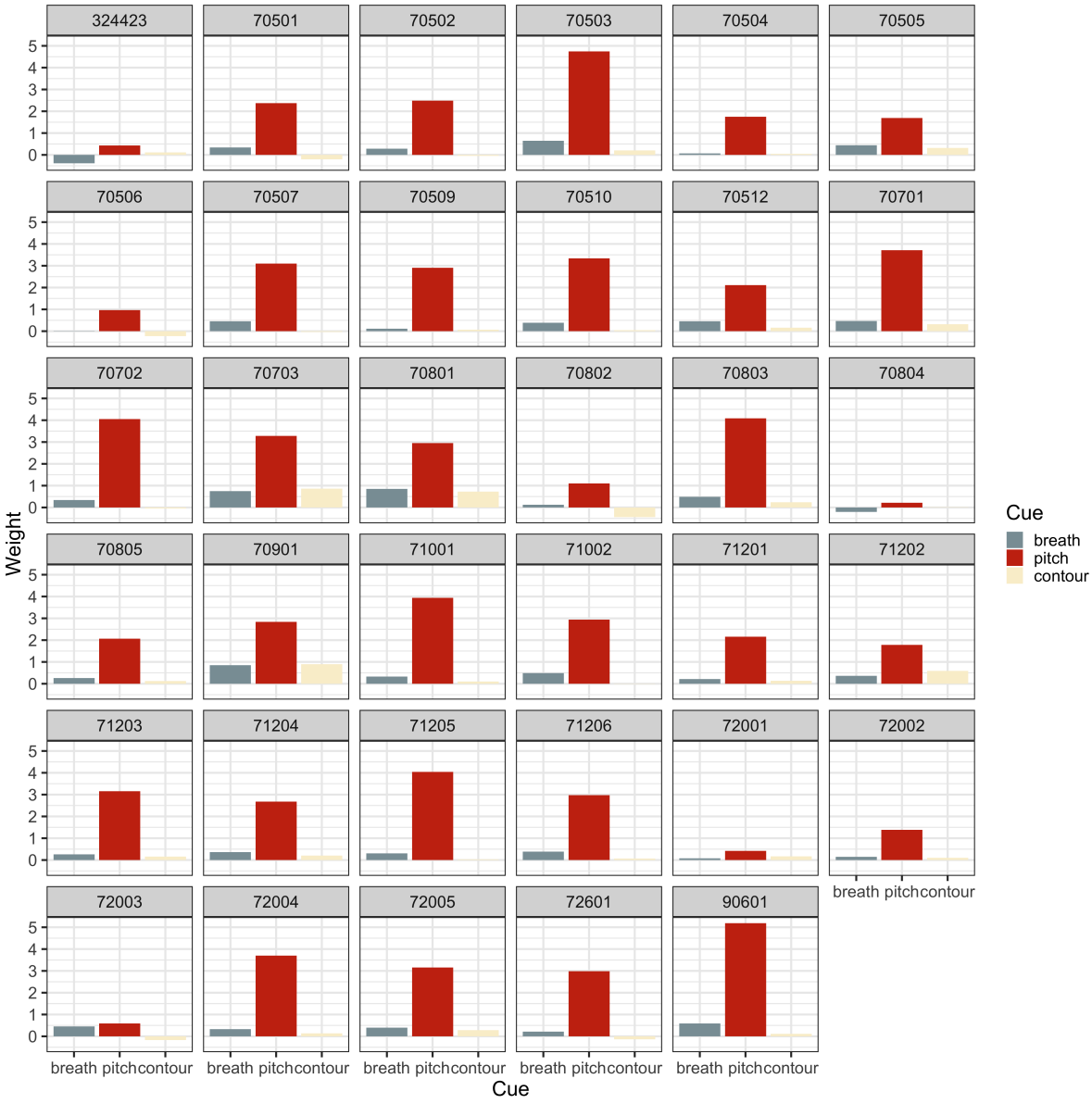


Figure A5: Individual cue weighting of Shanghai participants (Experiment 2).

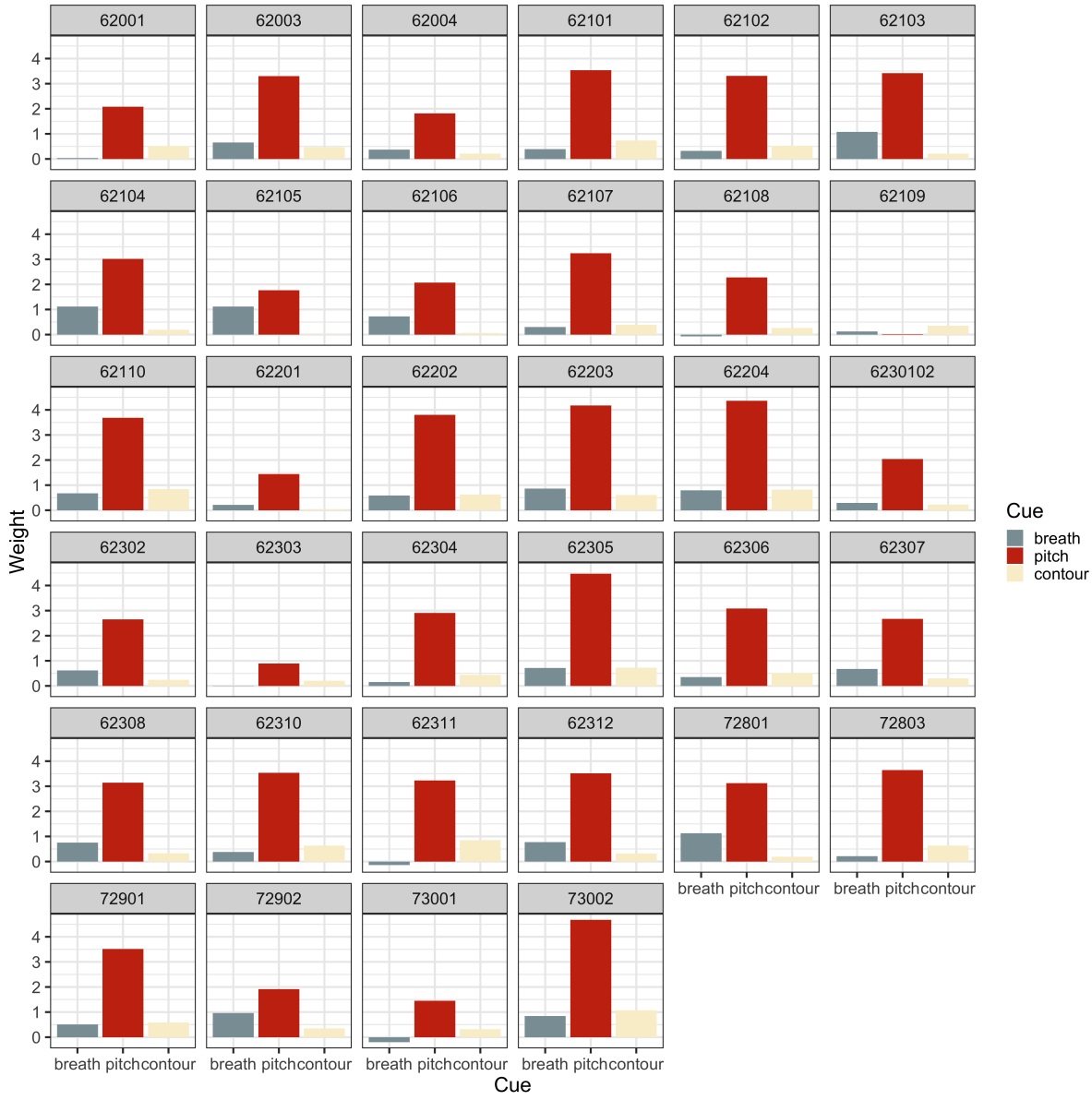


Figure A6: Individual cue weighting of Jiashan participants (Experiment 2).

Preface to Chapter 3

Chapter 2 investigates how multiple acoustic cues contribute to multi-dimensional tonal register contrasts and how dialectal experience shapes listeners' perceptual strategies. The results from the perceptual experiments reveal that listeners differ mainly in their overall cue acuity (e.g. there are listeners with flatter and steeper boundaries between sounds – across all cues). Moreover, listeners also differ in the primary cue they rely on, when the contrast is signaled without a dominant cue. Their perceptual strategies are further affected by their dialect background.

Chapter 3 examines a broadly similar topic, perception of tonal contrast in Chinese languages, using a different approach—computational modelling. This study focuses on the tonal contrast in continuous speech, where there is large variability in the phonetic realization of tones. The goals are to examine the importance of each cue in continuous speech, and to examine the structure of the tonal space induced by a computational learner, across multiple cues.

Chapter 3

Modelling perceptual tonal space in Mandarin Chinese continuous speech

3.1 Introduction

Mandarin Chinese is a tone language where tones are used to distinguish between lexical meanings for words containing otherwise identical segments (Duanmu, 2007; Zhang and Hirose, 2000). In order to make sense of an utterrance, listeners need to first recognize the tones to identify the words. Like other phonological categories, tones are variable and multidimensional, meaning that their acoustic realization depends on context and the information that speakers use to identify tones may be spread across multiple acoustic-phonetic dimensions. Traditionally, tones have been described using only pitch; however, various other acoustic cues are known to be involved in tone recognition. Thus, how multiple cues can contribute to the recognition and representation of tone is important for understanding tones themselves. The goals of this paper are 1) to use computational models to evaluate the quantitative importance for tone identification of different cues in the speech signal for Mandarin tone identification (Study 1), and 2) to better understand the 'percep-

tual' representation of Mandarin tones in continuous speech by an artificial learner (Studies 2 and 3).

In this paper we build on the vast literature on Mandarin tones. Our research contributes to the understanding of the recognition of tones in continuous speech through using a rich set of inputs - a minimal transformation of the acoustic signal - and examining the contribution of mulitple cues. Here we distinguish *properties of cues* from *cues*. We use *cues* in this paper to refer to sources of information available in the speech signal (e.g. pitch, a complex time series representation) or the linguistic context (e.g. the segments in a syllable rhyme), while *properties of cues* refer to functions or summary measures of the *cues* (e.g. the mean value of F0 over a rhyme). Previous work on tone production addresses the importance of (pre-computed) *properties of cues*, but not *cues*. Similalry, previous work on tone perception investigates the tonal representation focusing on a selection of *cues* and their *properties*. There is less work that provides a general account of how cues are combined perceptually, and how they might be represented in a perceptual space, which is abstracted from the acoustic space. Furthermore, both perception and production studies have tended to focus on tones produced in isolation, while it is generally less understood how tones are realized and perceived in continuous speech. In our work, we address these issues, building a model of tone classification from continuous speech and using it to study the importance of various cues and their organization in perceptual space.

This paper is organized into three studies. In Study 1, we evaluate the quantitative importance of three cues in classifying tones – pitch, intensity, and duration. Those cues have previously been found to be the most important for isolated tone recognition, and all have been shown to be useful on their own to human listeners in distinguishing tones (Coster and Kratochvil, 1984; Kong and Zeng, 2006; Whalen and Xu, 1992). We evaluate their importance using computational modeling, namely

a type of deep neural network (Long Short-Term Memory; LSTM) which handles variable length input, so syllables with any duration can be used as input. Moreover, we use the entire spectrum and F0, leaving the model to decide what aspects of the input are meaningful, rather than pre-encoding certain *properties of cues*, such as mean F0. Using this data-driven approach, Study 1 investigates the overall contribution of each cue to the tonal contrast (Goal 1), which serves as the first step to understanding the tonal space.

Study 2 and Study 3 focus on examining a low dimensional 'perceptual' representation of the tone space. In Study 2, we add a low-dimensional layer to the model from Study 1, forcing the model to learn a low-dimensional representation which can be understood as a perceptual tonal space for the model. This model can be thought of as an 'ideal listener' making the best classifications it can with the acoustic signal available. We determine the optimal number of dimensions for this low-dimensional representation empirically, providing insight into the geometry of the tone space. Finally, unlike perceptual experiments where the effects of cues can only be probed through listeners' responses, the ideal-listener approach allows us to perform follow-up experiments to probe the structure of the learnt representations more easily than for human listeners. While we do not argue that the model's representation *is* humans' perceptual tonal representation, the learnt representation nevertheless provides insights on how humans' perceptual tonal representation *could* be. We compare the representations learnt by the model to theories of tone representation from the literatures.

Finally, we perform a follow-up analysis in Study 3 to examine the organization of the pitch cue in particular in the two-dimensional space. Our results provide in-depth evidence that tones are represented as pitch height and pitch contour – rather than start and end points implied by traditional notation – and suggest how other cues may map onto this representation.

This paper makes several novel contributions. First, for all three studies, we use models that take as input the full continuous speech signal and do not make any simplifying assumptions. Second, we use a data ablation methodology inspired from perceptual experiments to asses the contributions of different cues to a (tonal) contrast (Goal 1) in a bottom-up manner, with no restrictions on *how* cues contribute. Third, we apply a dimensionality reduction method to infer the perceptual representation of the tonal contrast of an "ideal listener" (Goal 2), including the number of dimensions needed, and how to interpret individual dimensions in terms of acoustic cues.

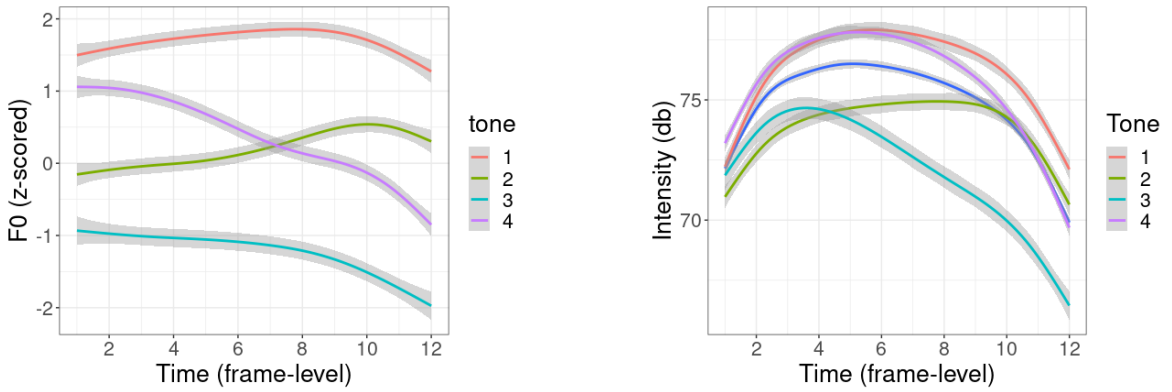
3.2 Cues for distinguishing tones in Mandarin Chinese

Mandarin Chinese has a four-tone inventory, described below in Table 3.1 using Chao numbers (Chao, 1930). In this system, 5 refers the highest pitch and 1 refers the lowest pitch, indicating the onset and offset (and turning point for T3) of the pitch trajectory: Tone 1 is a high level tone (55, which starts high and ends high), Tone 2 is a low rising tone, Tone 3 is low dipping with notation variants being a low tone (21) [1] (see Duanmu, 2007, for a survey and discussion of the variants), and Tone 4 is high falling. There are other notational variations to describe the tone inventory with Chao numbers being one of the most widely-used conventions (see Duanmu, 2007). While the numbers indicate the pitch trajectories of the tones, the transcriptions are phonemic representations, which do not necessarily reflect the acoustic realization of the contours, especially in spontaneous speech. Figure 3.1 shows the pitch trajectory and intensity trajectory of tones in the corpus of continuous speech used in the current study.

[1] Researchers disagree in terms of whether to represent Tone 3 underlyingly as a low tone or a low dipping tone. Typically, the low dipping realization happens more in the final positions, while the low realization happens more in non-final positions, such as the first syllable in a disyllabic word.

Tones	T1	T2	T3	T4
Chao number	55	35	21(4)	51
Description	High level	Low rising	Low (dipping)	High falling
Intensity	mid	mid	low; double-peak	high
Duration	mid	long	long	short

Table 3.1: Tone inventory in Mandarin Chinese.



(a) Pitch (z-scored)

(b) Intensity

Figure 3.1: The trajectories of (a) pitch (z-scored) and (b) intensity in the test set of the *Mandarin Chinese Phonetic Segmentation and Tone* corpus (Yuan et al., 2015) used in this paper. Both x axes represent normalized time (12 measures extracted over the rhyme at evenly spaced intervals) and the y axis represents (a) the z-scored F0 according to the training set and (b) the raw intensity (blue line refers to average intensity across four tones). Shading represents 95% confidence intervals.

As in many tone languages, Mandarin Chinese primarily uses pitch (here synonomous with F0) to distinguish tones (e.g. Howie and Howie, 1976; Moore and Jongman, 1997; Wang, 1967). However, a variety of other cues are also facilitative for tone distinction, such as intensity, duration, spectral information, and voice quality. The rest of this section gives a brief overview of the past studies on various cues signalling tonal contrast in both isolated tones and continuous speech.

3.2.1 Pitch

A number of properties of pitch have been investigated in the literature, including average pitch height, pitch contour (i.e. the shape of the pitch trajectory without taking the absolute values into consideration), the turning point in pitch trajectory, and the pitch at tone onset and offset. Both production and perception studies have shown that pitch height and pitch contour together facilitate tone identification when tones are produced in isolation (Gandour, 1984; Howie and Howie, 1976; Massaro et al., 1985; Tupper et al., 2020). Specifically, in a production study based on a corpus of monosyllabic word productions, Tupper et al. (2020) investigated 12 pitch properties of the tonal distinction, and found that average pitch height and pitch contour were the most important, while other correlated properties were redundant. Some studies have indicated that pitch contour may be more important a cue than pitch height. Leung and Wang (2020) conducted both production and perception experiments using isolated monosyllabic words to evaluate relevant *properties* of the pitch cue. They found that pitch contour has higher perceptual importance, and the degree of importance is correlated between production and perception across listeners, while pitch height is less important with no production-perception correlation.

While the height and contour properties of pitch are important, the temporal and frequency location of turning points (the dip) is particularly informative when distinguishing Tone 2 (henceforth T2) and T3 which have similar contours when produced in isolation. T2 is often realized with a dip in order to reach its low-pitch target, and hence is confusable with Tone 3. However, the dip realized in T2 is typically shallower (e.g. Moore and Jongman, 1997; Shen et al., 1993) and happens earlier (Moore and Jongman, 1997; Shen and Lin, 1991) than T3. However, this property is only important for isolated word productions, and is less relevant for continuous speech, as the dips are often not realized (see Figure 3.1).

In continuous speech, the realization of tones is more variable than in isolated tones due to linguistic and prosodic context. Tones can be influenced by adjacent tones through phonological processes or tone sandhi. For example, when T3 is followed by T3, the first T3 will be realized as T2 (see Duanmu, 2007, for an overview). Moreover, neighbouring tones also affect tone realization through carry-over effects and anticipatory dissimilation, found both phrase-internaly and across phrases (Xu, 1997). The realisation of tones further interacts with prosody. For example, when a syllable recieves focus prosody, its pitch range is usually expanded and the focus further affects the realization of the neighbouring tones (Chen and Gussenhoven, 2008; Xu, 1999).

In summary, pitch height and contour are found to be important to all tones produced in isolation, while the temporal and frequency location of the turning point can further help distinguish tones with similar contours. The realization of pitch is more variable in continuous speech.

3.2.2 Intensity

Intensity patterns vary for different tones, as summarized in Table 3.1: T4 tends to have the highest overall amplitude while T3 is the lowest (Chuang and Hiki, 1972), marked by a double-peak intensity pattern (Lin, 1988), with the dip showing low intensity.

In fact, various perception studies found that intensity alone is informative for perceiving tonal contrast. Listeners are still able to use the intensity information to distinguish among T2, T3, and T4, even when pitch and formants are removed and duration is neutralized (Whalen and Xu, 1992). When listening to amplitude modulated noise, listeners acheived 65% accuracy (Fu et al., 1998). Listeners with cochlear

implants, while having issues perceiving the pitch contour, can nevertheless use intensity information to distinguish tones (Meng et al., 2018).

When produced in a sentence context, Chang (2010) found that in sentence-medial position, the intensity contours are largely similar to those produced in isolation, except for T3 which has a falling intensity contour, instead of the double-peak contour. Focus prosody is also realized in intensity, but to a lesser extent than pitch (e.g. Yang and Chen, 2020).

3.2.3 Duration

Tones also differ consistently in their durations, summarized in Table 3.1. For isolated tones, T3 and T2 exhibit longer durations while T4 is the shortest (see review in Jongman et al., 2006), and some find T3 to be the longest (e.g. Yang et al., 2017). Blicher et al. (1990) showed that for isolated T2 and T3 which are similar in pitch trajectory, listeners tend to perceive longer tones as T3.

While the duration pattern is clear in isolated tones, it is less distinguishable in continuous speech. Yang et al. (2017) examined three types of speech: isolated monosyllables, formal text-reading passages, and casual conversations. They found that while listeners reach 65% correct tone recognition solely relying on duration information in isolated tones, they are only able to correctly identify 23% of tones in conversational speech. The tones are also much shorter in the two types of continuous speech than isolated tones, especially in conversational speech where the duration of the four tones overlap.

Duration also differs by the tone's position in the sentence, and according to whether or not it is focused (Nordenhake and Svantesson, 1983). Deng et al. (2006) analyzed the read speech of one speaker and found that tones in sentence-medial and sentence-final positions share the same pattern: T4 is the shortest while T2 is

the longest. However, the absolute duration of the tones are longer when they are both in word-final and sentence-final positions, compared to word-final sentence-medial positions. In contrast, Chang (2010) examined read speech (target words in carrier sentences) from 19 speakers and found that in sentence-medial position, while T2 is the longest, T3 was the shortest.

3.2.4 Miscellaneous Cues

In addition to the cues described above, spectral envelope cues are found to be informative in the absence of pitch, such as whispered speech (Kong and Zeng, 2006). Perception studies have also shown that while creaky voice is not always present in T3 production, it nevertheless facilitates T3 identification (Cao, 2012). Moreover, other than acoustic cues, it has been shown that some tones are more likley to occur with certain rhymes (e.g. T4 occurs often with /uai/ as in /shuài/ 'handsome'). Such segmental information is found to facilitate tone identification as well (Shuai and Malins, 2017; Wiener and Ito, 2015). Furthermore, vowel formants in identical syllables have been shown to vary with lexical tone both in terms of average formant frequency and formant contours over time (Erickson et al., 2004; Kong and Zeng, 2006) meaning that the supralaryngeal articulations themselves can encode the tones.

3.3 Study 1: Relative contributions of cues in continuous speech

3.3.1 Introduction

As summarized in the previous section, while there are extensive studies on examining the contributions of individual cues, our understanding of the quantitative importance of a number of cues is still limited. The relative importance of cues in continuous speech is even less well understood, however, continuous speech is what hear most often hear during speech communication. This section reviews previous studies on the relative importance of various cues and highlights how the current study advances our understanding of the tonal contrast.

Relative importance of cues

To our knowledge, studies examining the relative contribution of cues have been limited and mainly focusing on isolated tones. A recent study by Tupper et al. (2020) uses perceptual experiments and computational approaches to addresses a similar question by including 22 acoustic correlates (cue properties) of pitch, intensity and duration. Their statistical modelling results revealed that intensity and duration were not "distinctive acoustic cues" for the tonal contrast, nor were they found to be in the top five most important acoustic correlates used by listeners. Their sparse PCA analysis showed that pitch properties related to the trajectories are the most important, followed by intensity properties, while duration is the least important.

Fu and Zeng (2000) conducted perceptual experiments to examine the relative importance of periodicty (F0 alone with no harmonics), intensity and duration in isolated tones produced by six talkers. They created stimuli using signal correlated noise and different modifications (time normalization, low pass filtering and ampli-

tude normalization) to isolate each cue. They found that duration was the least effective cue. Listeners were only able to achieve 34.4% accuracy with duration alone and adding duration to other cues didn't improve accuracy. Intensity and pitch were both more effective. Listeners achieved 58.5% and 53.9% accuracy respectively.

Studies also found correlations between intensity and pitch (Fu and Zeng, 2000; Fu et al., 1998; Whalen and Xu, 1992). Fu and Zeng (2000) found that intensity contours were correlated with pitch contours in their productions of isolated tones, although such correlations showed large variability accross tones, speakers and syllables. Moreover, when testing their manipulated stimuli they found that productions with higher correlation are more accurately perceived by listeners. Similarly, Whalen and Xu (1992) found pitch-intensity correlations for T2, T3, and T4.

In summary, past studies on the relative contribution of pitch, intensity, and duration in ditinguishing isolated tones found that pitch is relatively important and duration is relatively less important, while there is no consensus for the role of intensity. Moreover, pitch and intensity are found to be correlated, and higher correlation contributes to higher identification accuracy in degraded stimuli.

Current study

Study 1 contributes to our understanding of the tonal contrast by 1) increasing the *scale* of both the speech data and the richness of the feature representation of the speech signals and 2) exploring the tone realizations in continuous speech. As summarised above, previous production studies mostly focus on particular properties of selective cues, which do not fully reflect the spectral and temporal richness of the speech. On the other hand, while perception experiments expose listeners with the full speech signal, they often use a limited number of stimuli. Moreover, it is less understood the importance of various cues in continuous speech, which are much more variable.

In the current study, we increase the richness of the speech data by examining a spoken corpus of 30 hours of speech (Yuan et al., 2015) and the richness of the signal by using MFCCs sampled throughout the course of the syllable as well as the full F0 pitch track. Past studies often use simplified measures for intensity compared to pitch, which may not accurately reflect the contribution of intensity. Intensity is found to be the most important in Fu and Zeng (2000), where listeners are exposed to full intensity trajectories, while it is much less important in production studies where intensity is only captured by a few properties. Equating the complexity of intensity to pitch is thus needed to better evaluate the relative contribution of the cues.

As for the tone realization in continuous speech, the studies we review in the previous sections reveal that all the three cues are realized differently compared to isolated tones. This variation from the canonical form, involving multiple cues, raises the general question of how similar cue weights in continous speech are to the more studied isolated syllables.

In order to test the relative importance of cues, we use a data ablation method inspired from perceptual studies to neutralize relevant cues in the corpus. We manipulate the cues directly in the audio of the training data so that the corresponding information is not available to the model. We compare predictions made by the model trained on the natural corpus and the models trained on the manipulated corpora. The degree of accuracy reduction when removing a single cue is taken to represent the importance of the cue to signaling tonal contrasts.

3.3.2 Methods

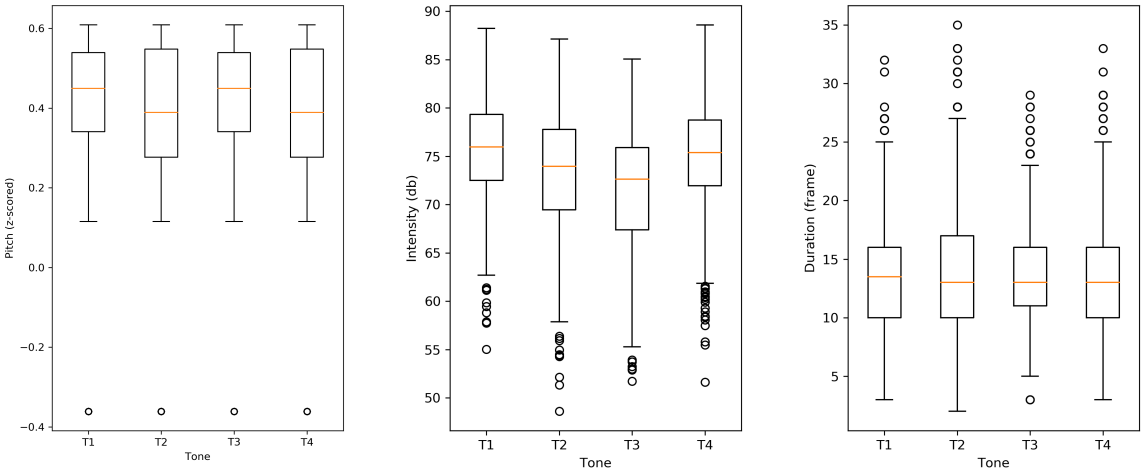
Corpus

We used *Mandarin Chinese Phonetic Segmentation and Tone* (Yuan et al., 2015), a corpus developed based on *1997 Mandarin Broadcast News Speech* (Huang et al., 1998). Note that for the tone annotations in the corpus, T3 is annotated as the surface form T2 when it underwent T3 sandhi. The test set provided by the corpus consists of 300 utterances from six speakers, and the training set consists of the remaining 7549 utterances with 10% of the syllables in the training set were split out as a validation set. Following Howie and Howie (1976), we treat the syllable rhyme (including the neuclei vowel and/or the nasal endings) as the tone bearing unit (e.g. /à/ in /dà/ 'big'; /àn/ in /bàn/ 'half') for each tone to extract MFCCs and F0 features and use individual rhymes as input to the model.

Table 3.2 summarizes the mean values of pitch, intensity and duration in the corpus for the training and test sets respectively. The distribution of the cues in the test set is visualized in Figure 3.2. The average pitch trajectories by tone are shown in Figure 3.1.

	T1	T2	T3	T4
F0 Train (Hz)	197.9	159.1	140.5	175.9
F0 Test (Hz)	209.9	169.6	153.4	184.3
Intensity Train (db)	75.6	73.9	72.0	75.0
Intensity Test (db)	75.7	73.3	71.5	75.0
Duration Train (frame)	12.5	13.3	13.3	12.5
Duration Test (frame)	13.6	13.8	13.7	13.0
Training tokens	20970	22266	14839	33086
Test tokens	748	760	507	1189

Table 3.2: Mean value of F0, intensity and duration summarized by tone in the corpus. Frames are extracted every 10 msec.



(a) Pitch (z-scored) **(b)** Intensity **(c)** Duration

Figure 3.2: The distribution of (a) mean pitch (z-scored), (b) mean intensity, and (c) mean duration of the tones in the test set. The orange lines indicate the medians.

Feature extraction

For each rhyme we extracted features of size t $\times$ 40, where t is the number of frames (indicating duration), and 40 refers to 39 MFCCs plus one pitch estimation. Specifically, the 39 MFCCs (the first 13 cepstral coefficients with Δ and $\Delta\,\Delta$, computed every 10ms using a window of length 25ms) were extracted using the *python_speech_feature* package (Lyons et al., 2020), and pitch values were estimated using Parselmouth (Jadoul et al., 2018). The 40 frame-level features were z-scored by subtracting the mean and dividing by two standard deviations based on the corresponding manipulated training set.

Data Ablation

We use *data ablation*, a method inspired by perceptual experiments to create testing and training corpora varying in the neutralization of different cues: pitch, intensity,

and duration. Each ablation results in one of eight unique conditions: 1) natural data; 2) data with pitch neutralized; 3) data with intensity neutralized; 4) data with duration neutralized; 5) data with both pitch and intensity neutralized; 6) data with both pitch and duration neutralized; 7) data with both intensity and duration normalized; 8) data with all three cues normalized. For each model trained, cues are neutralized for both training and test sets.

The cue neutralization was done using the following procedures. First, in order to neutralize pitch, the entire dataset was resynthesized using the Pitch Synchronous Overlap and Add (PSOLA) method (Moulines and Charpentier, 1990) implemented in Praat (Boersma and Weenink, 2019). F0 was constant at 200Hz throughout each tone, ensuring that pitch information was identical across all tones. Second, to neutralize intensity differences across tones, we used Praat to flatten the intensity contour to a constant 70db across all tones. Third, to eliminate the effect of durational differences, we normalized all tones to be 12 frames long (the mean of the original data set). We extracted 12 evenly spaced measures for each tone, while the specific interval between frames differed from tone to tone. [2].

Model

To classify tones, we used LSTMs (Hochreiter and Schmidhuber, 1996), a variant of recurrent neural networks (RNNs). We used LSTMs because of their general applicability to sequence processing problems and the fact that they easily handle variable length inputs (Graves et al., 2013).

The same model configuration, shown in Figure 3.3, was adopted for training all eight dataset conditions: a bi-directional LSTM with 1024 hidden state dimen-

[2]We chose 12 frames because it is the mean length of the training set when extracted every 10msec (default). The specific procedure of extraction was performed by first taking MFCCs and F0 estimation every 1ms. The total frames of each rhyme were then divided into 12 equal intervals and the first frame of each interval was kept, resulting in 12 frames for each tone instance.

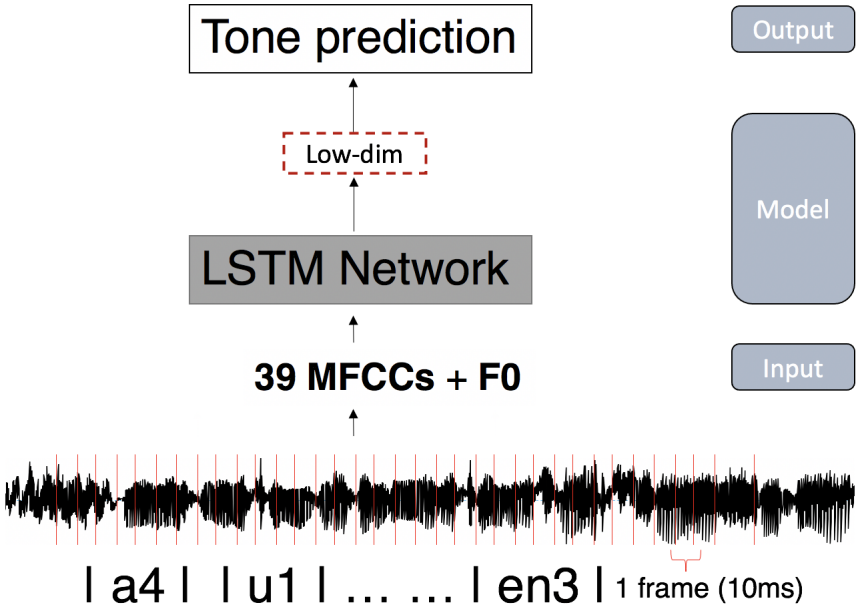


Figure 3.3: Illustration of the model. The low-dimension layer is only used for training the model in Study 2.

sions, and with the final hidden state output pushed through a 4-dimensional linear layer (corresponding to four tone classes) followed by a softmax layer to produce the probabilities of the four tones. For training, we used cross-entropy loss using the Adam optimizer, with dropout rate of 0.2 and batch size of 32. We trained the models in `Pytorch` (Paszke et al., 2019) until the validation loss failed to improve, and tested on the test set.

Note that the model and training scheme were chosen not for producing a state-of-the-art tone classification system, but for best addressing our research questions (Goals 1 and 2), and thus favour interpretability over accuracy. While it has been widely shown that the predicted accuracy of the individual token (e.g. phone, tone, word) is higher when using the entire sentence as the input, the context information actually introduces further confounds when examining certain cues, making it hard to distinguish whether the effect is introduced by the target tone or the surrounding

tone(s). Thus, in order to ensure that the learnt representation is explainable, we only feed the model the individual rhyme tokens witout context, at the sacrifice of accuracy.

3.3.3 Results

The models were evaluated by comparing performance on the tone categorization task. Table 3.3 below shows the weighted accuracy and weighted F1 scores of the models trained and tested on the eight different dataset conditions. As our data is not perfectly balanced for the four tones, it is necessary to ensure that high accuracy is not the result of always classifying the data as the most frequent category, thus we use weighted accuracy and weighted F1 scores to evaluate each model's predictions. Specifically, weighted accuracy is the accuracy weighted by the class size. The F1 score takes into account both recall and precision. For a specific tone *i*, recall calculates the number of correctly identified tone *i* results divided by the number of all tokens that should have been identified as tone *i*; precision calculates the number of correctly identified tone *i*s divided by the number of all predicted tone *i* results, including those not identified correctly. By incorporating precision into the measure, the F1 score better captures the quality of the model performance. After calculating the F1 score for each tone (tone *i* vs. non-tone *i*), the scores were then summarized by weighting according to each tone's frequency to produce the final weighted F1 score shown in Table 3.3. Each model's tone predictions were compared against random predictions generated proportional to the frequency of each tone, and all eight pairs (each condition vs. random baseline) were significantly different under the Wilcoxon test ($p < 0.001$).

A comparison of the models reveals that pitch is the most important cue, as neutralizing pitch results in the largest decrease in performance compared to the other

	Model	Accuracy (weighted)	F1 (weighted)
1)	Natural Speech	77.3%	77.3%
2)	– Pitch	67.2%	67.1%
3)	– Intensity	75.0%	74.8%
4)	– Duration	74.1%	74.3%
5)	– (Pitch+Intensity)	64.8%	64.8%
6)	– (Pitch+Duration)	60.3%	60.3%
7)	– (Intensity+Duration)	73.8%	73.7%
8)	– All	58.0%	58.0%
9)	Random baseline	25.1%	27.5%

Table 3.3: Test accuracy of data with different cues neutralized.

two cues. This is further confirmed by comparing the models with two cues removed: neutralizing duration and intensity (model 7) results in a relatively small decrease of accuracy, comfirming that pitch is indicative of tone category. On the other hand, intensity has the least impact, as neutralizing intensity only leads to a 1.6% drop in accuracy, and retaining only intensity (model 6) causes a sharp drop. Overall, neutralizing either one cue (models 2,3,4) or a few cues (models 5-8) consistently results in poorer performance, suggesting that all cues have independent predictive power. Yet, it is surprising that when all three important cues were removed (model 8), the model still achieved relatively high performance (58%). Some of this is likely due to additional acoustic information such as voice quality and formant information, which are indeed learnable from the MFCCs and can vary by tone (Erickson et al., 2004; Kong and Zeng, 2006).

Furthermore, the model could implicitly learn rhyme-tone correlations from the segmental information encoded in the MFCCs (Shuai and Malins, 2017; Wiener and Ito, 2015, 2016). In order to quantify the role of rhyme segments for tone classification, we did a follow-up test training a tone classification LSTM with trascriptions of the segmental information and without any acoustic information. The logic is the following: the model could solely rely on acoustics, or it could also be implicitly do-

ing phone recognition. By training a model that has explicit knowledge of phones but no acoustics, we can at least get a sense of what accuracy can be achieved from implicitly doing phone recognition under a best case scenario. Then at minimum, the difference between the two accuracy scores could be due to something in the acoustics that we did not test. The model achieved 37% accuracy, suggesting that segmental information is indeed informative to tone classification. The remaining 21% could be contributed by other acoustic cues unmanipulated in this study, for example, formants and voice quality.

While the results in Table 3.3 show the overall impact of cues on tone classification, they do not reveal whether certain cues impact some tones more than others. Table 3.4 presents the weighted accuracy and F1 scores broken down by tone showing the diferential impact of cues across the tones. First, neutralizing pitch affects all tones, especially T1 and T2. Second, neutralizing intensity affects T2 and T3 the most, yet results in no change for T1. Third, unlike the previous two cues, neutralizing duration shows different patterns for the two different accuracy measures. Only T2 and T4 are substantially affected in terms of weighted accuracy, while all tones are affected relatively equally in terms of F1.

Model	T1		T2		T3		T4	
	wAcc	F1	wAcc	F1	wAcc	F1	wAcc	F1
Natural Speech	85%	79%	85%	78%	78%	65%	85%	81%
– Pitch	-9%	-15%	-7%	-12%	-2%	-3%	-8%	-10%
– Intensity	-0.6%	0	-3%	-4%	-4%	-5%	-1%	-2%
– Duration	-2%	-4%	-2%	-3%	-1%	-3%	-3%	-2%

Table 3.4: Change of weighted accuracy (wAcc) and F1 score of predicting each tone.

3.3.4 Discussion

The results show that pitch is indeed the most important cue for the tonal contrast in most cases, and intensity and duration matter to different degrees between tones, consistent with previous findings. However, our results differ from previous studies in terms of the relative importance between duration and intensity. While previous studies found duration to be the least important (Fu and Zeng, 2000; Tupper et al., 2020), our findings show that duration is more important than intensity. Moreover, the fact that all sets of predictions produced by different datasets in Table 3.3 are significantly different from the random baseline shows that all three cues are informative for the tonal contrast on their own. The findings on the predictability of rhyme segments reveals that the model may implicitly learn the distribution statistics of the rhyme-tone co-occurrences, even though the models are not directly fed with transcriptions in the input. This section discusses the importance of the three acoustic cues as well as the segmental information, with regard to their impacts on classifying the four tones.

Pitch

While pitch is the cue that primarily affects the tonal contrast, its importance for tone prediction varies substantially by tone. In particular, T1's weighted accuracy decreases by 9.4% (F1 by 15%) while T3 only decreases by 2.2% (F1 by 3%) when pitch is neutralized. Recall that each time we manipulate the corpus, we train a new model so that the model can learn a new set of weights based on the acoustic information available. Thus, the fact that T1 is affected to the greatest degree by neutralizing pitch indicates that when pitch information is present, it is the most useful cue for classifying T1. Such reliance on pitch may result from T1 having generally less variation in pitch trajectory. It stays high and is only affected by tonal

co-articulation and specific prosodic or intonational interactions, making it a reliable cue. T3, being a canonical dipping tone while realized as a falling tone for the majority of the cases in continuous speech, shows more variations and is especially confusable with the other falling tone T4. The other reason that T3 has a larger reliance on intensity may be due to T3's more distinctive intensity contour, which we discuss next.

Intensity

Contrary to previous findings, we found that intensity is the least important. The disparity may arise from the type of speech: while previous studies finding intensity to be important used isolated tones (Fu et al., 1998; Meng et al., 2018; Whalen and Xu, 1992), our study used continuous speech.

Our results further show that the importance of intensity also depends on the tone. Specifically, for T1, intensity has little effect, while T2 and T3 are most affected. In terms of T1, the model hardly changes its prediction when intensity is neutralized, suggesting that intensity is not indicative for identifying T1 in continuous speech. T3, on the other hand, shows the opposite reliance on pitch and intensity to other tones. Among the three cues, intensity affects T3 the most. The reason for T3 being unique is likely twofold: 1) T3's falling intensity contour (Figure 3.1 (b)) differs from all other tones; 2) the similar falling pitch contour of T3 and T4 makes pitch less reliable in identifying T3. In fact, we found that 80% of the tokens originally predicted correctly as T3 in natural speech were predicted as T4 when intensity was neutralized.

In summary, intensity is an important cue for T3, which distinguishes it from T4, a tone with a similar pitch contour. However, intensity is not as important for other tones, and has little effect for T1.

Duration

While it may seem contradictory that the two measures, weighted accuracy and F1, show different qualitative results in Table 3.4, they actually reveal different aspects of the ways in which neutralizing duration influences the four tones. The F1 scores, which take into account both recall and precision, show that all four tones are affected by the duration neutralization. However, only T2 and T4 are substantially decreased in terms of weighted accuracy, which only captures the recall weighted by the tone's frequency. The two patterns suggest that neutralizing duration results in decreased accuracy in predicting T2 and T4, and more false prediction of T1 and T3 in general. Nevertheless, the effect of duration on tone classification is small. We return to the discussion of LSTM's ability to encode duration (i.e. sequence length) information in Study 2 on the mapping of duration and the low-dimensional representation.

Segments

Finally, for the segmental information, while we did not test directly whether or not the model indeed learnt segmental information from MFCCs, we found evidence that segment-tone co-occurrence may help tone classification. When trained using rhyme transcriptions with no acoustics, the model still achieved relatively high accuracy, which suggests that the model trained on acoustics could be learning about segments as well. This is consistent with previous literature showing that human listeners are also found to use such information (Shuai and Malins, 2017; Wiener and Ito, 2015, 2016).

Study 1 Summary

In summary, all acoustic cues and the rhyme segments are informative for tone classification. While pitch is still the overall primary cue in continuous speech, it is not always the most important for all tones. T3 is found to be an exception, which is more affected by intensity. One caveat is that these results do not take into account any contextual information. The role of some cues may have been increased if other sources of variation, like prosodic position, were taken into account in the model.

3.4 Study 2: low-dimensional perceptual tonal representation

3.4.1 Introduction

Study 1 examined the importance of various acoustic cues for the tonal contrasts in continuous speech. Study 2 further investigates a low-dimensional representation of Mandarin tones, focusing on the geometry and the linguistic interpretation of such a representation. We are particularly interested in how multiple cues are integrated in the low-dimensional space. While humans need to incorporate information from various cues to identify tones, it is hard to directly examine their perceptual representations. An alternative approach is to use a computational model, which does the same task as humans, with the learnt representation being its tonal representation. Unlike the human brain which can only be indirectly probed through responses, the computational model provides a learnt representation one can directly examine. While we do not argue that the model's representation *is* humans' perceptual tonal representation, the learnt representation nevertheless provides insights on how humans' perceptual tonal representation *could* be.

Earlier work has shown that neural networks can be used to learn low-dimensional latent representations. These representations retain the most important information for performing the downstream tasks, such as phoneme recognition (Weber et al., 2016). While these models are not designed specifically to represent human cognition, they still learn a representation with a similar linguistic structure as has been proposed for humans when doing the same task. Weber et al. (2016) examined a two-dimensional representation extracted from a phone classification model trained on vowels. They found that the dimensions of the representation corresponded to the first and second formants of the vowels – precisely the cues humans use to distinguish among vowels. Moreover, they found that the learnt representation of vowels are similar to the quadrilateral vowel space in phonetics. Other studies show that the clustering of phones using a learnt low-dimensional representation corresponds roughly to linguistic feature-based segmental categories such as stops and fricatives (Bai et al., 2018; Grósz et al., 2020). Given that segmental representations learnt by computational models are similar to human's perceptual space, we further wonder to what extent suprasegmental representation such as tones are learnt in a similar manner as humans. Specifically, we are interested in three aspects: the number of dimensions, the mapping between the cues and the dimensions, and the geometry (i.e. the organization of the tones in the representation space).

Previous studies have suggested that the Mandarin tone representation is two-dimensional (Chandrasekaran et al., 2007, 2010; Gauthier et al., 2007; Peng et al., 2012; Rhee and Kuang, 2020). Moreover, various studies, while differing in methodology and the acoustic measures they use, all find a quadrilateral space formed by the four tones, where T1 is adjacent to T2 and T4 and is opposite to T3. Chandrasekaran et al. (2010) asked listeners to discriminate pairs of tones and recorded reaction times. They used multi-dimensional scaling (MDS) of the RT data to construct a two-dimensional perceptual space of the tonal contrast. The space again had

the same geometry with T1 and T3 opposite from each other and T2 and T4 opposite from each other. They hypothesized that the two dimensions correspond to average pitch height and pitch contour: plotting the stimuli along these two acoustic dimensions matched the MDS geometry well. Rhee and Kuang (2020) examined tone F0 trajectories at different phrase positions and found similar quadrilateral organization of the tones. Morever, even when multiple cues (pitch, intensity, and duration) are measured, Tupper et al. (2020) found that a sparse PCA analysis resulted in the same quadrilateral space.

Together these findings suggest that the production and perception of Mandarin tones may be organized in a two-dimensional space, with a fixed quadrilateral geometry constructed by the four tones. We ask the question: if a computational learner is trained on a tone classification task and forced to learn a low dimensional solution, what tonal representation would it learn?

Current Study Study 2 explores the structure and the linguistic interpretation of a learnt low-dimensional tone representation. We do this by inserting a low-dimensional layer into the model trained on natural data from Study 1. We then visualize and investigate this low-dimensional layer as the model's tonal representation. First, we are interested in the number of dimensions needed in the low-dimensional layer for the model to accuractly recognize the tones. Second, we investigate whether and how the cues examined in Study 1 — pitch, intensity, duration, and rhyme segments — map onto the tone representation, and whether or not the mapping of the cues and the geometry of the tones are similar to humans. In order to explore the mapping of cues onto the tone representation, we use the same data ablation methods as Study 1 and observe the change of representation when each cue is neutralized. By comparing the low-dimensional representation before and after the ablation (i.e. natural data vs. data with one cue neutralized), we observe how the representation

shifts solely as a result of the missing cue. The movement in the representation thus reveals the mapping of the cue in the low-dimensional space.

3.4.2 Methods

Model

The model configuration takes the model from Study 1 and adds an additional low-dimensional layer before the final output used for tone prediction (Figure 3.3). It takes as input the last hidden outputs of the bidirectional LSTM and outputs to the softmax layer for classification.

The size of the low-dimensional layer was tested from 1 to 128 units to find the smallest number of dimensions (least complex model) that maintained model performance. Shown in the Results section later, we found that we were able to use a layer with just *two* dimensions without sacrificing much accuracy (2.6% decrease of model performance, shown in Table 3.5 in the next section).

The two dimensional representation is the only information the model has access to for making the tone classification. Each input token is converted from the original input of size $40 \times t$, where t refers to the duration, to the two-dimensional vector, which represent the token in the abstract space. In other words, the two dimensions are compressed information the model learns and constructs from the speech input, but do not refer to any specific acoustics or linguistic space. The model then applies a linear transformation to the two-dimensional space to predict the tone. We further examine the geometry of this space and each dimension of the representation in terms of its linguistic interpretability.

Data Ablation

We use the same manipulated data sets as in Study 1: datasets with pitch, intensity, and duration neutralized. In addition, for the purpose of testing if segments are represented in the space and how, we further performed ablations where we neutralized the rhyme of each syllable. This results in four ablated datasets with one type of information being neutralized in each.

It is important to clarify that in contrast to Study 1, the target model we test on in Study 2 is always the model trained on natural speech, and the test data consists of natural and neutralized tokens. It is necessary to consistently test on the same model trained on natural speech for two reasons. First, in order to probe the representation learnt by the low-dimensional layer, we need to hold the low-dimensional layer constant (by holding the training data constant), but ablate the test data. Secondly, the scale of the representation needs to be consistent in order to compare the representation before and after cue neutralization.

For rhyme neutralization, all rhymes in an utterance were resynthesized to share the same rhyme segment(s) as the first rhyme of the utterance. All manipulations were done in Praat. The pitch trajectories and the voice source remained unchanged for the neutralized rhymes. The resulting corpus consists of utterences of words sharing the rhyme of the first word in the utterance. The resynthesis relies on the source-filter theory of speech production (Fant, 1970), where changing the 'filter' – the shape of speakers' articulators – changes the articulated sound, while the speakers' voice (including F0) is unchanged. In our resynthesis, the first rhyme of each utterance served as the filter which is intended to represents the contribution of formants to the spectral envelope[3]. The rest of the rhymes in the same utterance only

[3]The filter is extracted by using `To Formant (burg)...` in Praat, and using the default values of the five parameters: Time steps = 0, Max. number of formants = 5.0, Maximum formant = 5500Hz, Window length = 0.025, Pre-emphasis frequency = 50Hz.

retain the source information, achieved by inverse-filtering the original signals with their own LPCs to remove the effect of formants [4] . The same formant filter (i.e. the first rhyme) is applied to all the subsequent rhymes, creating identical segmental information with the duration being the shorter one of the two (the filter or the source).

Feature extraction

All input features were extracted after the manipulations were done. We used same the 40-dimensional features (39 MFCCs + F0) as in Study 1. The feature extraction and preprocessing specifics are the same as Study 1 except for the z-scoring process. Recall that for Study 1, we examined performace on models trained with only partial information (manipulated data sets) to see how well they could predict tone category. The test set was thus z-scored based on the corresponding manipulated training set. In Study 2, however, the purpose is to examine the representation of the low-dimension layer trained on natural speech. In this case, the manipulated test sets were z-scored using the mean and standard deviation of natural training set.

3.4.3 Results

Low-dimensional representation

Our experiments suggest that two dimensions are sufficient to make tone classification while only lowering the accuracy by 2.6%, as shown in Table 3.5. Restricting the dimension to one would considerably lower the performance. However, when

[4]The LPCs are extracted for each subsequent rhyme using `To LPC (autocorrelation)`, with default values of the four parameters: Prediction order = 16, Window length = 0.025, Time steps = 0.005, Pre-emphasis frequency = 50Hz.

further increasing the dimensions, the accuracy does not improve significantly. This suggests that tones can be represented in a two dimensional space.

Dimension(s)	Accuracy (%)
1	62.1
2	74.7
3	76.5
4	76.6
5	76.4
128	75.8
Study 1 benchmark	77.3

Table 3.5: Accuracies for models with different number of dimensions in the low-dimensional layer.

Further investigation of the two-dimensional representation reveals that the tones are organized in a quadrilateral geomtery, where T1 is adjacent to T2 and T4 and is opposite to T3. Figure 3.4 shows each test token plotted along the two dimensions of the low-dimensional layer. The left side of the plot shows the results of all test tokens while the right plot only includes the correctly classified tokens, showing a clearer pattern that the model has carved up the two dimensional space into four quadrants. The black dots represent the mean of each tone on the two-dimensional space, with the ellipses representing the 95% confidence interval. The means of the natural data serve as a bench mark which is used to compare with the manipulated data. In the current study, as we are primarily interested in interpreting the learnt two-dimensional representation, including falsely clasified data points would obscure the pattern, so we examine only the correctly classified tokens, which we assume are more representative of the learnt representation.

Given this geometry, one may speculate that the model may be classifying the tones by their pitch height and contour, or their oneset and offset. Specifically, if D1 represents pitch contour and D2 represents average pitch height, then the rising (and level) tones and falling tones are separated on D1 while high and low tones

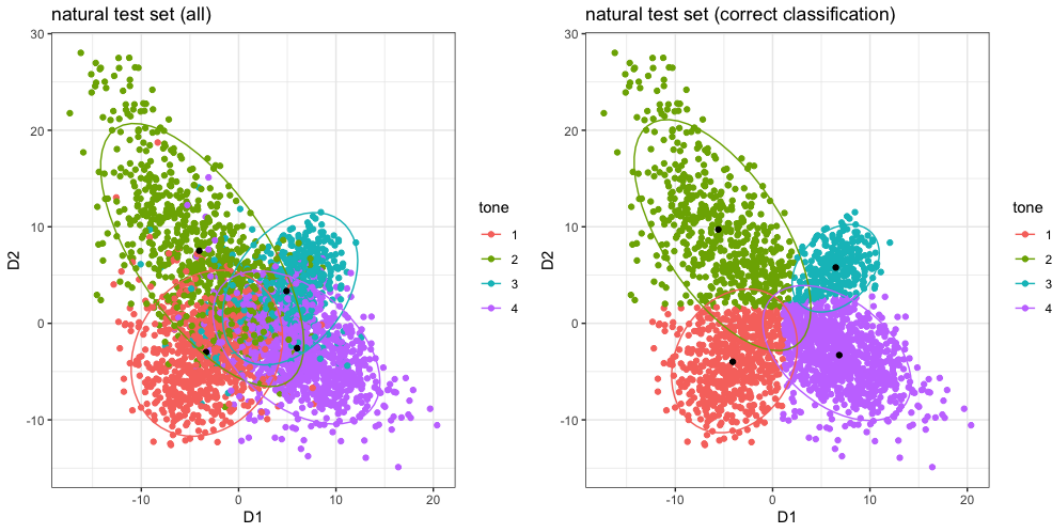


Figure 3.4: Two-dimensional representation of the natural test set (no ablation). Left: the entire test data. Right: the subset of test tokens with correct tone classification only. Black dots refer to the means of the tones and the ellipses refer to the 95% confidence interval. The current study only investigates the representation of the correctly classified tokens.

are seperated on D2. Alternatively, it is also possible that D1 corresponds to the offset and D2 corresponds to the onset, with smaller values indicating larger pitch values. We seek empirical support for these different interpretations by examining pitch ablation, conducted in Study 2 and 3.

Mapping between cues and the low-dimensional representation

In order to determine how the different cues in our study (pitch, intensity, duration and rhyme) are represented in the two-dimensional space learnt by the model, we analyse the direction in which the distribution of each tone changes when a specific cue is neutralized.

Figure 3.5 illustrates two hypothetical examples. Imagine that we neutralize some cue (e.g., pitch) in our test data. Whenever we neutralize a cue, we expect the mean values for each tone to shift closer to the global mean values for all tones, since

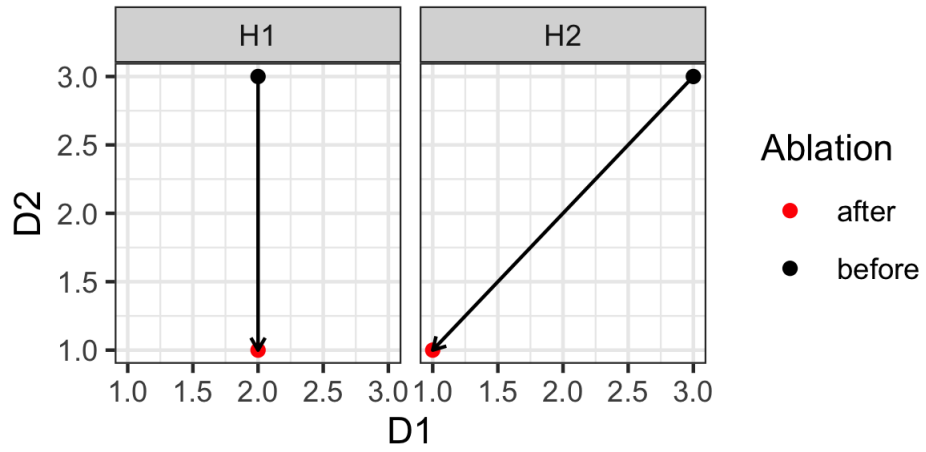


Figure 3.5: Illustration of two kinds of hypothetical movement in tone's representation in a two-dimensional space. The left plot shows one hypothesis where a cue is represented on one dimension and the right plot shows an alternative hypothsis where a cue is represented on both dimensions.

any neutralization should make all tones more similar. However, such shifts may differ in magnitude along the two dimensions of the representation. The left plot of Figure 3.5 illustrates the case where the neutralized cue is represented primarily along dimension D1 of the two-dimensional space. In such a case, neutralization will lead to a shift of the mean value of D1 for each tone towards the global mean value. The example illustrated in this plot assumes that the mean value on D1 of this tone was initially greater than the global mean, and thus the tone shifts downward. The right plot of Figure 3.5 shows another possibility. In this case, the neutralized cue is represented equally on dimensions D1 and D2 of the representation. In the example illustrated here, we assume that the tone's mean value is greater than the global mean on both dimensions. Since the two dimensions both represent the neutralized cue, we predict a shift towards the center of the plot. Any real data will be noisier than such hypothetical data of course. Nevertheless, by analyzing the

direction and magnitude of change under each tone mean, we will show how our two-dimensional space represent tonal information.

Segments While the focus of our study is the acoustic space of tonal representation, we first examine how much segmental information contributes to tone classification. As reported above, study 1 only demonstrated that a model *could* learn the correlation between rhymes and tone, but does not examine whether or not the model makes use of this information. Figure 3.6 illustrates that the distribution of tones indeed shifts when rhymes are neutralized, confirming that rhymes are encoded in the representation, and that the model indeed learns the segmental information.

Neutralizing rhymes affects all tones, as shown by the movement of the means for all tones. Moreover, the movement is realized on both dimensions, suggesting that segmental information is represented on both dimensions to facilitate tone identification.

Duration As shown in Figure 3.7 below, the representation of individual tones move towards the center in both dimensions to roughly the same degree. The magnitude of movement largely corresponds to the difference between the mean duration of the tone and the neutralized duration (12 frames). Figure 3.2 (c) shows that the four tones do not vary much in duration. In fact, t-tests suggest that only T4 is significantly different from the other tones. The magnitude of T4 movement is indeed the smallest among the four tones, while the shifts in the rest of the tones are similar in magnitude.

Intensity Figure 3.8 shows how tones are represented when intensity is not available in the acoustics. The tones respond differently: T2 and T3 move mostly along D1 while T1 and T4 move in both dimensions. In other words, different tones seem

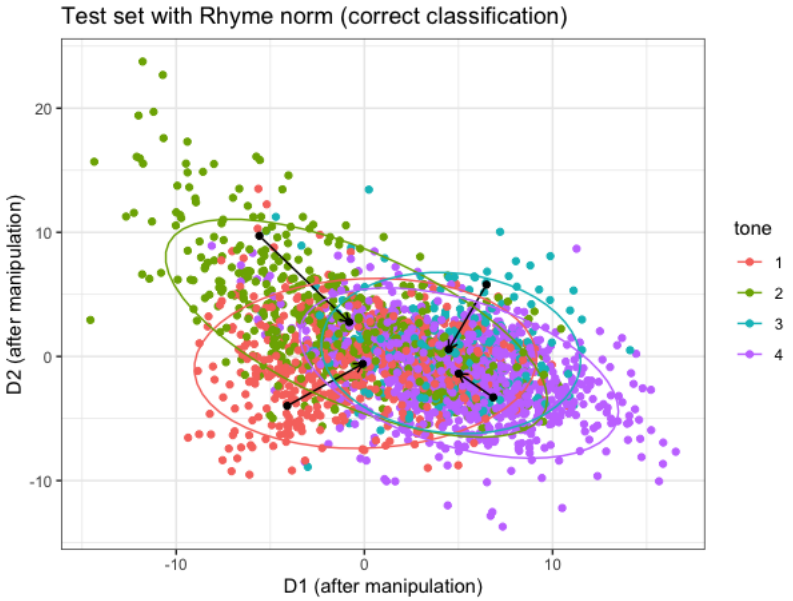


Figure 3.6: Two-dimensional representation of the rhyme-neutralized test set, z-scored relative to the natural training set, showing only tokens which can be correctly classified without manipulation. Arrows represent the effect of the loss of rhyme information relative to the natural data. The endpoints of the arrows are the means of the ablated data (rhyme-neutralized), and the start points of the arrows represent the means of the natural data. The ellipses refer to the 95% confidence interval after ablation.

to encode intensity on different dimensions. When comparing the movement of the representation with the characteristics of each of the tones in continuous speech, summarized in Table 3.2, there is also no clear relationship. It is possible that intensity is not explicitly encoded in a compact two-dimensional representation, and we return to discuss the possible explanations in the Discussion section.

Pitch The movement after pitch ablation is shown in Figure 3.9. As in the previous figures, the arrows point from the means of the natural data to the means of the ablated data. The four tones change differently in response to the loss of pitch information and these changes reflect the characteristics of the tones in continuous

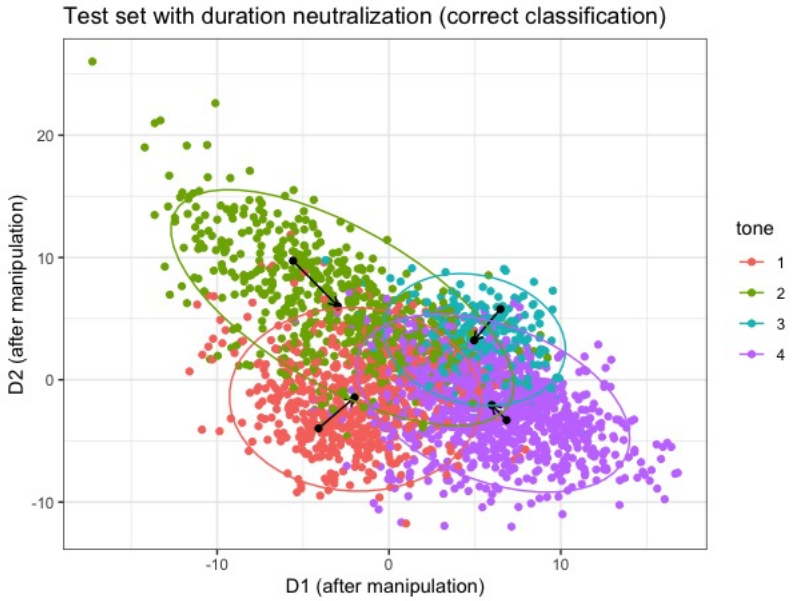


Figure 3.7: Two-dimensional representation of the duration-neutralized test set (12 frames), z-scored relative to the natural training set, showing only tokens which can be correctly classified without manipulation. Arrows represent the effect of the loss of duration distinction relative to the natural data. The endpoints of the arrows are the means of the ablated data (duration-neutralized), and the start points of the arrows represent the means of the natural data. The ellipses refer to the 95% confidence interval after ablation.

speech. We first summarize the pattern observed in Figure 3.9 and proceed to test specific hypotheses of pitch representation in the next section.

A close examination of the movement of different tones suggests that pitch may be decomposed on the two dimensions as average pitch height (D2) and relative pitch contour (D1). Specifically, average pitch height refers to the pitch averaged over the entire tone and relative pitch contour refers to the shape of the pitch contour with respect to the average pitch. These two components together capture the characteristics of pitch for any tone and best explains the movement of all tones in Figure 3.9. While we go in detail to verify this hypothesis (and other alternative

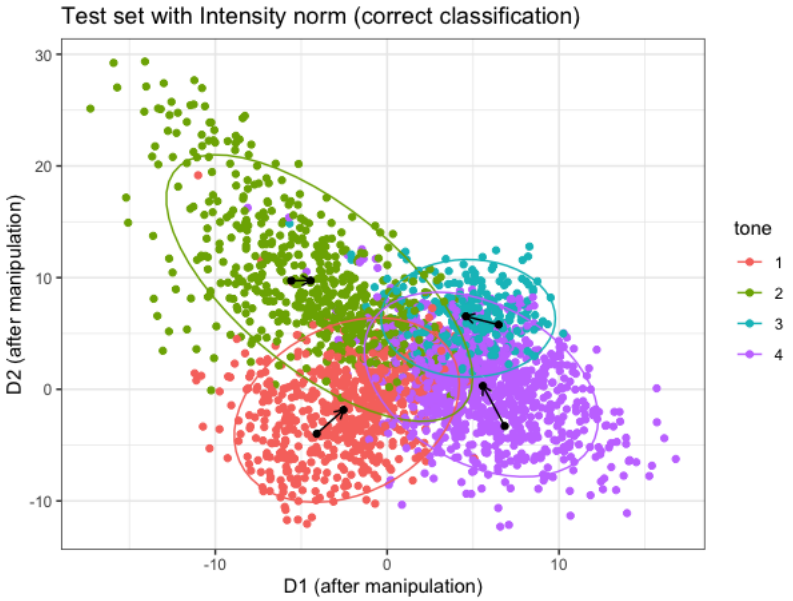


Figure 3.8: Two-dimensional representation of the intensity-neutralized test set, z-scored relative to the natural training set, showing only tokens which can be correctly classified without manipulation. The arrows represent the effect of the loss of the intensity distinction relative to the natural data. The endpoints of the arrows are the means of the ablated data (intensity-neutralized), and the start points of the arrows represent the means of the natural data. The ellipses refer to the 95% confidence interval after ablation.

ones) in the next section, here we delineate how the two components can be inferred from Figure 3.9.

First, T1 exhibits the smallest change among the four tones and only moves along D2 (average pitch height). The small magnitude can be explained by T1's average pitch height being the most similar to the neutralized pitch (see Table 3.2 in Study 1), among the four tones. In terms of the direction of movement, as a level tone, T1 only moves along D2. This matches how T1 differs from the neutralized pitch: both being flat in the pitch trajectory, T1 is only slightly higher than the neutralized pitch. Second, T3 and T4 which are both falling tones in continuous speech (see Figure 3.1a), both move leftwards along D1 (pitch contour). In contrast to the leftward

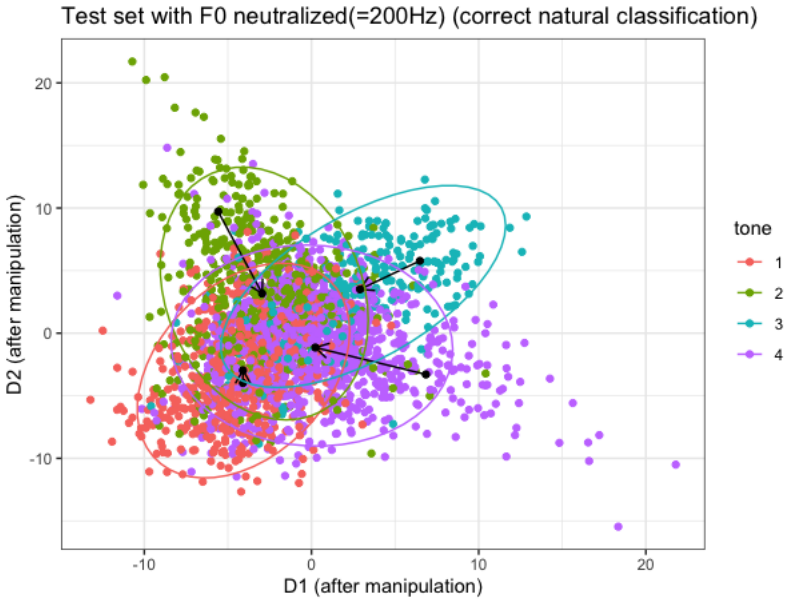


Figure 3.9: 2D representation of the pitch-neutralized test set, z-scored relative to the natural training set, showing only tokens which can be correctly classified without manipulation. The arrows represent the effect of the loss of pitch distinction relative to the natural data. The endpoints of the arrows are the means of the manipulated data (pitch-neutralized), and the start points of the arrows represent the means of the natural data. The ellipses refer to the 95% confidence interval after ablation.

movement of falling tones on D1, the rising tone T2 moves in the opposite direction. This is indeed the pattern we expect to observe if D1 were to represent relative pitch contour (whether it is rising or falling). T2's large movement along D2, however, is unexpected, given that T2 is not the tone that differs the most in average pitch height from the neutralized pitch. Thus, we would not expect it to display the largest movement along D2 (the reason may be again the outliers, discussed previously in duration neutralization). Nevertheless, the overall pattern suggests that pitch is represented on the two dimensions as average pitch height and relative pitch contour. To further validate this claim, and reject other alternative hypotheses of the two-dimensional representation of pitch, we did a follow-up study focusing on testing pitch-specific hypotheses, described in the next section. As both studies

concern the low-dimensional representation of tones, we discuss the results of pitch in the two studies together at the end of Study 3.

3.4.4 Discussion

The geometry of the tone space

Our results conform to previous findings in multiple ways. First, as in previous studies, we also found that two dimensions are sufficient to represent the four tones, only losing 2.6% accuracy. This is consistent with previous literature, where Mandarin tone representations are two-dimensional, based only on pitch cues (e.g. Rhee and Kuang, 2020) or a combination of cues (e.g. Tupper et al., 2020).

While two dimensions can reasonably represent tones, there is indeed a marginal 1.8% boost to the accuracy if the dimension is increased to three. The loss of accuracy suggests that some (properties of) cues, which can otherwise be better represented on three dimensions, are forced to be compressed on two dimensions. We speculate that the loss of accuracy is likely to result from the compression of cues and hence render the relevant cues less interpretable (particularly intensity), which we discuss in detail below.

Second, we found that the tones form a quadrilateral, with T1 being adjacent to T2 and T4 while being opposite to T3. Such geometry is maintained no matter what linear transformation is performed on the space. Most importantly, the same geometry is also found in other studies on both production and perception (Chandrasekaran et al., 2007; Rhee and Kuang, 2020; Tupper et al., 2020), using multidimensional scaling for dimensionality reduction. Rhee and Kuang (2020) examined tone F0 trajectories at different phrase positions and found similar quadrilateral organization of the tones. Chandrasekaran et al. (2010) also presented similar findings to the current study. They showed that the perceptual space of the tonal contrast

which represents the reaction time of listeners for tone identification is also two dimensional, with the dimensions corresponding to average pitch height and pitch contour. Furthermore, they also found a similar geometry to ours, with the tones encoded at similar positions in the space.

It is worth noting that although these studies differ in many aspects – using different input (full acoustics or F0-only), examining different aspects of the tones (production and perception), and using different dimensionality reduction methods, they nevertheless find the same structure of the tone space. Compared to our findings, the geometry found in different studies share similar encoding of the tones in the space. This suggests that the tone space is constructed similarly in both production and perception, and that all cues are organized in a similar geometry.

Interpretation of the representation

Segments Rhyme neutralization further supports previous findings that the model implicitly learnt the association between segments and tone. As shown in Figure 3.6, all tones move towards the centre when the rhymes are neutralized to be the same within each utterance. It also matches previous findings that rhymes are correlated with tones, which is first inferred from the spectral information (e.g. vowel formants) and used to facilitate tone identification.

Duration We observe that duration does not seem to be encoded in a single dimension that is easily interpretable (Figure 3.7). It could be that duration is encoded indirectly through the encoding of other features. Algorithmically, duration is the sequence length of the input, which is essentially accounted for by the *counting behaviour* in the model. Weiss et al. (2018) found that LSTMs indeed have the ability to track the length of the input, but it is not entirely clear whether the model directly encodes duration as a cue, or is a byproduct of encoding the sequence information.

Moreover, we did not find evidence in support of duration being encoded explicitly, that is, the duration ablation does not affect the representation of tones in an interpretable way. Therefore, it is more likely that duration is encoded indirectly.

Intensity While the mapping between intensity and the two dimensions is even less clear and requires follow up study, the results neverthess provide negative evidence that average intensity is not represented on any of the dimensions. Recall that the intensity ablation has lowered the intensity to 70db, which is below any of the tones' averages. Therefore, if average intensity were to be represented on one dimension, it must be the dimension where all tones move, i.e. D1. However, the magnitude of the movement does not correspond to the intensity difference: T4 has a much higher average intensity thant T3, yet it moves less than T3 along D1.

The inconsistency further points to a possibility that intensity may be represented better on a third dimension if allowed, and the compression results in the cue being uninterpretable. This mainly arose from the speculation that the intensity contour and the pitch contour mismatch (Figure 3.1), while pitch is much more important than intensity. We discuss this point later in the general discussion together with the pitch cue.

Summary In summary, we find that the low-dimensional tone representation is structured in a manner such that pitch height and pitch contour are mapped to the two dimensions respectively (further investigation in the next section), with duration being encoded implicitly. Intensity shows more tone-specific patterns and is likely to be represented as average intensity on one of the dimensions, although further study is needed. Segmental information, on the other hand, is represented on both dimensions. We hold our discussion on the pitch representation for Study 3, where we investigate specific hypotheses on pitch representation.

3.5 Study 3: pitch-specific hypotheses

3.5.1 Introduction

In Mandarin tones, pitch is a dynamic cue that has a set of properties (e.g. onset and offset, average height, direction). Several proposals have been made about how this continuous space can be used to identify the four tones. As a way to probe the pitch representation in the two-dimensional space, we evaluate specific claims in the literature. We test two classes of hypotheses which disintegrate pitch into two orthogonal dimensions differently: the first decomposes pitch into onset and offset and the second decomposes pitch into average pitch height and relative pitch contour, the tendency shown in the results in the previous section. Note that each class of hypothesis has a number of variants. While we mention some variants, we do not intend to test certain hypothesis specifically. Instead, we focus on the high-level idea these variants share.

Onset-offset Hypotheses

The *Onset-offset hypothesis* mainly treats the pitch representation as onset and offset. As shown in Table 3.1 earlier, tones are conventionally notated by the onset and offset. While the specific representation may vary as discussed in the introduction, they agree in that tones need to specificy pitch targets for onset and offset. *Prediction:* Onset is associated with one dimension and offset with the other.

One variant of this class of hypothesis differs in terms of the importance of the offset, which proposes that offsets of contour tones are underspecified (Yip, 2001). Yip argues that the level tone (T1) needs to be specified by both onset and offset as the production targets, while contour tones (i.e. T2, T3, T4) only need to be specified by the onset, with the rest of the tone "drifting away" from the onset. Yip's proposal

divides the pitch range of the tone inventory into four levels: it first divides the range into two halves by the register features [+/-Upper], with each register being further divided into two using the feature [h/l]. For contour tones, after the onset is specified, the rest of the tone drifts towards the middle of the register. In other words, the contour is a by-product of the onset specification. Level tones, on the other hand, have both onset and offset specified and the pitch trajectory stays level. *Prediction:* Onset is associated with one dimension, while offset may vary for different tones.

To summarise, if the pitch is indeed represented in terms of onset (and offset), we should at least find that onset is represented on one and only one dimension (i.e. Figure 3.5 left).

Height-Contour Hypothesis

The second class of hypothesis essentially divides pitch into average pitch height and the shape of the contour. While there are again a number of variants in terms of how one can account for the shape (slope, curvature etc.), we chose to use the entire contour so that the representation is not restricted to any specific hypothesis.

As described above, several studies have argued for the height-contour hypothesis from different domains including but not limited to speech perception (e.g. Hallé et al., 2004; Tupper et al., 2020; Wiener, 2017) and neuroscience (e.g. Chandrasekaran et al., 2007). In order to account for the contour, various studies adopt slightly different measures. Some researchers use simplified measures, such as the difference between the onset and offset (Yang, 2015) and the difference between the maximum and minimum pitch divided by duration (e.g. Flemming and Cho, 2017). Other researchers fit a function to the contour, such as parabolics (Tupper et al., 2020). While the specific measures differ among those studies, they all find that contour is informative to Mandarin tone identification. Moreover, listeners' use of average

pitch height and the shape of the contour are orthogonal. There is also evidence that which dimension listeners use depends heavily on the language background. Native Mandarin listeners typically rely more on the contour while non-tone language listeners rely more on average pitch height, and switch to use pitch direction more as they learn Chandrasekaran et al. (2010). *Prediction:* Average pitch height is associated with one dimension and the contour shape is associated with the other.

3.5.2 Methods

Models

The model configuration is the same as Study 2, with the only difference being a smaller input size. Instead of the 40-dimensional input, we use only F0 (no MFCCs). The main motivation is to eliminate confounded changes which can potentially be introduced when ablating pitch. For example, MFCCs carry some acoustic information such as voice quality that are correlated with pitch. Such information is likely also changed when the pitch is resynthesized. As we only wish to examine the effect of one cue, it is problematic to have more than one cue being changed. Among many solutions, we chose to only include F0, as the learnt representation is similar enough to the representation learnt on the 40-dimensional input (shown later in Figure 3.10 in Results). Thus we assume that the findings in Study 3 can be generalized to the representation learnt in Study 2, even though the two representations are not the same.

Data Ablation

Following the rationale of Study 2, we trained the model on natural data and only ablated on the test data. For the resynthesis of pitch, we again used the PSOLA method in Praat as we did in Study 2.

Onset-offset Hypothesis We use two ways to evaluate this hypothesis, one on the correlation between onsets and offsets and the two dimensions, and the other on examining the change of the two-dimensional representation when the first half of the rhyme (indicating onset) and the second half (indicating offset) are neutralized to 200Hz. The two evaluations reflect two possibilities of how one can account for onsets and offsets. While onsets and offsets refer to the very beginning and the end of the tone, using one frame at the very beginning and the end may reflect more about the transition between adjacent tones, rather than the tone itself. It is desirable to have a more stable representation to actually reflect the target tone which we did by neutralizing each half of the tone.

Height-Contour Hypothesis In order to test the correlation between the two cues (i.e. average pitch height and the shape of the contour) and the dimensions, we did an ablation for each cue to create two datasets. For the first ablated dataset (pitch height neutralization), we subtracted the mean of the tone across all utterances in the training set and added the maximum pitch of the utterance the tone token occurred in. This ensures that tones are disassociated from their height differences (through mean-subtraction) while keeping all pitch values above zero (through adding the max pitch from the utterance the tone occurred in). For the second ablated dataset (contour neutralization), we extracted the mean pitch of each tone and flattened the entire tone to have the same pitch as its mean. In other words, all resulting tones are level tones at the frequency of their original mean pitch.

3.5.3 Results

Figure 3.10 shows the two representations of the tones trained on natural data with different inputs. The left panel shows the representation using pitch-only training data and the right panel using 40-dimensional training data. While the two dimen-

sions learnt in the two plots are not the same due to the different training input, they display similar patterns in terms of the distribution of the four tones, and thus we assume that the findings using pitch-only data is generalizble to using the full signal.

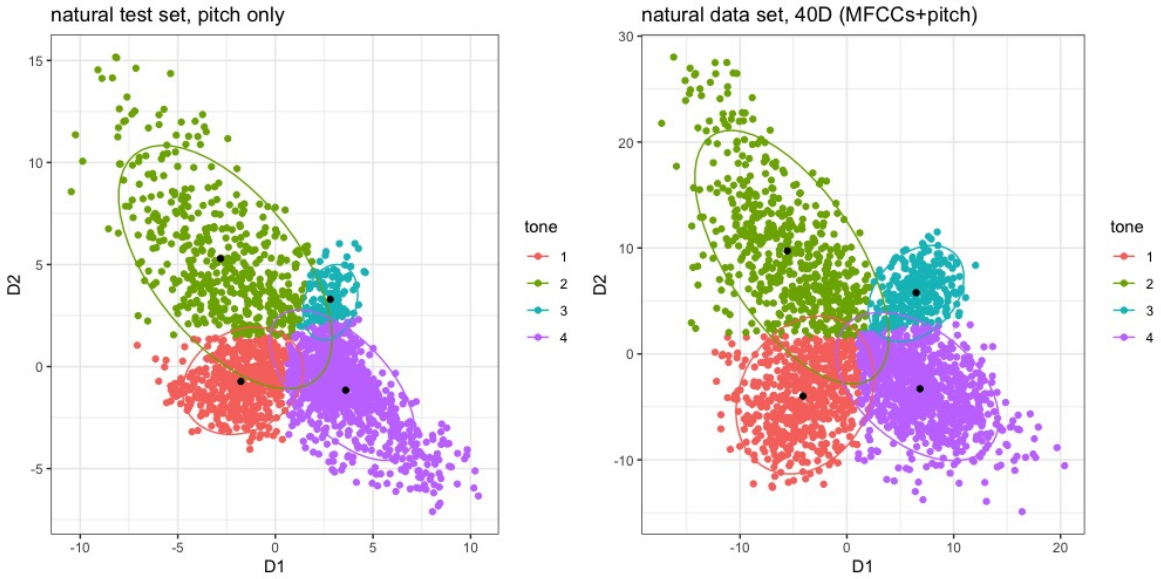


Figure 3.10: Mean and 95% Confidence Interval of each tone in the natural test set, plotted in the low-dimension space learned by the model. Left: the model trained with just pitch as input (Study 3). Right: the model trained using 40-dimensional input (Study 2). Only correctly classified data are plotted.

Onset-offset Hypothesis The first evaluation of the Onset-offset hypothesis measures the correlation between the onset and the offset and the two dimensions repectively. If this hypothesis is correct, we expect to find onset F0 correlating with one and only one dimension and offset F0 correlating only with the other dimension. However, as shown in Figure 3.11, both onset and offset of T1, T2, T4 correlate with both D1 and D2, and both the direction and magnitude of these correlations are similar across the two dimensions within each tone. However, the onset and offset of T3

do not strongly correlate with either dimension. These results suggest that onsets and offsets are not orthogonal in this two-dimensional representation.

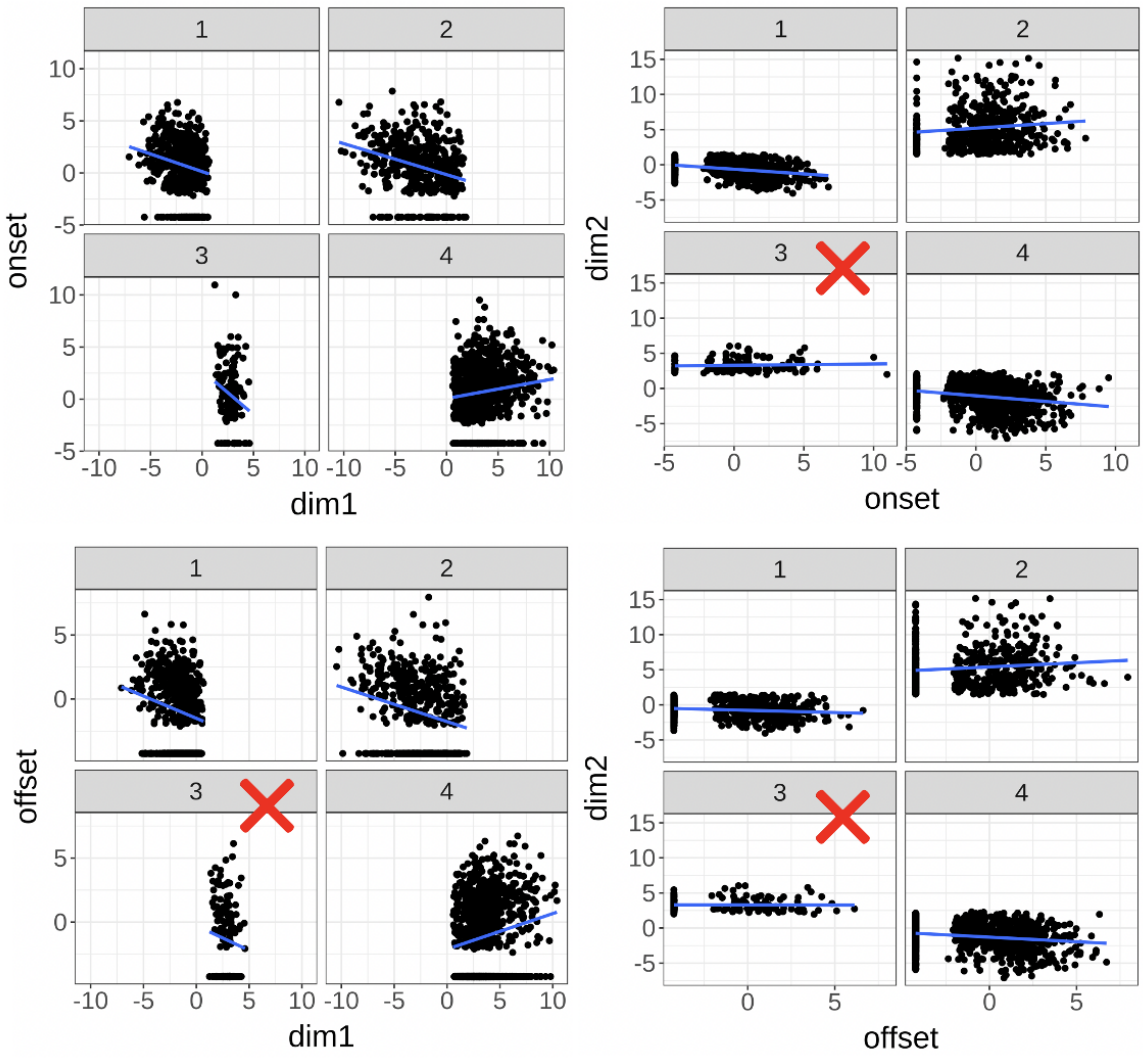


Figure 3.11: Correlations between onset and offset and the two dimensions (note different scales).

Figure 3.12 below shows the effect of onset neutralization (left) and the offset neutralization (right) respectively. The arrows represent the change of means of each tone before and after the neutralization. The plots reveal similar findings as the previous correlation test: for T1, T2, and T4, the changes are diagonal for both

the onset and the offset, indicating that they are encoded on both dimensions. Only T3 shows a clearer distinction: onset is correlated mostly with D2 (vertical move) and offset is correlated mostly with D1 (horizontal move). These results again suggest that onsets and offsets are not represented independently on single dimensions, but on both dimensions, at least for T1, T2 and T4. Thus, they do not match the predictions by the Onset-offset hypothesis.

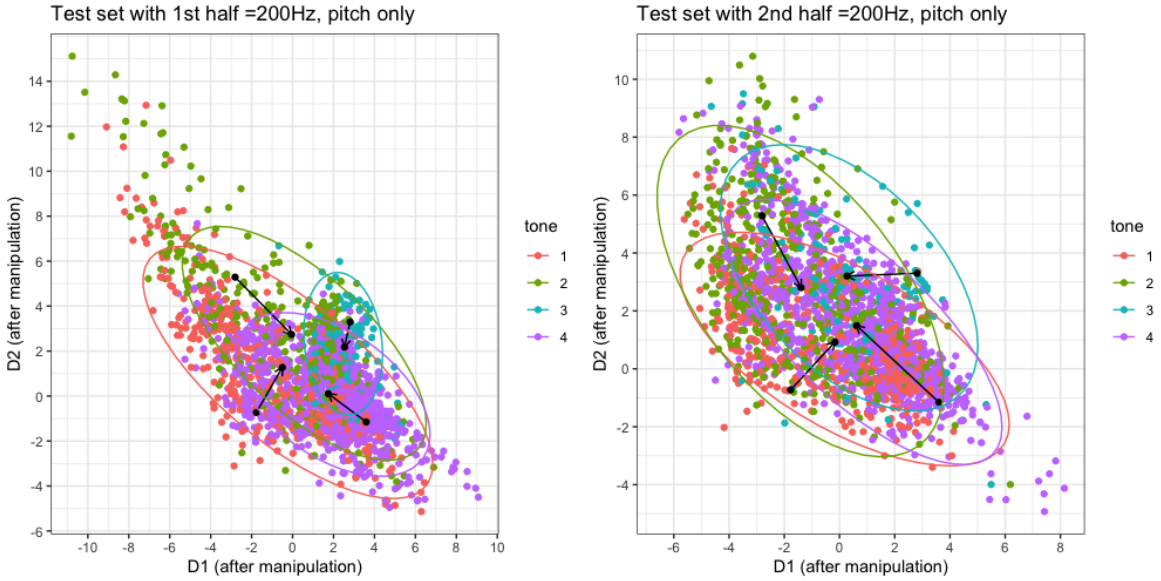


Figure 3.12: Mean and 95% Confidence Interval of each tone before and after onset neutralization (left) and offset neutralization (right). The start of the arrow is the mean of the natural data and the end of the arrow is the mean of the neutralized data.

Height-contour Hypothesis The height-contour hypothesis states that pitch can be represented as pitch height and pitch contour on two dimensions. Figure 3.13 below shows the representation for height neutralized (left) and contour neutralization (right). Again, the starting point of the arrows are the means of the tones in natural production, and the end points of the arrows are the means of the tones after neutralization.

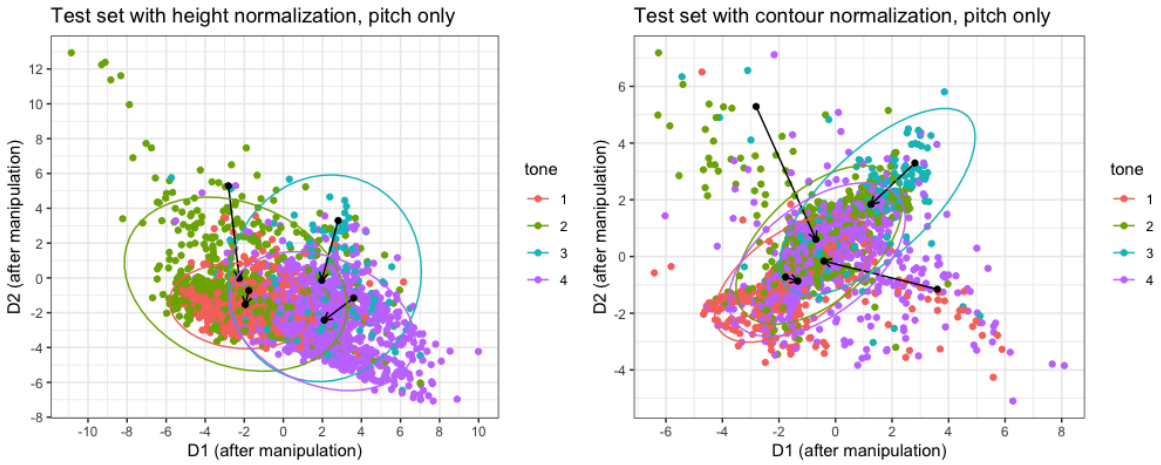


Figure 3.13: Mean and 95% Confidence Interval of each tone after height neutralization (left) and contour neutralization (right). The start of the arrow is the mean of the natural data, the end of the arrow is the mean of the neutralized data.

The pattern is clearer than for onset-offset neutralization. As can be seen on the left-hand side of Figure 3.13, pitch height neutralization leads to a vertical movement of the points, indicating that D2 largely codes this aspect of pitch information. Moreover, the magnitude of the change roughly corresponds to the average pitch height of the four tones, shown in Table 3.2: T1 and T4 are relatively higher, corresponding to smaller magnitude, while T2 and T3 are relatively lower, corresponding to larger magnitude. Recall that the height normalization is done by adding the maximum F0 of the *utterance*, after subtracting the mean F0 of the rhyme. As T2 and T3 are much lower, the manipulation thus does more change to T2 and T3 then to T1 and T4, which is indeed what is shown in the left plot.

The right plot, while not showing the pattern as clearly as the left one, nevertheless shows distinct pattern of movement between rising and falling contours on D1. The phonetically realized falling tones (i.e. T3 and T4) move leftwards while the rising tone T2 moves in the opposite direction. Moreover, T1 is realized with a slightly rising contour (shown in Figure 3.1 (a)) and indeed patterns with T2 in

terms of the rightward direction of movement. The mapping between the contour shape and the movement along D1 verifies that D1 is correlated with relative pitch contour. There is, however, a relatively large movement along D2 for T2 and T3. We discuss the potential explanation for the unexpected movement in the next section.

3.5.4 Discussion

In Study 2, we studied the low-dimensional representation of tones and how it changes when various acoustic and segmental cues are removed. While the model manages to learn a two-dimensional representation which can be solely relied on for tonal classification, the representation is highly compressed and somewhat abstractly structured, with many cues not showing a clear one-to-one mapping with the dimensions. However, in study 3 we find evidence for a structured projection of pitch in the space — pitch height is represented roughly on one dimension and pitch contour on the other. This fits with recent findings on the most discriminable cues for tone perception in isolated tones. Tupper et al. (2020) found that among various measures describing pitch trajectory, the average height and contour shape are most important, and other measures of F0 are redundant or correlated with the more critical cues.

One exception worth noting is T2's large movement along D2 when pitch contour is neutralized, where the major movement is expected only on D1. One possible explanation is that the phonetically realized T2s actually consist of two groups: 1) underlying T2s and 2) underlying T3s that undergoes T3 sandhi. Specifically, T3 sandhi refers to a T3 turning into a T2 when it precedes another T3 (Mei, 1977; Wang and Li, 1967). While those sandhi T3s (surface T2s) display similar rising contour as underlying T2, researchers show that they differ with underlying T2s in multiple ways (e.g. Chien et al., 2016). In terms of the model's representation, it is possible

that the model encodes the two groups of surface T2s differently, which leads a more dispersed representation, and hence cause a larger movement in D2.

Onset and offset, on the other hand, are found to be represented on both dimensions. Similar findings are also reportded in Chen and Xu (2020), which shows that the intermediate features, i.e. pitch targets as specified in conventional tone inventory notation like Table 3.1, are not necessary for tone recognition.

3.6 General Discussion

This study examines the relative contribution of cues to the multi-dimensional Mandarin tonal contrast in continuous speech. Our results are partially consistent with previous findings: similar to previous studies, we found pitch to be the most important cue. However, duration, which previous work found to be less important than intensity, was more important according to our findings. We further found that the model may implicitly make use of the rhyme-tone correlation to facilitate tone identification.

Our second finding was that tones can be represented in a two-dimensional space, also consistent with previous research. Moreover, some cues show a clear mapping with the two-dimensional space, while others are less decomposable. Pitch can be represented on the two dimensions as pitch height and relative pitch contour, while it is less clear how intensity, duration and rhymes are represented.

3.6.1 Relative Contribution of cues

Pitch Consistent with previous studies, Study 1 also found pitch to be the most important cue in continuous speech, causing the largest decrease in tone identification accuracy when it is missing. Moreover, the effect of pitch on tone identification varies by tone, and is notably less important for T3. Instead, intensity is more used

for distinguishing T3. One plausible explanation is that T3 shares a similar falling contour as T4, while its intensity contour is more distinguishable from the rest of other tones, and thus is made more use of.

Duration The duration patterns in continuous speech in our study show that the four tones largely overlap with each other, with only T4 being significantly shorter than other tones. This is consistent with previous production studies on various types of continuous speech, where duration of tones converge (Yang et al., 2017). Yet, it is surprising that it turns out to be the second most important cue, while in previous studies on isolated tones, it is found to be the least important (e.g. Fu et al., 1998). Given the reduced distinctiveness of duration in continuous speech, we suspect that the change of cue ordering may result from intensity being less distinguishable, which we discuss below.

Intensity The most striking findings of intensity is that its role in the tonal contrast is extremely small in continuous speech, as opposed to previous studies on isolated tones where it is more important than the duration cue. This might be caused by the mismatch between the intensity contours and the pitch contours. Previous studies reveal that pitch and intensity are positively correlated to different extents for different tones, talkers and rhymes. Furthermore, when the correlations between the two cues are higher, tone identification is also better (Fu and Zeng, 2000; Luo and Fu, 2004). Our results, however, show that intensity and f0 contours are not very similar (with T3 being an exception). While we do not know if models treat the (dis)similarity between pitch and intensity as meaningful for tone identification, given the shared manner of decomposing pitch between the models and humans, it is reasonable to assume that such inconsistency between the two cues causes the models' lowered reliance on intensity.

Moreover, the T3 results further support the argument above. While the intensity contours of the rest of the tones do not align well with their corresponding pitch contours, T3 exhibits a falling contour for both pitch and intensity. Correspondingly, the effect of intensity on the accuracy of T3 is considerably larger than other tones. If the enhanced importance of intensity is merely to disambiguate T3 from T4 which has a similar pitch contour, then we would also expect T4 to rely more on intensity. However, it is not the case according to our results, with T4 showing distinct pitch and intensity contours. This lends support to the idea that correlated f0 and intensity patterns enhance identification in both humans and models.

Segments While we did not directly neutralize the rhyme segments in study 1 to test their contribution to classification quantatatively, the higher-than-chance accuracy for the model trained just on rhymes nevertheless indicates that segmental information *may* and *can* be used to facilitate tone identification.

3.6.2 Low-dimensional representation

Together with previous literature, our results suggest an ideal listener would represent tones in a two-dimensional space, which encodes pitch as well as intensity, duration, and rhyme information. Such representation allows for a quasi-optimal solution for Mandarin tone identification in continuous speech.

Among all cues examined, we found pitch to be the only cue that can be decomposed on two dimensions in a clear manner. In fact, given the importance of the pitch cue, it is reasonable to assume that the two-dimensional representation is structured primarily based on pitch. Through testing specific pitch representation hypotheses, we found that pitch is decomposed into average pitch height and pitch contour, rather than onset and offset, which is consistent with previous perception literature (Gandour, 1984; Howie and Howie, 1976; Massaro et al., 1985).

This shows that a computational learner trained on a specific task using noisy and rich input signals in fact learns similar properties of cues as humans. Conversely, the similarity shared by the model and humans suggests that Mandarin listeners perceive tones using a strategy that is found to be optimal, at least as far as pitch is concerned. Our findings complement previous studies on low-dimentional representations (Bai et al., 2018; Weber et al., 2016), suggesting that not only segments, but also suprasegmental information can be learnt by a computational learner in a way that is similar to humans. This finding further implies that a similar method can be used to study tone representations in various other tone languages, and may provide insights for langauges with relatively more complex tone inventories. More broadly, this method can be applied to other studies of human perceptual representations to evaluate among competing hypotheses.

In terms of non-pitch cues, the results are less straightforward, and merit further investigation. As disccused in earlier sections, duration is likely to be implicitly encoded through encoding pitch. As for intensity, the difficulty in interpreting the mapping to the low-dimensional represenation may be because intensity was partly encoded on additional dimensions and some information was lost when intensity was compressed onto just two dimensions. Specifically, the intensity contour is drastically different from the pitch contour. Given that pitch is the primary cue and that it is clearly decomposable on the two dimensions, it is possible that intensity is encoded as a "redundant" or correlated cue to pitch, with only the correlated aspects being encoded. In fact, Tupper et al. (2020) also found that intensity is encoded as a correlated cue with pitch, although we expecte the correlation between the two should be lower for our stimuli than in the isolated tones they used. Further analysis is needed to examine the mapping of intensity to the low-dimensional representation.

Another striking finding of the two-dimensional representation concerns the geometry of the tones – the abstract organization of tones in this space – is surprisingly consistent with previous literature. While the input and the methodology vary from study to study, the organization of tones is consistent: T1 is adjacent to T2 and T4, and is opposite to T3. In study 2, this geometry is maintained no matter no matter which cue is neutralized, suggesting that it is not a specific cue – e.g. the most important pitch cue – that induces the quadrilateral geometry, but all cues.

3.6.3 Future directions

Building on the current study, there remains a lot to explore further in the future. First, as tones in continuous speech vary largely with the context, it would be interesting to examine the representation of tones when neighbouring context is included. In this paper, we deliberately discard context information so that it does not interfere with the representation of individual tones. Building on this first step, future research can examine how context modifies tone representation.

Second, in order to better evaluate how the model's representation aligns with human perception, especially the less understood cues such as intensity and duration, one may conduct a follow-up perception study using the stimuli from the same corpus used in this paper. In fact, how human listeners perceive tones in continuous speech is generally understudied. Given that the duration is less distinguishable in continuous speech, which is in conflict thhe previous work on isolated tones, it is interesting to study how humans adjust their use of cues compared to perceiving isolated tones, and if the use varies by the tone's position in the sentence.

Third, as mentioned earlier, the computational analysis of tone space can be applied to other tone languages or even other multidimensional contrasts. Moreover, according to the findings so far from our study and the previous research, the learnt

representation by the computational learner is similar to humans, which in turn provides insight on how humans integrate multiple cues in their perceptual representation.

3.7 Conclusion

This chapter investigates the contribution of various cues to the identification of Mandarine tones in continuous speech, and their geometric representation in abstract space. Unlike previous studies that use pre-extracted linguistic measures, this study adopts a data-driven method using raw speech. It explores the representation of tone by examining a low-dimensional layer learnt in a deep neural network tone classification model. The model can be seen as doing the same task as humans, with the low-dimensional layer being its tonal representation. Unlike the human brain which can only be indirectly probed through responses, the computational model provides a learnt representation one can directly examine. The main findings of the three studies reveal that 1) pitch is the most important cue, followed by duration, with intensity being the least important; 2) there exists a two-dimensional tone representation which compresses all cues, and tones in this representation form a quadrilateral geometry; 3) pitch is decomposed into average pitch height and relative pitch contour in the two-dimensional space.

Preface to Chapter 4

Chapter 3 examines the more general case of variability in continuous speech relative to isolated syllables. The tonal realization is indeed different from isolated production in all three acoustic cues examined. However, the learnt tonal representation by the model from continuous speech is similar to previous perception and production studies on both isolated and continuous speech. A quadrilateral geometry of tones in a two-dimensional space is consistently found regardless of the specific cues examined.

Chapter 4 studies a more restricted type of variability – phonological assimilation. Human listeners are found to display language-specific compensation patterns, where English listeners compensate more for place assimilation than voicing assimilation, while French listeners show the opposite pattern. The chapter uses a set of parameterzied ASR models to examine the type and complexity of linguistic knowledge needed for compensation for phonological assimilation.

Chapter 4

Modelling Perceptual Effects of Phonology with ASR Systems

4.1 Introduction

This study aims to understand phonological alternations, one type of variability in spontaenuous speech, in speech perception through computational modelling. It also investigates how much linguistic knowledge Automatic Speech Recognition (ASR) systems can capture and how their learnt knowledge compares to humans. While these two broad goals have different focuses – human perception and ASR models – they are intrinsically related. Human behavioral results provide the benchmark for evaluating ASR models, and evaluating our explicit implementation of an 'ideal listener' using ASR models improves our understanding of human speech perception. The comparison between models and humans also sheds light on how we may improve ASR systems to better acount for pronunciation variability.

Studies have found that non-canonical variants constitute 27% to 75% of instances for some sounds in conversational speech (e.g. Dilley and Pitt, 2007). There are a variety of sources that lead to non-canonical pronunciation, such as dialectal

differences and talker idiosyncrasies. Towards the broader goal of understanding how humans cope with variability in speech, this study focuses on one type of variability, from phonological alternations, defined as predictable sound changes when the context meets certain conditions. In particular, we focus on phonological assimilation, one widespread type of language-specific phonological process. For example, in English *green beans*, the *n* in *green* tends to be pronounced *m*, with the place of articulation assimilated to that of the following *b* (labial). While English-speaking listeners perceptually compensate for this assimilation, perceiving the *m* as *n*, listeners whose native language is French, which does not exhibit place assimilation, do not show this behaviour.

Our general interest in studying variability in speech using computational modelling arises from the discrepancy between humans and models when presented with the same challenge. While state-of-the-art ASR systems reach near-perfect performance when given clear read speech, they have a harder time when dealing with more noisy and variable speech (Davis and Scharenborg, 2016; Spille et al., 2018). Humans, on the other hand, have no trouble processing speech with extensive variability. This makes an interesting case for cognitive modelling to explore what knowledge or capacity makes a learner good at recognizing noisy speech signals.

In the field of speech perception, while many behavioral studies have investigated how humans process spoken language at different levels—from specific acoustic cues to understanding entire sentences—important questions remain about how these different levels of processing are integrated and interact with each other. For example, it is hard to isolate one's phonological knowledge, as it is already acquired and can not be 'undone'. Computational models allow for full control of the system, such that one can manipulate specific components to see how each change affects

the final outcome, and hence quantitatively investigate the importance of the corresponding component in human cognition.

On the modelling side, the results also inform us whether or not machine learning models constructed for very specific tasks (here, ASR models built for phone recognition) can nonetheless learn generalized knowledge about the sound system of a language, which resembles what humans use in speech perception. Human performance provides a cognitive benchmark to which we can compare ASR models, to better understand which aspects models manage to learn about and which aspects they do not.

In this study, we build ASR models of 'ideal listeners' capturing different kinds of linguistic knowledge, which are evaluated on the same stimuli from a behavioral experiment using human listeners (Darcy et al., 2009), and compared to the human results. Of the two broad goals introduced above, we put a stronger emphasis on what the computational modeling results suggest about the role of different kinds of linguistic knowledge in human speech perception. The next section gives an overview of the assimilation phenomena, the behavioral experiment results which serve as the 'human benchmark', and the computational experiments carried out in the current study.

4.2 Background

Humans are able to detect predictable sound changes in running speech and restore the original sound. This is one instance of *perceptual compensation*, the general ability of humans to 'undo' predictable changes in pronunciation due to context (Beddor and Krakow, 1999; Mitterer and Blomert, 2003; Ohala et al., 1990). The mechanism has been discussed in several frameworks: a top-down approach focusing on lexical knowledge, a bottom-up approach focusing on acoustics, and an intermediate

approach. However, important questions still remain about what kind of language-specific knowledge is necessary for listeners to perceptually compensate. In particular, while the intermediate approach makes use of both lower-level phonetics (the *phonetic knowledge*) and higher-level phonology such as phonotactics and syllable structure (the *phonological knowledge*), the relative contribution of the two types of knowledge is not clear. Is one of them more important than the other, or are both truly necessary to show human-like compensation behavior? In this paper, we will implement the intermediate approach of Language-Specific Phonological Inference, modeled after Darcy et al. (2009), in ASR models. Examining what these 'ideal listeners' do lets us better understand the role of different types of knowledge (phonetic, phonological) in perceptual compensation.

In terms of the methodology, we build ASR models with different *types* (i.e. phonetic, phonological) and *complexities* of linguistic knowledge. We treat these computational models, containing different kinds of knowledge, as 'ideal listener' models of humans, which we give as input the same experimental stimuli as Darcy et al.'s behavioral experiment. Examining which linguistic knowledge is necessary for these models to show human-like patterns provides a possible answer to what knowledge is necessary for humans to realize perceptual compensation.

In this section, we first introduce with examples the phenomenon examined by Darcy et al. – language-specific phonological assimilation in English and French – and the theoretical account they proposed. We then present the paradigm used in their behavioral experiments, which we follow in our computational experiments. We further summarize their key findings, which serve as the human benchmark to which computational models are compared in the current study, then outline the experiments performed in our study.

4.2.1 Language-specific phonological compensation

In running speech, certain phones undergo predictable changes when the context meets certain conditions. Phonological assimilation is one such case, where sounds at the begining or ends of words are changed so that they are more similar to adjacent sounds across a word boundary.

Different languages show different assimilation patterns across word boundaries; for example, English shows regressive nasal place assimilation and French shows regressive voicing assimilation. In Table 4.1, (1.1) illustrates an example where the /n/ in *own* is 'assimilated' to the following /p/ as [m], as [m] is the labial equivalent of [n]. English listeners are able to perceptually compensate for the change of the sound and restore its original form. In other words, English listeners 'undo' the assimilation in the sense that they hear the assimilated sound (i.e. [m]) as the unassimilated version (i.e. [n]). On the other hand, French listeners fail to compensate for place assimilation, which does not occur in French. However, French does have another type of assimilation in voicing. In (2.1), the voiced /b/ sound in *robe sale* is pronounced as *ro[p] sale* due to assimilation to the following [s] sound, which is voiceless. French listeners show a compensation effect in these cases, while English listeners do not (Darcy et al., 2009). In summary, listeners are able to compensate for the type of assimilation that occurs in their native language.

The theoretical approach underlying Darcy et al. (2009) is Language-Specific Phonological Inference, which treats compensation as a language-specific mechanism that undoes the effect of assimilation rules that apply during phonological planning in production. This theory predicts that the pattern of compensation depends on the listener's language (Coenen et al., 2001; Gaskell, 2003; Gaskell et al., 1995; Gaskell and Marslen-Wilson, 1996, 1998), which accounts for the fact that

Language	Assimilation Type	Target Word	Sentence (Phonetic realization)	Condition
English	Place	own /n/	(1.1) [...] it's my *ow[m]* plan.	Viable Change
			(1.2) [...] it's my *ow[m]* choice.	Unviable Change
			(1.3) [...] it's my *ow[n]* choice.	No Change
French	Voicing	robe /b/	(2.1) la *ro[p]* sale [...]	Viable Change
			(2.2) la *ro[p]* noire [...]	Unviable Change
			(2.3) la *ro[b]* rouge [...]	No Change
English	Voicing*	big /g/	(3.1) the *bi[k]* fountain [...]	Viable Change
			(3.2) the *bi[k]* river [...]	Unviable Change
			(3.3) the *bi[g]* lighthouse [...]	No Change
French	Place*	lune /n/	(4.1) la *lu[m]* pâle [...]	Viable Change
			(4.2) la *lu[m]* rousse [...]	Unviable Change
			(4.3) la *lu[n]* jaune [...]	No Change

Table 4.1: English and French examples. Star (*) refers to *illegal* (i.e. non-native) assimilation types (English voicing assimilation and French place assimilation) illustrating stimuli from Darcy et al. (2009). Non-starred types refer to *legal* (i.e. native) assimilation types.

French and English listeners fail to compensate for place and voicing assimilation, respectively.

However, there are confounding factors which make it difficult to isolate the exact cause of this failure to compensate. Previous studies (to Darcy et al.'s) had used non-native utterances, introducing both non-native phone sequences (including phonological knowledge and phonetic knowledge) and non-native assimilation mechanisms (e.g. Mitterer and Blomert, 2003; Weber, 2001). One possibility is that it is the non-native (and hence ill-formed) phone sequences and phone categories that result in the failure to compensate, and if given native utterances showing 'non-native' assimilation patterns (e.g. voicing assimilation in native English speech), listeners would be able to compensate. It is also possible that listeners will not be able to compensate for a non-native assimilation pattern even in this case. Thus, in order to understand the source of language-specific compensation, one needs to dis-

entangle the source of ill-formedness in the experiment stimuli, which is accounted for in Darcy et al. (2009)'s experimental design.

4.2.2 Summary of Darcy et al. (2009)

Darcy et al. (2009) refines our understanding of language-specific phonological compensation by avoiding the confounds introduced by using non-native utterances. An important contribution from Darcy et al. (2009) is using an experiment paradigm that separates the two components, where the productions are *legal* phone sequences (i.e. in the participant's native language) but with *illegal* (i.e. non-native) assimilation patterns, deliberately produced by the speakers (e.g. *Unviable Change* conditions in Table 4.1). They found that even with legal phone sequences, listeners still fail to compensate as much for non-native assimilation patterns as native listeners do.

Main Experiment

In their study, Darcy et al. (2009) conducted a word detection task where listeners heard native utterances with native and non-native assimilation patterns. In the experiment, listeners first heard the target word produced by a male speaker in the standard form without undergoing assimilation. After 500ms of silence, listeners heard a sentence produced by a female speaker, containing the same target word which could be one of the three types (illustrated in Table 4.1) – *Viable Change*, *Unviable Change*, and *No Change* – and one of the native or non-native assimilation types. Specifically, *Viable Change* refers to the assimilation occurring in the correct phonological context (following phone), whether it is native (such as 1.1) or non-native (such as 3.1). *Unviable Change* refers to the assimilation occuring in an incorrect phonological context (such as 1.2 where *ch* is not a bilabial sound and hence does

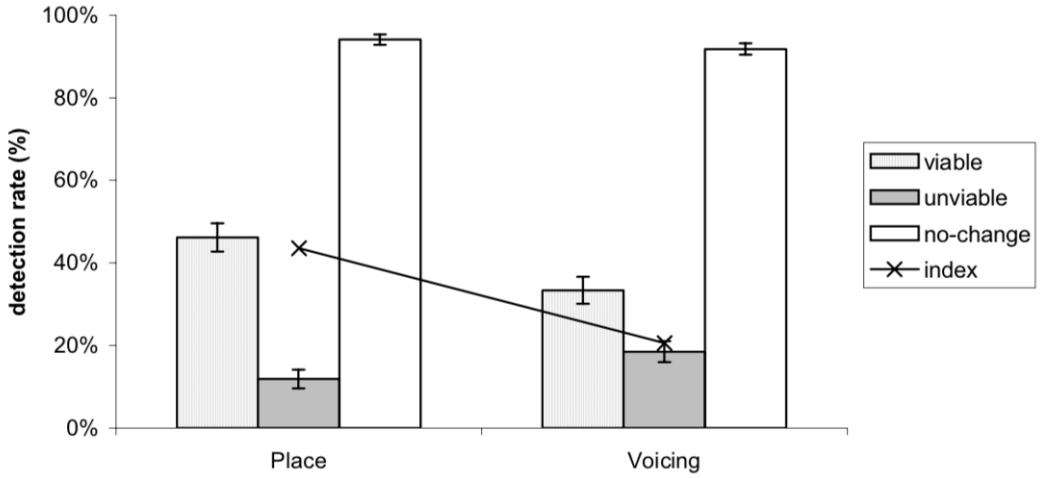
not licence the preceeding *n* to undergo place assimilation). *No Change* refers to the canonical pronunciation, where no assimilation occurs. After hearing both stimuli, the listeners decided whether or not the sentence contained the target word they had heard earlier.

The results reflect listeners' ability to identify the words with the canonical pronuncition in different contexts (*No Change*), detect assimilation (*Viable Change*) and spot unlicenced variants (*Unviable Change*). Identifying *Unviable Change* versus *No Change* can also be thought of as their ability to detect minimal pairs.

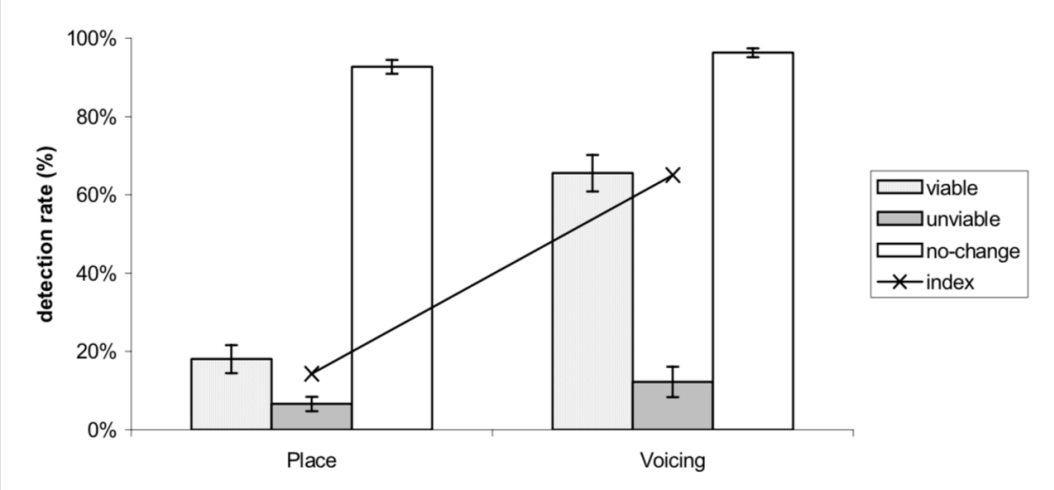
The barplots in Figure 4.1 represent listeners' percentage of detecting the same target word in the sentences in different conditions and assimilation types. In order to capture the degree of assimilation detection relative to the two baseline conditions (i.e. *Unviable* and *No Change*) and make results of English and French listeners comparable, the authors define a Compensation Index (Equation 4.1). The index calculates a ratio of *Viable Change* to *No Change*, while controls for the perceptual biases or error in the *Unviable Change* conditions. Intuitively, a higher Compensation Index indicates more detection of assimilation, and hence more compensation.

$$Compensation\ Index = \frac{(\%detection_{viable} - \%detection_{unviable})}{(\%detection_{no-change} - \%detection_{unviable})} \tag{4.1}$$

The results reveal that even with native phone sequences, listeners still fail to compensate for non-native assimilation. Figure 4.1 shows that English listeners show both higher detection rate for *Viable* conditions and higher Compensation Index for place assimilation, compared to the non-native voicing assimilation. They still detect more *Viable* cases than *No Change* cases for voicing assimilation, indicating that they can compensate to some extent, but the degree of compensation is much lower than for native place assimilation. The opposite pattern is found for French listeners, where the native type of assimilation is voicing assimilation.



(a) Native English listener responses from Darcy et al. (2009)



(b) Native French listener responses from Darcy et al. (2009)

Figure 4.1: The detection rates of the target words in the sentences and Compensation Indices from Darcy et al. (2009).

We define two patterns observed from human listeners' responses, which we use later to compare with our models' responses. First, human listeners are able to distinguish *minimal pairs* (i.e. *Unviable* and *No Change*) in all four cases, no matter whether the assimilation type is native or non-native. Second, the Compensation

Indices show the following pattern: English listeners show a larger Compensation Index for place assimilation than for voicing assimilation, while French listeners show a smaller Compensation Index for place assimilation than for voicing assimilation.

Control Experiment

As the stimuli had been produced by speakers who were asked to purposely produce non-native assimilation patterns (i.e. illegal phone sequences), such as *bi[k] fountain*, and *Unviable* assimilation patterns (i.e. wrong pronunciation), such as *ow[m] choice*—neither of which occurs in natural speech—Darcy et al. conducted a control experiment to make sure that the unnatural productions can be unambiguously perceived. In the control experiment, a different set of participants heard the phonetic realizations of the target words, spliced out from the carrier sentences, and were asked to identify the final consonant by choosing between the original sound and the assimilated sound. If neither of the two matched, the participant could write down the sound they heard.

The control experiment verified that all target words can be unambiguosly perceived. As shown in Table 4.2, while listeners were not 100% perfect in their performance, they indeed perceived the *Viable* cases to be 'different consonants', patterning with the *Unviable* cases. On the other hand, *No Change* cases were mostly judged to have 'similar consonants'. The above pattern holds for both languages regardless of the type of the assimilation, although English listeners were less certain about the stimuli than French listeners (% different for all conditions were closer to chance). In summary, without the following context, listeners were not able to compensate for any kind of assimilation.

Language	Condition	Consonant different from unchanged target (%)	
		Place (SD)	Voicing (SD)
English	Viable change	74 (3)	78 (1)
	Unviable change	78 (2)	77 (1)
	No change	23 (4)	17 (3)
French	Viable change	92 (0.9)	95 (0.7)
	Unviable change	90 (1)	97 (0.5)
	No change	9 (2)	2 (0.2)

Table 4.2: Results from the two control experiments (English and French) in Darcy et al. (2009) (combining Table 5 and Table 3 from the original paper).

Summary

Taken together, these results indicate that compensation for assimilation depends on a viable context. Furthermore, listeners use language specific knowledge to restore the assimilated phone. This indicates that compensation for assimilation must rely on learned knowledge, not just generic processing of the acoustic signal. However, this study does not isolate what aspects of language specific knowledge are involved. As mentioned earlier, even when processing legal (native) phone-sequence utterances, listeners need to make use of both their lower-level phonetic knowledge and higher-level phonological knowledge, such as phonotactics and syllable structure. Different kinds of linguistic knowledge are not clearly delineated in Darcy et al. (2009)'s experiments. We now turn to our study, which investigates their relative contribution.

4.2.3 Current Study

The current study aims to better understand the role of different kinds of linguistic knowledge in language-specific compensation: lower-level phonetic knowledge, higher-level phonological knowledge, and how they interact with the type of assim-

ilation pattern (native vs. non-native). We use ASR models to simulate the experiments in Darcy et al. (2009). One of the advantages of using computational models is to control what knowledge is available to the learner. For example, while humans always have a lexicon and cannot be unlearnt, any experiments done to human participants will get lexical interference from their lexical knowledge. Computational models, however, allows us to investigate exclusively on the effects of phonology without the confounds introduced from lexical bias. In the current study, we use computational learners trained without a lexicon and hence tease apart the lexical knowledge from other linguistic knowledge.

In order to examine the role of phonetic versus phonological knowledge, we model the two using separate components in the ASR system (as detailed later). For each type of knowledge (i.e. model component), we implement models with different degrees of complexity to examine how much linguistic knowledge a learner (i.e. the model) needs to achieve human-like behavior, and how varying the knowledge available to the learner affects performance.

In terms of evaluating how human-like the models are, we use a two-step process, corresponding to the two patterns humans show summarised at the end of section 4.2.2. We first filter out the models that are unable to distinguish minimal pairs (i.e. *No Change* vs. *Unviable*). We then evaluate the remaining models based on their Compensation Index patterns to select the most human-like models.

The current study includes three experiments, summarized in Table 4.3. Experiment 1 corresponds to Darcy et al. (2009)'s main perceptual experiment, which examines whether or not language-specific compensation is possible given legal phone sequences with non-native (and native) assimilation.

Experiment 2 corresponds to Darcy et al. (2009)'s control experiment. The goal is to further understand the ASR models by testing if they perform similarly to humans in cases where humans are not able to compensate for assimilation.

In Experiment 3, we simulate a follow-up study to Darcy et al. (2009), testing if any compensation is possible given *illegal* (i.e. non-native) phone sequences with native (and non-native) assimilation. In other words, we conduct the same experiment as Experiment 1 but using non-native speech: presenting the French sentence stimuli to English models, and vice versa. The findings shed light on how illegal utterances (forcing listeners to use non-native phonetic/phonological knowledge) interact with the assimilation type.

Experiment	Description of the test stimuli
Experiment 1	*legal* phone sequences + (non-)native assimilation
Experiment 2	*legal* phone sequences *without* following context
Experiment 3	Non-native speech: *illegal* phone sequences + (non-)native assimilation

Table 4.3: Summary of experiments in the current study.

4.3 Methods

Given the goal to investigate the importance of phonetic knowledge and phonological knowledge for perceptual compensation, HMM-GMM ASR models (Jelinek, 1997) are a good fit, mainly for their interpretability of learnt linguistic knowledge. Such models typically consist of two components, *acoustic models* and *language models*. The learnt knowledge of the two components best matches the two types of linguistic knowledge (i.e. the phonetic and phonological). We choose HMM-GMM models over state-of-the-art end-to-end DNN models due to the lack of interpretability of the latter type, as these are less modular, and much harder to interpret. Thus, even if the model shows good performance and human-like behavior, it is hard to explain what linguistic knowledge the model learns and how the model is able to achieve this performance.

We trained a set of phone recognition models on English and French respectively to represent native listeners. The acoustic model (AM) maps phones to acoustics, which can be seen as a listener's phonetic knowledge. The language model (LM) captures the statistical distribution of phone sequences, which represents a listener's phonological knowledge, such as phonotactics and syllable structure. The language model also indirectly captures lexical knowledge to some extent (as likely sequences of phones).

An important clarification concerns the choice of the LM being over phone sequences in the current study. Typically, an ASR model contains a word-level LM, as in most cases the model's task is to transcribe speech to words. However, since the purpose of the current study is to examine the phonetic and phonological knowledge used in perceptual compensation, a word-level LM is not suitable for reasons of 1) practicality and 2) experimental design.

First, implementing a word-level LM puts a strong bias towards recognising legal words. In other words, all production variants, including wrong productions, must be 'corrected' as real words, as the ASR model can only choose from its lexicon. In the current experimental design, we purposely include nonwords which we would expect models to distinguish from real words. However, the nonwords will actually always receive zero probability, as they do not exist in the lexicon. This would result in the model never being able to consider nonwords, even when the stimuli do contain them, and is thus inappropriate.

Second, the original experiment is designed to draw listeners' attention "on the detail of pronunciation of words, i.e. on the *form* of words and not to the mere presence or absence of a target word in the sentence" (Darcy et al., 2009, p. 279), as a word is still recognizable even if it is altered from its canonical form (see discussion in Darcy et al., 2009). In this sense, we prefer the phone-level LM over the word-level LM, as it emphasizes the form rather than the presence or absence of the word.

4.3.1 Models as ideal listeners

In human speech perception and automatic speech recognition, the task (for a listener/AST model) is to infer the most likely sequence of phones and words given the acoustics they hear. As explained in the previous section, we focus on phones in this study; the corresponding ASR task is called 'phone recognition'. The mathematical formalization of the speech perception process is shown in Equation 4.2, where the decoding process can be seen as model's equivalence of speech perception. $P(q|X)$ refers to the most likely sequence of phones given the speech signal, where q stands for any possible phone sequence and X stands for the acoustics. $\hat{q}$ is the phone sequence which maximizes the posterior probability, given the observed acoustics $P(q|X)$.

$$\hat{q} = \arg\max_{q} P(q|X) \tag{4.2}$$

The equation is further broken down according to Bayesian inference to show how the acoustic model and language model jointly determine the posterior (4.3). $P(X|q)$ is the likelihood of the acoustics given the phone sequence, captured by the acoustic model (AM). Linguistically speaking, $P(X|q)$ is a probability model which specifies how likely different acoustic realizations (X) of a phone sequence (q) are, over a set of acoustic features which parametrize the speech signal, usually a discretized spectrogram such as MFCCs. $P(q)$ is the prior probability of the phone sequence, captured by the language model (LM).

$$\hat{q} = \arg\max_{q} P(X|q)P(q) \tag{4.3}$$

Since compensation for assimilation is context-sensitive, one needs contextual information through the phonological knowledge (LM), or phonetic knowledge (AM),

or a combination of the two. In order to capture different degrees of contextual knowledge, we implemented a variety of AMs and LMs with different complexities. We illustrate below the specific types of AMs and LMs, and our predictions for the corresponding models.

Language Model Phone-level language models are *n-gram* WFSA (weighted finite state automata), which model the distribution of *n*-phone sequences. We trained four types of LMs of different complexities, by varying *n*: 1) a flat (null) LM, where the probability of the next phone is the same for all phones; 2) a unigram LM, where the probability of the next phone is equal to the probability of the phone occuring in the language; 3) a bigram LM, where the probability of the next phone is conditioned on the previous phone; 4) a trigram LM, where the probability of the next phone is conditioned on the previous two phones.

Conceptually, the four types of LMs capture an increasing degree of complexity of contextual phone patterns. As shown in Table 4.4, we divide the four LMs into two categories by their abilities to capture context information. The *simple* category (i.e. flat and unigram) has no-or-limited phonological knowledge, does not capture any information about phone sequences, and thus cannot encode any contextual information. The *complex* category (i.e. bigram and trigram) has more complex phonological knowledge, and can encode information about phone sequences. The LMs incorporated in successful models represent the complexity of phonological knowledge needed in terms of modelling phone sequences. We discuss below the specific predictions made by different LMs together with the AMs which we introduce next.

Acoustic Model We trained three types of acoustic model for mapping phones to acoustics. The least complex is a *monophone AM*, a *context-independent* phone-to-acoustics mapping, which does not take adjacent phones into consideration. For

example, if phone *m* occurs in two contexts *a_a* and *b_b* and is pronounced slightly differently, a single representation of phone *m* must cover both of those contexts. **Prediction**: as this type of AM has limited contextual knowledge, we expect that it may only be successful when combined with a *complex* LM that learns some knowledge of phone sequences (i.e. bigram and trigram), but not *simple* LMs (i.e. flat and unigram).

Second is a *triphone AM*, an acoustics-phone mapping which is *context-dependent*. Using the same example as above, the phone *m* is modelled differently in the two contexts (i.e. *a_a* and *b_b*), while the final output is still *m*. **Prediction**: since this type of AM captures information about the neighbouring context, it is possible that no additional phonological knowledge is needed. We predict that triphone AMs in combination with *simple* LMs may be successfull, but are less similar to humans compared with *complex* LMs.

Third, a *triphone speaker-adapted (triphone-SA) AM*, which corrects the representation of the phone *m* for different speakers. For example, the phone *m* produced by two different speakers will be modelled differently, while non-speaker-adapted AMs only share the same represenation across speakers. While speaker adaptation is not the central focus of the current study, we nevertheless include this type of AM because it explicitly models one ability humans are able to use to deal with variability in speech. **Prediction**: given that there are only two speakers in the experiment stimuli, we predict that the performance of triphone-SA AMs is similar to triphone AMs.

Baseline In order to better understand the improvements made through incorporating different kinds of linguistic knowledge into the model, we included raw acoustics (i.e. MFCCs, the input for other ASR models) as the baseline, where no linguistic knowledge is learnt. This baseline model receives no training and has to

rely on solely the acoustic information for the experiments. It is *not* a phone recognizer as those described above. Instead, the baseline merely calculates the raw acoustic distance between the words in the sentences and the target words in isolation. **Prediction**: the baseline model captures the information carried in the speech signal alone. We expect that it will be outperformed by other ASR models that learn different degrees of linguistic knowledge.

		AM		
		Monophone	**Triphone**	**TriphoneSA**
LM	**Flat Unigram**	Context-independent phonetic knowledge No/ limited phonological knowledge	Contextual phonetic knowledge	
	Bigram Trigram	Context-independent phonetic knowledge Phonological knowledge	Contextual phonetic knowledge Phonological knowledge	
Baseline	No phonetic or phonological knowledge (raw acoustics)			

Table 4.4: Description of the 12 models used in Experiment 1, which differ by LM and AM, plus one baseline model (MFCCs). The 12 models are dicided up by broad types of AM and LM; see text.

4.3.2 Procedure

Training

All training was done using *Abkhazia* (Schatz et al., 2016), a Kaldi-based (Povey et al., 2011) speech recognition package. Training data were 46 hours in English from Librispeech for the English models and 36 hours in French from the data used for the Zero Resource Speech Challenge 2017 (Dunbar et al., 2017). Input features were 39-dimension Mel Frequency Cepstral Coefficients (MFCCs) with Δ and $\Delta\Delta$ extracted from the audio, with window length of 25ms and step size of 10ms (meaning one frame = 10 msec). The training can be done by following the recipes [1] provided by *Abkhazia*.

[1] In the command line tool, the `language` and `acoustic` recipes for LM and AM respectively.

Experiment Simulation

After training all the models, we conducted the same experiments as Darcy et al. (2009). The process is illustrated below in Figure 4.2. The task for the model is to decide whether or not the sentence presented to it contains the same target word produced in isolation (by a different speaker). In particular, the model receives the sentence stimuli and decodes over the entire sentence. The decoding process (Equation 4.3) can be seen as analagous to speech perception, where the model finds the best phone sequences. We extracted the frame-level phone posteriors (i.e. the estimated probability of phones at each frame), to represent the model's 'mental representation' of perceived sounds. The same decoding was done to the target word in isolation. For the baseline acoustics, as it is not a phone recognizier, it does not output phone posteriors. Thus, we instead use MFCCs directly to represent information carried by raw acoustics.

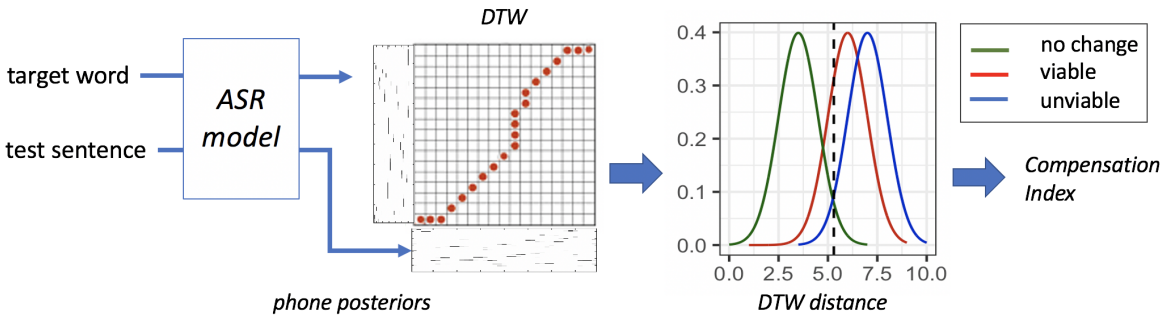


Figure 4.2: The procedure of model evaluation for ASR models. Note that for the baseline acoustics, phone posteriors are replaced by MFCCs.

After extracting the phone posteriors, we extracted the frames corresponding to the target words in the sentences and calculated the distances between the *word in carrier sentence* (i.e. the words spoken by the female talker in the carrier sentence) and *word in isolation* (i.e. the target word produced by the male talker). Distances

were calculated using Dynamic Time Warping (DTW), where a larger value indicates larger difference between the pair. DTW is an algorithm for measuring the similarity between two temporal sequences which may vary in length, and calculates the total *frame-level distance* along a path that optimally stretches the time axes to align the two words.

To define a frame-level distance metric: as each frame corresponds to a vector of probabilities (posterior over phones), we used Kullback– Leibler (KL) divergence, which quantifies how much a one probability distribution differs from the true distribution. We used the isolated word as the true distribution, so the KL divergence measures the differences in the distributions of phone probabilities for carriers and target words in isolation.

We treat the DTW distance calculated from the ASR models as equivalent to humans' proportion of *same* (vs. *different*) judgements. As a higher DTW distance indicates a larger difference between the pair, we interpret higher DTW distance as a greater chance of *different* judgements. The next section presents how the ASR models are evaluated, and comparison with human responses.

4.3.3 Evaluation of models

We evaluate two aspects of each model's performance: 1) its ability to compensate for assimilation, the higher level qualitative pattern, and 2) its similarity with human performance at the stimuli level.

First, as any model that manages to compensate for assimilation should first be able to differentiate between the two phones, we test whether or not the models can distinguish the minimal pairs (i.e. *No Change* and *Unviable Change*). We only keep the *MP-distinguishing models* for further evaluation. Next, we examine these

for their ability to compensate for assimilation, and whether or not they show the same language-specific pattern as humans.

In terms of the ability to compensate for assimilation, we define three levels of model goodness, filtered according to the two patterns exhibited by humans (Section 4.2.2): 1) *bad* models, which fail to distinguish the minimal pairs (i.e. *Unviable* vs. *No-Change*); 2) *MP-distinguishing* models, which distinguish minimal pairs in either languages, but may or may not show the same language-specific compensation patterns as humans; 3) *Compensating* models, a subset of *MP-distinguishing* models, which display language-specific compensation effects in *both* languages. The selection process is illustrated in Figure 4.3 and we explain the three levels in more detail below, referring to the schematic plots in Figure 4.4.

Models with the highest goodness level, 'compensating models', correctly reflect the language-specific compensation pattern. Nevertheless, because there may be several 'compensating' models, this does not show which AM-LM configuration best approximates human behavior. In order to examine how 'human-like' the compensating models are, we fit a set of mixed-effects logistic regression models. We use the variance explained by the statistical model as an indicator of how well the ASR model predicts the human responses.

***Bad* Models** Bad models refer to those that fail to discrimiate between *Unviable* and *No-Change* conditions. If a model does not distinguish the unambiguous minimal pairs in the first place, then it does not make sense to discuss 'compensation' using the ambiguous *Viable* condition where the phone potentially undergoes assimilation, since this can only be judged relative to the *Unviable* and *No-Change* endpoints. It is important to first discard such models, as they can show false-positive correct compensation patterns. Figure 4.4 (left) illustrates a case where including bad models can be problematic. While the model is unsuccessful in telling apart

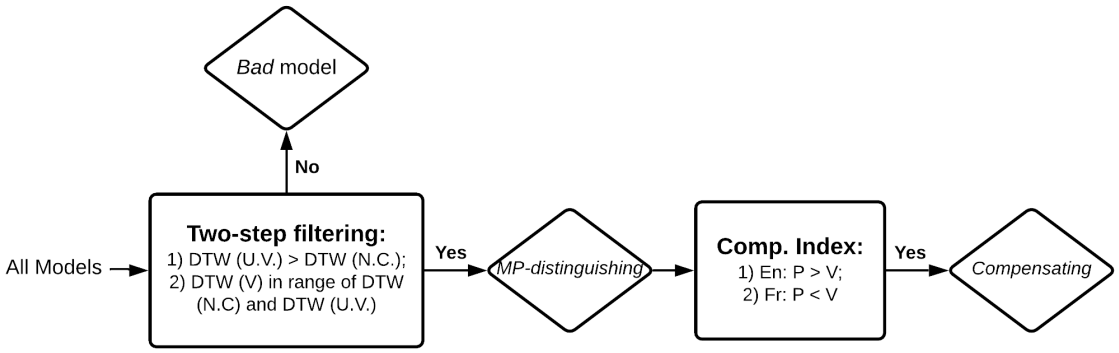


Figure 4.3: The selection process to determine *bad* models, *MP-distinguishing* models and *compensating* models. U.V=*Unviable*, V=*Viable* and N.C.=*No Change*.

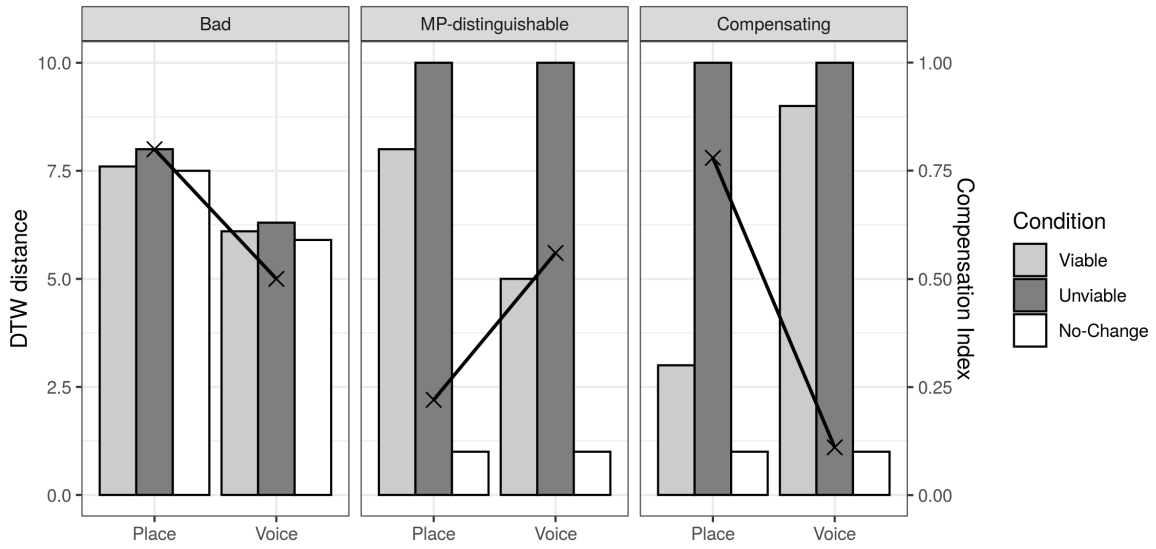


Figure 4.4: Schematic plots for three types of English models: Bad (left), MP-distinguiable (middle), Compensating (right). Left y-axis: DTW distance, higher means larger difference (lower detection rate). Right y-axis: Compensation Index (crosses, calculated from DTW distances).

Unviable and *No-Change*, it shows the expected Compensation Index pattern for an English listener—higher Compensation Index for place assimilation than voicing assimilation—even though the Compensation Index is meaningless in this case.

Thus, the Compensation Index only makes sense to be used for models that can distinguish minimal pairs, which we call *MP-distinguishing* Models.

Before inroducing the filtering process for *MP-distinguishing* models, we first present our variant of the Compensation Index calculated for each ASR model. Recall that Darcy et al. (2009) calculates the Compensation Index from participants' detection rates (i.e. proportion of the stimuli perceived to be 'same' for each condition) for the three conditions (Equation 4.1).

Our definition of Compensation Index (Equation 4.4) for the ASR models varies slightly from the original Compensation Index. Equation 4.1 requires first calculating $\%detection_{viable}$, $\%detection_{unviable}$ and $\%detection_{no\text{-}change}$. However, unlike humans' *same* vs. *different* responses which can be directly counted as 1's and 0's, the models output DTW distances. One can either force a threshold on the model to decide if a stimulus is same or different (then calculate % different), or use DTW distance as a similarity measure. A previous version of this study (Jiang et al., 2020) uses the former; however, further investigation suggested these thresholds may not be reliable for various reasons (e.g. the threshold can be affected by extreme values). Therefore, in the current analysis, instead of asking the models to make deterministic decisions on the phone category, we use the DTW distance to reflect more fine-grained similarity measures.

$$Compensation\ index_{model} = \frac{(DTW\ distance_{viable} - DTW\ distance_{unviable})}{(DTW\ distance_{no-change} - DTW\ distance_{unviable})} \quad (4.4)$$

***MP-distinguishing* Models** Conceptually, *MP-distinguishing* models are those that manage to distinguish the minimal pair, and at the same time identify the *viable* cases within the similariy range (i.e. DTW distance) defined by the minimal pair. In other words, the *Viable* cases are not more different than *Unviable* cases, and not more

similar than *No-Change* cases. However, such models may not necessarily show the human-like compensation pattern. The subset of *MP-distinguishing* models showing correct language-specific compensation are termed *Compensating* models, introduced in the next section.

An example of *MP-distinguishing* model (non-compensating) is demonstrated in Figure 4.4 (middle). Shown by the large difference in DTW distances, the *Unviable* conditions are clearly distinguished from *No-Change* conditions for both types of assimilation. Moreover, the *Viable* conitions are less similar than *Unviable* conditions (smaller DTW distance) but more different than *No-Change* conditions. However, the Compensation Index shows a higher compensation for voicing assimilation than place assimilation, opposite to what a native English listener would have perceived.

In order to apply this intuition of model selection in qualitative evaluation, we adopt a two-step filtering process (Figure 4.3). First, the qualified *MP-distinguishing* models should show significant difference in the DTW distance distributions between the minimal pairs (i.e. *unviable* vs. *no-change*). Second, the DTW distance distribution of the *viable* cases are not significantly larger than *unviable* cases and not significantly smaller than *no-change* cases (i.e. *viable* cases are not 'more different' than *unviable* cases, and at the same time not 'more similar' than *no-change* cases). A model is *MP-distinguishing* only if both criteria are met.

Compensating Models Compensating models are a subset of *MP-distinguishing* models that correctly reflect language-specific compensation effects. An English model needs to show higher compensation for place assimilation than voicing assimilation, and vice versa for a French model.

Figure 4.4 (right) presents a possible compensating model (English). First, it meets all criteria of a *MP-distinguishing* model: clear distinction between the minimal pair, with *Viable* condition being in the range of the two baselines. Moreover,

the compensating model shows higher Compensation Index for place assimilation than voicing assimilation.

Note that we only show the English model as an example for concise presentation. In practice, only those that show the correct compensation pattern under the *same* AM-LM combination for *both* English and French are regarded as compensating models.

Predictability of Human Responses After determining the set of *compensating* models which are roughly equivalent, we examine how quantitatively different these models are by using them to predict human responses, across all experimental stimuli. For each *compensating* model, we fit a mixed-effects logistic regression to predict human reponses (same vs. different) using the model's DTW distances. We use two measures—the marginal R^2 (Nakagawa and Schielzeth, 2013) and Area Under Curve (AUC) (Bradley, 1997) —to evaluate how well the fixed-effect terms predict the human responses, after accounting for variability among stimuli and participants. The model with the highest score best approximates human behavior.

4.4 Results

4.4.1 Experiment 1: Simulation of Darcy et al. (2009)

In order to get a good estimate of the mean DTW distances for each condition, we first fit a linear mixed-effects model (LMEM) (Bates et al., 2015) for each ASR model's responses, then obtain an LMEM-predicted DTW distance for each condition. A more straightforward option would be to use the average raw DTW distance for each of the three conditions. However, these means may not be representative of the DTW distance distribution for each condition, due to the large amount of variability from factors not of interest for comparing the conditions, such as the car-

rier sentence, the target word, etc. Our LME models control for these factors using random effects, giving more reliable estimates of DTW distance as a function of condition and assimilation type (the fixed effects). Unless otherwise indicated, the DTW distances discussed in the paper are all LMEM-estimated DTW distances.

We fit LMEMs to DTW distances for each of the ASR models for a total of 26 models (13 models * 2 languages).[2] The fixed effects are the three `Conditions` (*Viable*, *Unviable*, *No-Change*) and assimilation `Type`, and the interaction between them. The `Condition` variable is Helmert-coded and the `Type` variable is centered (coded as -0.5, 0.5). In terms of the random effects, we include a by-item (i.e. target word) random intercept to account for the variability from different target words. We further include a by-frame (i.e. carrier sentence) random intercept nested within item, as each `item` corresponds to three `frames`, and each `frame` has three `Conditions`). [3]

MP-distinguishing models

Recall that in order to select models showing language-specific compensation behaviour, we use a two-step selection process (Figure 4.3). The first step is to determine models that show the correct order of similarity between the *word in the carrier sentence* and the *word produced in isolation*. Translating into DTW distances, the *MP-distinguishing* models should show that 1) $DTW_{Unviable}$ is significantly larger than $DTW_{No\text{-}Change}$ and 2) DTW_{Viable} is not significantly larger than $DTW_{Unviable}$ or significantly smaller than $DTW_{No\text{-}Change}$.

In order to select *MP-distinguishing* models, we compare the similarity across the three conditions and evaluate the significance of differences between each of the two conditions using the LMEMs fit to the data. Specifically, we use the `lsmeans()`

[2]Model syntax: `lmer (dtwDistance ~ condition.helm * type.std + (1|item) +(1|item:frame),data = df)`. We fit English models to English participants and French models to French participants.

[3]More complex random effects structures, including random slopes, were tried, but produced ill-conditioned models, probably due to small sample size.

function from the `lsmeans` R package (Lenth, 2016) to calculate the statistical significance of the difference in means between *Unviable* and *No-Change* (marginalizing over items and frames). For the significant cases, we further check if the *Viable* mean is predicted to be significantly larger than the *Unviable* mean, or significantly smaller than the *No-Change* mean (marginalizing over items and frames). We found that all models showing a significant difference between the minimal pairs (*Unviable* and *No-Change*) also predict the correct similarity ordering for *Viable*; thus, no models were discarded in this second step. Table 4.5 reports all models from both languages, with uncorrected p-values for each type of assimilation. *MP-distinguishing* models, where p is lower than $\alpha = 0.01$, are shown in bold. Given that we did a large number of comparisons, we chose a relatively low α cutoff to reduce the likelihood of false positives.

The results (Table 4.5) show that linguistic knowledge is indeed needed in order to distinguish minimal pairs. English models can distinguish minimal pairs whenever the AM is triphone. French models, on the other hand, mostly manage to distinguish minimal pairs, with monophone-flat, monophone-trigram and triphone-unigram being the only exceptions.

As for the baseline acoustics (MFCCs), which do not make use of any linguistic knowledge, they fail to distinguish the minimal pairs in neither of the languages. Table 4.5 shows that MFCCs do not exhibit significant differences between minimal pairs. Moreover, Figure 4.5 reveals that the *Unviable* and *No-Change* clearly overlap for the majority of the cases. The indistinguishability of the acoustics suggests that the minimal pairs are acoustically ambiguous enough not to be separated based on pure acoustics, and that linguistic knowledge plays an important role for realizing the minimal pair contrast.

English			French		
Model (AM-LM)	p_{place}	p_{voicing}	Model (AM-LM)	p_{place}	p_{voicing}
MFCCs	0.673	0.758	MFCCs	0.205	***
monophone-flat	0.838	0.122	monophone-flat	0.011	0.015
monophone-unigram	0.930	0.068	**monophone-unigram**	***	***
monophone-bigram	0.715	***	**monophone-bigram**	***	***
monophone-trigram	0.898	***	monophone-trigram	0.027	***
triphone-flat	**0.009**	***	**triphone-flat**	***	***
triphone-unigram	***	**0.003**	triphone-unigram	0.028	***
triphone-bigram	***	***	**triphone-bigram**	***	***
triphone-trigram	**0.002**	**0.005**	**triphone-trigram**	***	***
triphoneSA-flat	0.089	***	**triphoneSA-flat**	***	***
triphoneSA-unigram	0.048	***	**triphoneSA-unigram**	**0.003**	***
triphoneSA-bigram	0.208	***	**triphoneSA-bigram**	***	***
triphoneSA-trigram	0.069	***	**triphoneSA-trigram**	***	***

Table 4.5: p-values for minimal pair contrasts in place and voicing, for each ASR model, for English and French data. Bolded models are *MP-distinguishing* models (where $p < \alpha = 0.01$) and underlined models are *MP-distinguishing* models found in both languages. Uncorrected p-values (from `lsmeans()` post-hoc comparisons) are reported by assimilation type. *** indicates $p < 0.001$.

Compensating Models

We have shown that in order to distinguish minimal pairs in both languages (bolded and underlined in Table 4.5), an ASR model needs a triphone AM. This suggests that contextual phonetic knowledge is required for discriminating minimal pairs, in English and French. We now check whether or not those *MP-distinguishing* models also show the language-specific compensation pattern: place $>$ voicing for English and place $<$ voicing for French. As the *MP-distinguishing* models share the same AM and differ only in their LM, this section further answers the question of how much phonological knowledge is needed for an ASR model to show language-

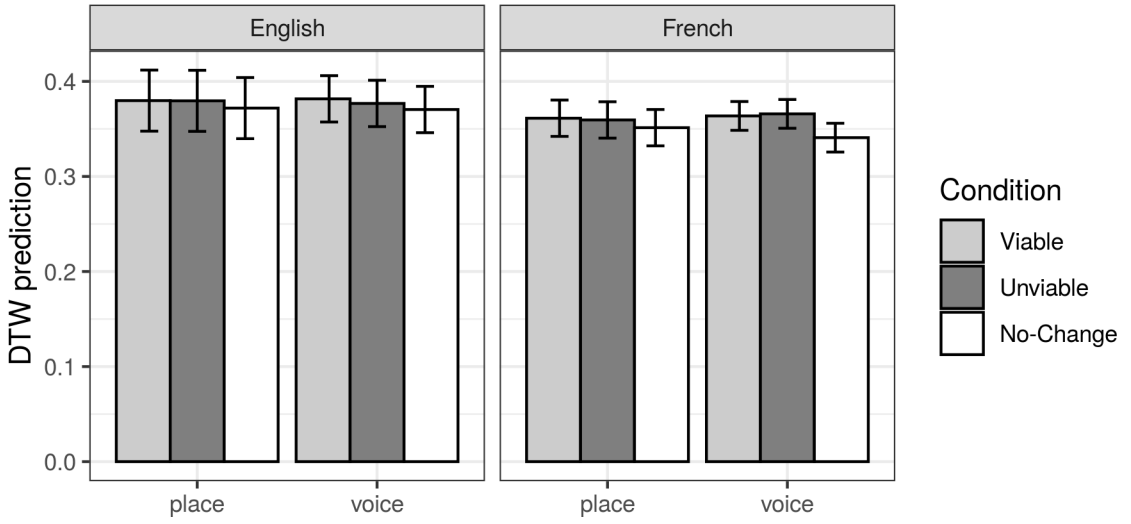


Figure 4.5: DTW distances (LMEM-estimated) from MFCCs for the three conditions in English and French. Error bars stand for 95% confidence interval.

specific compensation patterns, given that the model already has contextual phonetic knowledge.

One clarification concerning the Compensation Index used in this section is that negative values are adjusted to be 0. The reason is the following: recall that in the calculation in Equation 4.4, if *Viable* is less similar than *Unviable* ($DTW_{Viable} > DTW_{Unviable}$), the numerator is positive (while the denominator is negative), resulting in a negative Compensation Index. Since when selecting the *MP-distinguishing* models, we allowed for *Viable* to be less similar than *Unviable* as long as the difference is not significant, it is not surprising that the predicted DTW distance is sometimes in the incorrect order. As such difference is statistically insignificant, and essentially indicates that the model fails to detect assimilation, we adjusted the negative value to be 0 for easier interpretation. The updated adjusted Compensation

Index is shown in Equation 4.5.

$$Compensation\ Index_{model} = Max\Big\{\frac{DTW\ distance_{viable} - DTW\ distance_{unviable}}{DTW\ distance_{no-change} - DTW\ distance_{unviable}}, 0\Big\} \tag{4.5}$$

Compensation Indices (calculated from Eq. 4.5) for *MP-distinguishing* models in either language are reported in Figure 4.6. The left column shows English models and the right column shows French models. The specific model configuration can be determined by the row (representing LMs with increasing complexity from top to bottom) and the color (representing AMs). For comparison, human performance reported in Darcy et al. (2009) is shown in black which represents the human benchmark of their language-specific compensation pattern. [4] The results show that all triphone AMs except for the combination with unigram LM qualify as Compensating models. For example, the model with triphone AM and flat LM (first row) is compensating because it shows the Compensation Index pattern of voicing $<$ place in English and voicing $>$ place in French.

In terms of AMs, models with monophone AMs (red lines) fail not only because they cannot distinguish minimal pairs in English, but also because they display the opposite pattern to humans in French (voice $<$ place). Triphone speaker-adapted AMs (blue lines), on the other hand, fail to predict the language-specific pattern: they compensate more for voicing assimilation than place assimilation, regardless of the language.

On the other hand, LMs (shown in rows), which represent prior knowledge of phone sequences, also play an important part in whether a model shows the correct language-specific perceptual pattern. Specifically, models with a ungiram LM fail to give the correct pattern regardless of the AM, while models with the other three LMs

[4] Since the Compensation Index is computed differently for the two, the absolute *height* of the black vs. coloured lines (human vs. model) does not matter.

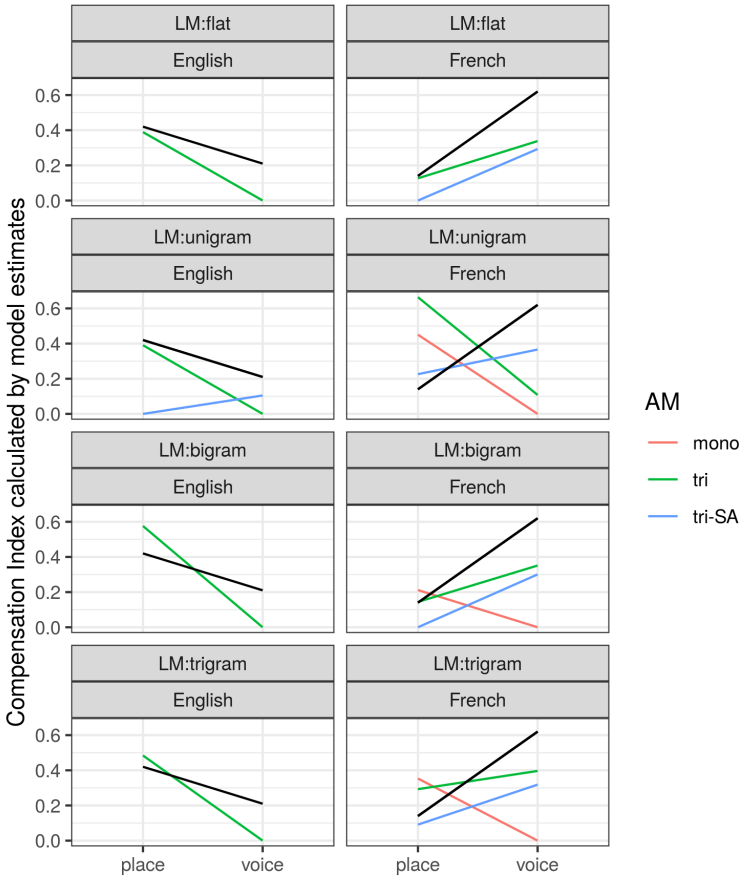


Figure 4.6: Compensation Indices of *MP-distinguishing* models in either language (bolded in Table 4.5, for ASR models varying in AM and LM (colored lines), and for human data (black lines). Compensation Indices are calculated using the LMEM-estimated DTW distances for ASR models, and using Darcy et al. (2009)'s experimental data for humans.

do show the qualitative pattern of compensation for assimiation—different slope directions for English/French—for some choices of AM.

The results show that all three *MP-distinguishing* models also conform to the *qualitative* language-specific compensation pattern. We further seek to examine whether the choice of LM makes any significant difference or all three of them are statistically

similar. In order to test the effects of LM, we further fit a LMEM [5] with LM as an independent variable and DTW distance as response variable to check the importance of LM. A likelihood ratio test comparing the original model and the model removing all interactions with the LM term does not result in significant difference (χ^2 = 14.448, df = 24, p = 0.9359). As the only qualified AM for compensating models is triphone SA, the LMEM is built only for DTW distances from ASR models that use triphone AM (not including triphone-SA).

To summarize, the only models which show human-like compensation behavior are those with a triphone AM and a non-unigram LM. The specific choice of the LM does not qualitatively affect the model's compensation ability. This result suggests that in order to realize language-specific compensation pattern, only contextual phonetic knowledge would suffice. We return to this in the Discussion section.

Predictability of human responses

Results in the previous section show that at a qualitative level, the three models (all with trigram AM) show human-like language-specific compensation patterns. In order to further test if any of these models, which differ only in LM, are 'better', we turn to a different evaluation – predictability of human responses for individual stimuli.

In order to test directly the models' predictions of human responses, we fit three mixed-effects logistic (MEL) models [6] using DTW distance to predict the human response(i.e. same/different) for each stimulus, one for each ASR model (i.e. triphone-flat, triphone-bigram, triphone-trigram). The fixed effects are the ASR model's `DTW distance` for the stimulus, the `type` of assimilation (i.e. place / voicing), and the

[5] Model syntax: `lmer(distance ~ lm.helm*lang.std*condition.helm*type.std + (1|item) + (1|item:frame), data=df)`

[6] Model syntax: `glmer (sameDiff ~ DTWdistance*type*language + (1|stimuli) +(1|participant), data = df, family = ''binomial'', control = glmerControl(optimizer = ''bobyqa'', optCtrl = list (maxfun = 100000)))`

`language` (i.e. English / French). The random effects include by-stimuli random intercept and by-participant random intercept, to account for the variability from particular stimuli and participant.

Conceptually, the measure of the model quality (i.e. to capture the predictability using models' DTW distances for human responses) should only reflect the contribution of *fixed* effects. In other words, we are only interested in *how well type, DTW distance, and language predict human responses*, after by-stimuli and by-participant variability.

We use two measures, the marginal R^2 and the Area Under the Curve (AUC). The marginal R^2 shows how much variance in the human response data is explained by the fixed effects terms in the MEL model. The AUC is a measure of the ability of a classification model to distinguish between classes (here, the ability to distinguish humans' *same* or *different* responses), used as a summary of the ROC curve. A higher AUC means that the model is better at distinguishing the classes. As we are only interested in the *fixed* effects, the AUC is also calculated using the predictions using *fixed* effects only.

The results (Table 4.6) show that the two measures give the same qualitative pattern (i.e. bigram $>$ trigram $>$ flat), although the differences across the three models are rather small. Despite the small differences, the bigram LM is found by both measures to be slightly better than the other two in predicting human responses.

	LM		
Measure	flat	bigram	trigram
marginal R^2	0.085	0.091	0.086
AUC	0.639	0.647	0.635

Table 4.6: Marginal R^2 and AUC score for three MEL models, corresponding to the three *Compensating* ASR models.

4.4.2 Experiment 2: Spliced-out word with no following context

Experiment 2 simulates Darcy et al. (2009)'s control experiment to test the (un)ambiguity of the *Viable* (assimilated) stimuli in the absence of the following context, using spliced-out target words. Recall that the purpose of this control experiment in the original paper was to verify that *Viable* stimuli were heard to have a different phone instead of the underlying phone without the following context. This shows that it is not the acoustics of the target words that allow listeners to perceive them as their underlying form, when presented in a viable context for assimilation. However, given the nature of the control experiment being a *deterministic* transcription task, it is possible that the *Viable* stimuli are somewhat different from the *Unviable* stimuli, but such differences are not sufficient to lead to a different transcription in the absence of the following context.

In order to get a more fine-grained evaluation of the acoustics of the target words, we use the DTW distances produced from our models and evalute them in the same way as in Experiment 1 (division into bad/MP-distinguishing, compensating). This tests whether the models are sensitive to acoustic differences in the target words. If no *compensating* models are found, it confirms that the acoustics in the target words alone are not sufficient for compensation to happen. However, if there are indeed *compensating* models, then it suggests that the acoustics in the target words are informative enough to 'compensate' (despite the absence of the following context), but humans either do not perceive these fine details, or do not use them, and only compensate when licensed by the following context.

Note though that the current experiment differs from Darcy et al.'s control experiment in the task performed by the listener. In their experiment, listeners simply transcribed the word-final phone they heard, and did not compare to the 'target word' produced in isolation. We do present the ASR models with the words in iso-

lation, in order to calculate DTW distances with the cut-out words (calculating a distance requires a pair of words). Given this difference between the experiment designs, we treat the human responses only as reasonable reference.

***MP-distinguishing* models** Four models are found to be *MP-distinguishing* in both languages, summarised in Table 4.7: triphone AM with bigram and trigram LM; triphone speaker-adapted AM with flat and trigram LM. The four models are plotted in Figure 4.7 (a) with comparison with humans doing the control experiment. The columns represent the languages and the rows represent models. The y axis represents models' DTW distance (bars) and the percentage detection rate for humans of perceiving different consonant from the unchanged target words (black dots). Higher values in both cases indicate larger difference.

The comparison between models and humans show the similar pattern that *Viable* and *Unviable* are both found to be more different than the unchanged target words, although models find *No Change* cases to be more 'different' compared with human judgements. The latter might be due to the difference in that humans did the deterministic transcription tasks while models used continuous DTW measures, which can add up and show larger dissimilarity. In summary, the control experiment again shows that in order to distinguish minimal pairs, one needs contextual phonetic knowledge.

***Compensating* models** As in Experiment 1, we calculated the Compensation Index for both models (Equation 4.5) and humans (Equation 4.1). Figure 4.7 (b) shows the Compensation Index for **all** *MP-distinguishing* models (colored) reported in Table 4.7 and human perception (black), with rows showing different LMs and colors representing different AMs. Note again that in this experiment, humans and models did not do the same task, although the two are similar. Thus, the human reference

English			French		
Model (AM-LM)	p_{place}	p_{voicing}	Model (AM-LM)	p_{place}	p_{voicing}
monophone-flat	0.141	0.001	monophone-flat	0.150	0.001
monophone-unigram	0.720	***	monophone-unigram	0.019	***
monophone-bigram	0.392	0.006	**monophone-bigram**	**0.007**	***
monophone-trigram	0.164	0.005	monophone-trigram	0.054	0.547
triphone-flat	***	0.367	triphone-flat	0.013	***
triphone-unigram	***	**0.010**	triphone-unigram	0.065	***
triphone-bigram	**0.008**	**0.004**	**triphone-bigram**	***	***
triphone-trigram	**0.003**	**0.001**	**triphone-trigram**	***	***
triphoneSA-flat	***	***	**triphoneSA-flat**	***	***
triphoneSA-unigram	0.025	***	triphoneSA-unigram	0.250	***
triphoneSA-bigram	***	***	triphoneSA-bigram	0.177	***
triphoneSA-trigram	***	***	**triphoneSA-trigram**	**0.005**	***

Table 4.7: p-values for minimal pair contrasts in place and voicing (spliced-out stimuli), for each ASR model, for English and French data. Bolded models are *MP-distinguishing* models and underlined models are *MP-distinguishing* models found in both languages. Uncorrected p-values (from `lsmeans()` post-hoc comparisons) are reported by assimilation type. *** indicates $p < 0.001$.

should be only treated as what humans *might* respond if they were to participate in the same experiment, rather than their actual response.

Out of the four *MP-distinguishing* models, two show language-specific compensation despite the absence of the following context: †he models with triphone AM and bigram or trigram LM. A further check on the effects of LM using an LMEM model [7] , as in Expertiment 1, shows that LMs again do not make a significant difference for predicting the DTW distance, by a likelihood ratio test comparing LMEM models with and without the interactions of LM term with all variables (χ^2= 7.535, df = 12, p = 0.820). The pattern differs between models and humans in two

[7]Model syntax: `lmer(distance ~ lm.std*lang.std*condition.helm*type.std+ (1|item) + (1|item:frame), data = aefr.tribitri)`

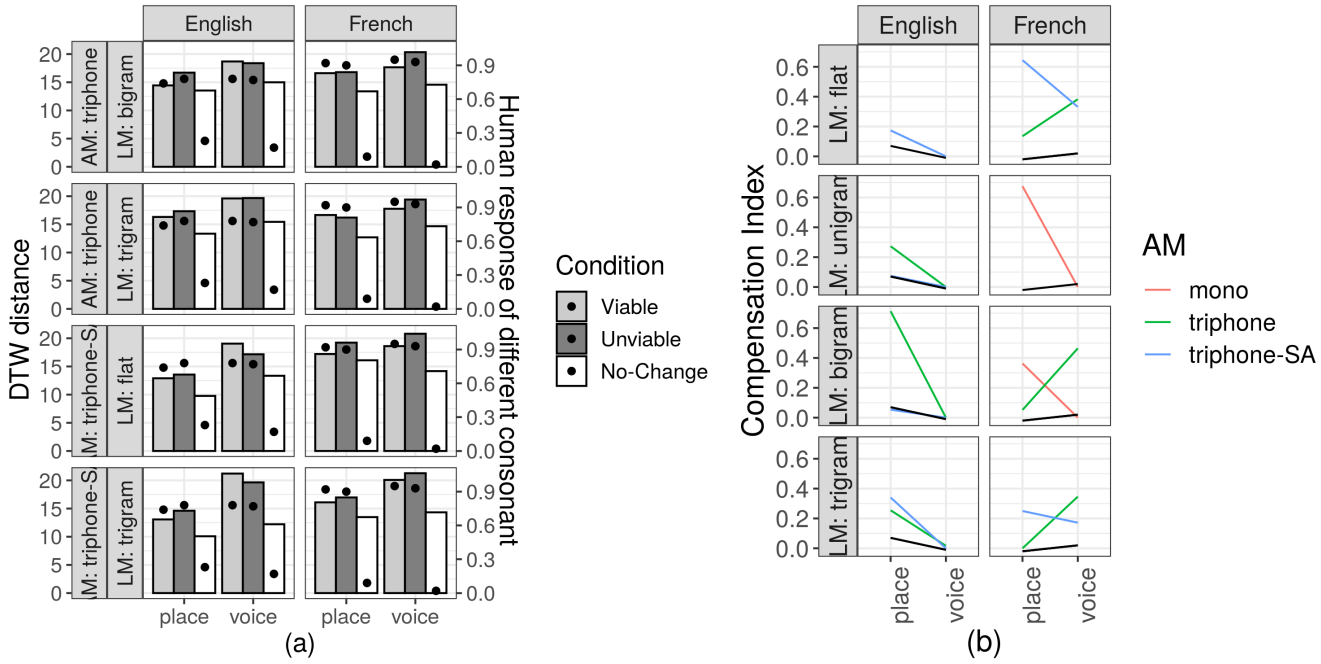


Figure 4.7: Results for spliced-out words. (a) model DTW distance (bars) and human detection rate of perceiving a different consonant from the unchanged target (points) for the three conditions, for *MP-distinguishing* models successful in both languages (varying in AM/LM). (b) Compensation Indices of **all** *MP-distinguishing* models, decoded without context across ASR models differning in AM and LM (colored lines), and for human data (black).

ways. The ASR models show overall higher Compensation Indices, while humans are essentially close to 0 with no compensation, meaning that the models can still compensate for assimilation *in the absence of the following context*. In addition, the ASR models show *language-specific compensation*, which humans do not do. The finding suggests that there is indeed sufficient information in the acoustics in order for a model to compensate for assimilation even without the following context. However, the amount of information is not enough for humans to guess that the presented word is assimilated, and hence compensate for the assimilation.

4.4.3 Experiment 3: non-native speech

Previous results show that language-specific compensation is possible given *legal* phone sequences with native assimilation (Experiment 1), even without the following context (Experiment 2), but not possible for non-native assimilation (e.g. the English voicing assimilation stimuli). Experiment 3 further examines how *illegal* phone sequences (including syllable structure) affect compensation, in terms of both native and non-native assimilation. This advances our understanding of how much listeners are affected by *illegal* phone sequences during non-native speech perception.

We test models on the stimuli produced by non-native speakers—that is, models trained on English listen to French stimuli and vice versa—using the same procedure as Experiments 1–2. No type of ASR model (AM/LM pair) is successful in distinguishing minimal pairs for both types of assimilation. As summarized in Table 4.8, certain models can differentiate non-native minimal pairs contrasting in voicing, but not pairs contrasting in place, regardless of the language. According to our criteria, no model is *MP-distinguishing*, so we do not explore further.

4.4.4 Summary

To summarise, Experiment 1 shows that when presented with assimilation patterns pronounced in their 'native language', some ASR models compensate for assimilation, in the same language-specific way as humans. This is mostly due to these models having context-dependent AMs, which represents contextual phonetic knowledge. In Experiment 2, we found that in the absence of following context, ASR models still manage to restore the original sound. This suggests that there are fine phonetic details in the assimilated words, which are sufficient for ASR models to 'compensate' for assimilation without the following context, but not for humans to do so. In Experiment 3, we tested whether the ASR models could compensate for

English model hear French stimuli			French model hear English stimuli		
Model (AM-LM)	p_{place}	p_{voicing}	Model (AM-LM)	p_{place}	p_{voicing}
monophone-flat	0.022	0.061	monophone-flat	0.955	0.066
monophone-unigram	0.078	0.516	monophone-unigram	0.978	***
monophone-bigram	0.066	0.352	monophone-bigram	0.493	***
monophone-trigram	0.310	0.027	monophone-trigram	0.303	***
triphone-flat	0.910	0.995	triphone-flat	0.678	0.973
triphone-unigram	0.855	0.957	triphone-unigram	0.213	0.861
triphone-bigram	0.962	0.693	triphone-bigram	0.822	0.072
triphone-trigram	0.503	0.552	triphone-trigram	0.181	0.140
triphoneSA-flat	0.648	0.842	triphoneSA-flat	0.558	0.015
triphoneSA-unigram	0.156	***	triphoneSA-unigram	0.407	0.046
triphoneSA-bigram	0.250	***	triphoneSA-bigram	0.037	**0.005**
triphoneSA-trigram	0.484	**0.005**	triphoneSA-trigram	0.022	**0.005**

Table 4.8: p-values for minimal pair contrasts in place and voicing (non-native stimuli), for each ASR model, for English and French data. No model is found to be *MP-distinguishing*. Uncorrected p-values (from `lsmeans()` post-hoc comparisons) are reported by assimilation type, with $p < \alpha = 0.01$ in bold. *** indicates $p < 0.001$.

assimilation in non-native speech. No model met the minimum standard of human-like perception: distinguishing both place and voicing minimal pairs.

4.5 Discussion

In this study, we use ASR models as 'ideal listeners' to investigate what kinds of linguistic knowledge are crucial for a learner to be able to compensate for phonological assimilation in a language-specific manner. The results show that some ASR models indeed compensate for assimilation in human-like ways, despite being trained on a different, more generic task (i.e. phone recognition). In particular, contextual phonetic knowledge (i.e. a triphone acoustic model) already encodes language-specific knowledge that enables a learner to restore the underlying sound, and that phono-

logical knowledge about the phone sequences (i.e. a complex language model) is not mandatory for realizing the compensation effect. Moreover, the 'ideal listener' is able to use the acoustic details in the assimilated word to 'compensate' and restore the original sound, even in the absence of the following context (spliced-out words), although this is not what humans do. This section discusses the roles of phonetic knowledge, phonological knowledge, and the following context, and how our findings address different theories in relation to phonological assimilation.

4.5.1 Contextual phonetic knowledge is crucial and language-specific

A major finding of this paper is that contextual phonetic knowledge, represented by the Acoustic Model, is important for a learner to show two basic properties of human speech perception: distinguishing minimal pairs and compensation for assimilation. This even holds when combined with a null Language Model (flat LM), suggesting that such contextual phonetic knowledge is language-specific and is sufficient for a learner to compensate for assimilation.

Minimal pair distinguishability The failure of distinguishing minimal pairs using raw acoustics (the baseline) indicates that pure acoustics are not enough to disambiguate a pair of words that differ only in one segment. The acoustic differences of the minimal pairs in our experiments come from two sources: the talker difference (recall that the sentence stimuli and the target words were produced by two different talkers) and the segmental difference of the contrastive segment. Without an AM and a LM, no linguistic knowledge is learnt, and thus a learner is not able to tease apart the linguistic difference from the talker difference. This suggests that some degree of phonetic and phonological knowledge is mandatory for realizing the minimal pair contrast.

For models that are *MP-distinguishing*, the results from both Experiment 1 and Experiment 2 suggest that phonetic knowledge facilitates distinguishing minimal pairs, while the degree of context-sensitivity depends on the language. For French models, contextual phonetic knowledge is not mandatory, as models with context-independent monophone AMs are able to distinguish minimal pairs in both Experiment 1 and 2. However, English models always require context-sensitive AMs (i.e. triphone AMs). The control experiment performed by humans in Darcy et al. (2009) also show less distinguishablitiy across the board, which suggests that the speech stimuli might be less clear and that contextual information is needed. Another possible explanation is that English has more contextual allophones than French, so the model needs additional contextual knowledge to determine the underlying phone. Thus, while contextual phonetic knowledge is necessary in some cases, we speculate that this may vary from contrast to contrast and language to language, depending on the degree of contextual phonetic variation that needs to be discriminated from.

Language-specific compensation In terms of compensation for assimilation, our results show that only models with contextual phonetic knowledge are able to show the same language-specific compensation pattern as humans (i.e. compensating models). Models with context-independent (monophone) AMs, even when combined with a trigram LM, never show the expected compensation pattern. This finding suggests that the context-specific phonetics of the phone in a phonological alternation needs to be explicitly learned, in addition to its context. Moreover, a learner can learn and solely rely on the contextual phonetic knowledge that is language-specific. As phonological assimilation only concerns two adjacent phones, it makes sense that contextual phonetic knowledge without additional phonological knowledge is enough for realizing compensation.

4.5.2 Phonological knowledge is less important

The role of the LM, on the other hand, is secondary to the AM. A learner does not have to have any phonological knowledge for compensation, in the sense that the flat LM in Experiment 1 showed qualitatively-identical human-like compensation patterns to the bigram and trigram LMs. Nevertheless, more complex LMs perform slightly better than the flat LM in terms of predicting human responses (Table 4.6). In addition, if a learner needs to better predict the following context, then more complex LMs are needed. The results from Experiment 2 (Figure 4.7) show that only when combined with bigram or trigram LMs can the ASR model restore the underlying phone. In this case, a flat LM does not suffice.

4.5.3 Fine-grained phonetic detail and the effect of the following context

While the experiments performed by the ASR models and humans are not exactly the same, Experiment 2 nevertheless shows one disparity between the two types of listeners: while humans do not judge the assimilated phones to be the same as the unchanged phones in the absence of the following context, ASR models can still restore the underlying phone.

One possible explanation is that ASR models manage to make use of acoustic information not used (as much) by humans. Other studies have found that humans can in some cases make use of subtle acoustic cues to assimilation. In an eye-tracking study, Gow and McMurray (2007) found that for English place assimilation, listeners are able to predict the following phone (the one that triggers the assimilation) before they hear it. It is possible that models are better at learning such cues than humans, and hence they restore the underlying phone even without the following context. A related but slightly different explanation is that humans

also perceive such fine phonetic details, but use them in a different way: due to the lack of the following context, they regard them as cues to a different phone instead of an assimilated variant of the target phone.

An alternative explanation attributes the 'compensation' without following context to the represenation of the phones encoded by the triphone AM: each phone has different representations in different contexts, some of which are 'used' when the AM successfully categorizes the phone—regardless of whether that context is actually present in the signal. In other words, the AM has already hallucinated the context during categorization, so there is no need to actually observe the following phone to 'compensate' for it.

The disparity between the model and human performance suggests a few possible differences between humans and the ASR models. Humans may not be as good as the models at making use of fine-grained phonetic details, or they may be equally good, but interpret the lack of following context as meaningful, while our (phone-level) ASR models without higher-level linguistic knowledge do not. A third possibility is that humans do not learn distinct representations for a phone corresponding to each of its distributional contexts.

4.5.4 Problems with non-native speech perception

As introduced earlier, non-native perception experiments pose a problem of confounding various non-native factors, such as non-native phone categories, non-native phonotactics, non-native syllable structure, etc. While it can be hard to disentangle those factors for a human listener, we are able to represent them separately as AM and LM in ASR models, and can hence evaluate the influence of specific component in non-native speech perception.

In Experiment 3, through our process of model goodness selection, we found that none of the models can distinguish minimal pairs in both types of contrasts. Our models show an interesting pattern of never being able to discriminate minimal pairs in place contrasts (Table 4.8). While the two languages share the same phone categories contrasting in place, they nevertheless fail to differentiate the minimal pairs in the other language.

One possible explanation might be that place contrasts rely heavily on formant transistions of the neighbouring vowels, however, as the two languages have considerably different vowel inventories, it may not match well to trigram AMs which are vowel-specific. Our finding further suggests that when a listener perceives non-native phonological alternations, they may face challenges as early as the phone recognition stage, even for phone categories overlapping with those in their native language.

4.5.5 Other hypotheses

The current study also sheds light on other frameworks proposed to account for language-specific phonological compensation, namely, Lexical Compensation and Phonetic Compensation.

The Lexical Compensation theory is a top-down approach which attributes compensation to the listener's lexical knowledge, but does not use much phonetic details. It treats all variations as random noise, which can be recovered using lexical or higher-order context (e.g. Marslen-Wilson and Welsh, 1978; Samuel, 2001). This hypothesis predicts that in the absence of a lexicon, compensation for phonological assimilation cannot happen, which our results contradict. Instead, our results show that a learner with only phone-level knowledge, without higher level knowl-

edge such as lexicon or word boundaries, can still show the same language-specific compensation pattern as humans.

Second, the bottom-up Phonetic Compensation approach accounts for compensation with a low-level phonetic mechanism. This approach states that sounds that simultaneously encode two places of articulation (using earlier examples, the *[n/m]* in *own plan*) are parsed onto adjacent segmental positions, when the following context explains one of the places of articulation (Gow, 2003; Gow and McMurray, 2007). In this case, the recovery of /n/ from [m] can be attributed to the attraction of the labial aspects of the acoustics to the following labial segment. However, the phonetic information here is proposed to be language-independent, which does not account for the language-specific compensation observed. Our results show that phonetic details are indeed important, but instead of being language-independent, we found that if a learner learns contextual phonetic knowledge, then the language-specifity can be encoded and hence allows for compensation for assimilation.

4.6 Conclusion

In this study, we mainly examined the roles of phonetic and phonological knowldege in speech perception, focusing on language-specific phonological assimilation. We use standard automatic speech recognition systems trained on English and French to represent different 'ideal listeners'. The models are implemented with different degrees of phonetic and phonological knowlege. By comparing different ASR models' performance with human performance on the same experimental data, we found that the successful human-like models employ context-sensitive phonetic knowledge and phonological knowledge, but do not require higher-level knowledge of a lexicon or word boundaries.

Chapter 5

Conclusion

The central question of this dissertation is: How are human beings able to make sense of speech and interpret it as meaningful units, despite extensive variation in the speech signal? This dissertation addresses this question by examining different levels of human speech processing, from low-level phonetics to higher-level abstract patterning: listeners' specific use of acoustic dimensions in various linguistic contexts, the perceptual representation integrating all acoustic dimensions for a phonological contrast, and the linguistic knowledge used for processing phonological changes. In order to investigate these aspects of perception, I combine perceptual experiments and statistical modeling to test specific hypotheses and use computational models to examine abstract perceptual representations and processing mechanisms.

More broadly, this dissertation explores the methodology of bridging computational modelling with linguistics. Unlike the human brain which can only be indirectly probed through responses, a computational model has its components clearly defined and implemented and can be trained to represent a 'listener'. When the model listener's 'perceived sound' (i.e. the prediction) corresponds to humans', the system can be seen as a possible explanation for human speech perception. More-

over, one can use the model to test hypotheses impossible to conduct on human participants: for example, testing the importance of the lexicon by not feeding lexical information to the model, while for humans the acquired knowledge cannot be undone.

5.1 Summary

5.1.1 Individual and dialectal differences in perceiving Wu dialects

This project investigates how multiple acoustic dimensions (or cues) contribute to multi-dimensional phonological contrasts at both the group level and the individual level, and how dialectal experience shapes listeners' perceptual strategies. We examined the tonal register contrast in two Chinese Wu dialects (Shanghai and Jiashan) focusing on three cues: pitch height, voice quality, and pitch contour. Participants heard ambiguous words differing in tonal register and identified which words they heard. We built mixed effects models to capture the listeners' cue weights, both at the group level and the individual level. Individual variability in cue weights was examined by analyzing the random effect estimates in the model, making use of the component of mixed-effects models which are often treated as a 'by-product' in analyses of linguistic data.

The findings reveal that listeners differ mainly in their overall cue acuity (e.g. there are listeners with flatter and steeper boundaries between sounds – across all cues). Moreover, for certain contrasts signaled without a dominant cue, individuals further differ in their choice of the primary cue. Finally, listeners' use of cues is affected by their dialect background. For a cue less important in their native dialect, listeners do not make better use of it even when the cue becomes more salient in the same contrast.

5.1.2 Modelling perceptual tonal space in Mandarin Chinese continuous speech

The second project investigates the perceptual tonal representation of Mandarin Chinese running speech, and how various acoustic cues map onto this representation. While pitch is conventionally used for tone notations, various other acoustic cues are also involved for signalling tonal contrast. We built a computational model trained on Mandarin continuous speech to represent tone identification by an 'ideal listener'. We employed Long Short-Term Memory models which handles variable length input, so syllables with any duration can be used as input. Moreover, we used high-dimensional input extracted from raw speech, leaving the model to decide on the information learnt in the signal. We forced the model to learn a low-dimensional representation, which can be seen as the model's perceptual representation of tones. By examining this representation, one can find out how the tonal space is constructed and provide a possible explanation for humans' perceptual tone space.

The results show that the models learn a two-dimensional tone representation compressing the high-dimensional information in the input, without sacrificing accuracy. A closer examination of the perceptual tonal representation reveals that pitch is represented as average pitch height and pitch contour in the two dimensions. Furthermore, we failed to find the onset of the tone being independently correlated with one dimension and offset correlated with the other dimension, which calls into question the conventional tonal notation using onset and offset as the tone's pitch targets. The method used also opens up a new approach for investigating tonal representation across languages in the future.

5.1.3 Testing perceptual compensation for phonological assimilation

This chapter aims to understand the role of phonological knowledge in speech perception through computational modeling. While many behavioral studies have investigated how humans process spoken language at different levels, important questions remain about how these levels are integrated and interact with each other. In this study, we examine the minimal knowledge required for restoring the original sound changed by phonological assimilation, and whether or not phonological rules are needed. We approach this problem through the cases of French voicing assimilation and English place assimilation.

We trained standard (HMM-GMM) automatic speech recognition models to represent English and French listeners, which do the same task as humans – receiving speech as input and converting them into a string of phones. The system contains two components, the acoustic model and the language model, which represent the system's knowledge of phonetics and phone distribution/ phonotactics respectively. By varying the complexity of these models, we got a set of 'model listeners' with different levels of phonetic knowledge and phonotactic knowledge. We tested the model with the same experimental stimuli as humans heard in a previous study (Darcy et al., 2009) to examine listeners' ability to detect native and non-native assimilation. We further compared the results between the set of models and the human responses to determine the most human-like models.

The results reveal that some types of models show language-specific assimilation patterns comparable to those shown by human listeners. Models that best predict the human pattern use contextually sensitive acoustic models and language models, which capture allophony and phonotactics, but do not make use of higher-level knowledge of a lexicon or word boundaries. The patterns are explained by a com-

bination of contextual acoustic modelling and phonotactic patterns, but nowhere in the system is there an application of explicit phonological rules.

5.2 General discussion

5.2.1 Implications to studying tonal languages

The first two projects in this dissertation use different approaches to study a similar question: how listeners use multiple cues to distinguish tonal contrasts in variable speech. We found converging evidence that computational models 'perceive' speech sounds similar to humans in multiple ways. Those models not only largely agree with humans in the relative contributions of cues, but also the perceptual representation of the phonological contrast.

The present studies contribute to our understanding of computational models' learnt tonal representation. Previous studies investigating models' learnt knowledge of speech sounds mostly focus on segmental information (e.g. Bai et al., 2018; Weber et al., 2016), and little is known in terms of whether or not suprasegmental information can also be learnt by computational models in a similar manner as humans. The positive evidence found in our studies suggests that a similar method can be used to study tone representations in various other tone languages, and may provide insights for those with relatively more complex tone inventories.

5.2.2 Compuational models as ideal listeners

Chapter 3 and 4 are two ways of using computational models to represent ideal listeners. The two chapters focus on speech variability of different kinds: the former concerns more general variation while the latter deals with the more restricted phonological assimilation. In chapter 3, we directly probe the model's perceptual

representation, which provides a possible explanation of humans' perceptual representations. In chapter 4, instead of investigating one model, we parameterize the different types of linguistic knowledges to get a set of models. Through examining the model performances and human performances on the same task, the human-like models demonstrate the mechanism and knowledge required for language-specific phonological compensation. In both studies, we find that models can perform similar to humans, suggesting that computational models trained on specific tasks nevertheless learn linguistic knowledge beyond the scope of the trained tasks. Specifically, the Mandarin tone indentification models learn the particular representation and geometry of the cues, and the English and French phone recognition models display language-specific phonological compensation patterns.

The positive evidence suggests that this method is not only helpful for understanding human speech perception, but also informative in terms of which specific aspects to improve for ASR models. The findings in the two studies so far do not show models fail to learn certain specific linguistic knowledge. However, if there were some consistent failures shown by the models, then one can implement explicit instructions for the models to learn the missing information, and hence improve the model performance.

5.3 Future directions

This dissertation shows that computational models can be used to explore how humans perceptually represent phonological contrasts encoded by multiple cues, and how various kinds of linguistic knowledge are combined and interact in phonological processes. The two ways of computational modelling point to different future directions.

First, the methodology used in chapter 3 can be applied to various other tone languages, which provides a more unified and systematic evaluation of the tone systems across languages. While tone languages typically involve multi-dimensional tonal contrasts, with specific cues differing from language to language, this method allows for a parallel comparison across languages. Various tone languages can be compared under the same criteria, such as complexity in terms of number of dimensions needed for tonal representation, and similarity between languages in terms of the geometry of the tones. The examination of the low-dimensional tonal representation further sheds light on how cues may be integrated – whether there are shared patterns across languages or whether it is language-specific.

Second, one may use more advanced models to investigate if they learn more or less linguistic knowledge. In chapter 4, we use the traditional HMM-GMM models, where different kids of linguistic knowledge are implemented in different components. The more recent end-to-end deep learning models, on the other hand, are less modular. While they achieve high performance, it is less understood whether or not these models learn linguistic knowledge and how. Extensive research have been done to analyse the syntactic and semantic knowledge learnt in deep learning models [citations], yet little is known in the field of phonology. Following similar logic to that of chapter 4, the analysis of the disparaty between models and humans furthers reveal how we may improve end-to-end models.

5.4 Conclusion

This dissertation investigates how humans perceive speech variability using a combination of behavioral and computational methods. We show that in order to distinguish a multidimensional tonal register contrast, listeners use different strategies for different dialect varieties. The perceptual strategies also differ across listeners, but in

a structured manner (Chapter 2). A computational approach to a similar problem of the Mandarin tonal contrast reveals that computational learners and human listeners arrive at similar solutions in the perceptual representation of perceiving tones (Chapter 3). Finally, we use modularized computational models to gain insights on the importance of different types of linguistic knowledge on language-specific phonological assimilation (Chapter 4).

Bibliography

Abramson, A. S. and Lisker, L. (1985). Relative power of cues: F0 shift versus voice timing. *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, pages 25–33.

Allen, J. S., Miller, J. L., and DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, 113(1):544–552.

Anderson, J. R. (2013). *The adaptive character of thought*. Psychology Press.

Ashby, M. and Przedlacka, J. (2014). Measuring incompleteness: Acoustic correlates of glottal articulations. *Journal of the International Phonetic Association*, 44(3):283–296.

Bai, L., Weber, P., Jancovic, P., and Russell, M. J. (2018). Exploring how phone classification neural networks learn phonetic information by visualising and interpreting bottleneck features. In *INTERSPEECH 2018*, pages 1472–1476.

Bang, H.-Y. and Clayards, M. (2016). Structured variation across sound contrasts, talkers, and speech styles. *Poster presented at LabPhon15: Speech Dynamics and Phonological Representation. Ithaca, NY*.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Beddor, P. S. and Krakow, R. A. (1999). Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation. *The Journal of the Acoustical Society of America*, 106(5):2868–2887.

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., and Brasher, A. (2013). The time course of perception of coarticulation. *The Journal of the Acoustical Society of America*, 133(4):2350–2366.

Belinkov, Y., Ali, A., and Glass, J. (2019). Analyzing phonetic and graphemic representations in end-to-end automatic speech recognition. *arXiv preprint arXiv:1907.04224*.

Belinkov, Y. and Glass, J. (2017). Analyzing hidden representations in end-to-end automatic speech recognition systems. *arXiv preprint arXiv:1709.04482*.

Blicher, D. L., Diehl, R. L., and Cohen, L. B. (1990). Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: Evidence of auditory enhancement. *Journal of Phonetics*, 18(1):37–49.

Boersma, P. and Weenink, D. (2019). Praat: doing phonetics by computer [computer program]. Version 6.0.50, retrieved 31 March 2019 from http://www.praat.org/.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Brunelle, M. (2003). Tonal coarticulation in Northern Vietnamese. In *Proceedings of the 15th International Congress on Phonetic Sciences*, pages 2673–2676.

Brunelle, M. (2009). Tone perception in Northern and Southern Vietnamese. *Journal of Phonetics*, 37(1):79–96.

Brunelle, M. and Finkeldey, J. (2011). Tone perception in sgaw karen. In *Proceedings of the 17th International Congress on Phonetic Sciences*, pages 372–375.

Bukmaier, V., Harrington, J., and Kleber, F. (2014). An analysis of post-vocalic /s-ʃ/ neutralization in Augsburg German: Evidence for a gradient sound change. *Frontiers in Psychology*, 5:828.

Cao, J. and Maddieson, I. (1992). An exploration of phonation types in Wu dialects of Chinese. *Journal of Phonetics*, 20(1):77–92.

Cao, R. (2012). *Perception of Mandarin Chinese Tone 2/Tone 3 and the role of creaky voice*. University of Florida.

Chandrasekaran, B., Gandour, J. T., and Krishnan, A. (2007). Neuroplasticity in the processing of pitch dimensions: A multidimensional scaling analysis of the mismatch negativity. *Restorative Neurology and Neuroscience*, 25(3-4):195–210.

Chandrasekaran, B., Sampath, P. D., and Wong, P. C. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America*, 128(1):456–465.

Chang, C.-Y. (2010). *Dialect differences in the production and perception of Mandarin Chinese tones*. PhD thesis, The Ohio State University.

Chang, Y.-c. and Hsieh, F.-f. (2012). Tonal coarticulation in Malaysian Hokkien: A typological anomaly? *The Linguistic Review*, 29(1):37–73.

Chao, Y. R. (1928). Xiandai Wuyu Yanju [Studies in the Modern Wu dialects]. *Peking: Tsing Hua College Research Institute, Monograph*, 4.

Chao, Y.-R. (1930). A system of tone letters. *Le Maître Phonétique*, 45:24–47.

Chen, S., Wiltshire, C., and Li, B. (2018). An updated typology of tonal coarticulation properties. *Taiwan Journal of Linguistics*, 16(2):79–114.

Chen, Y. and Gussenhoven, C. (2008). Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics*, 36(4):724–746.

Chen, Y. and Gussenhoven, C. (2015). Shanghai Chinese. *Journal of the International Phonetic Association*, 45(3):321–337.

Chen, Y. and Xu, Y. (2020). Intermediate features are not useful for tone perception. In *Proceedings of the 10th International Conference on Speech Prosody 2020*, pages 513–517.

Chen, Z. (2010). 吴语清音浊流的声学特征及鉴定标志—以上海话为例. An acoustic study of voiceless onset followed by breathiness of Wu (吴) dialects: Based on the Shanghai (上海) dialect. *Studies in Language and Linguistics*, 30(3):20–34.

Chien, Y.-F., Sereno, J. A., and Zhang, J. (2016). Priming the representation of Mandarin Tone 3 sandhi words. *Language, Cognition and Neuroscience*, 31(2):179–189.

Chodroff, E., Godfrey, J., Khudanpur, S., and Wilson, C. (2015). Structured variability in acoustic realization: a corpus study of voice onset time in American English stops. In *Proceedings of International Congress on Phonetic Sciences*.

Chodroff, E. and Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, 61:30–47.

Chuang, C.-K. and Hiki, S. (1972). Acoustical features and perceptual cues of the four tones of standard colloquial Chinese. *The Journal of the Acoustical Society of America*, 52(1A):146–146.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., and Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, 78(4):685–709.

Clayards, M. (2018). Differences in cue weights for speech perception are correlated for individuals within and across contrasts. *The Journal of the Acoustical Society of America*, 144(3):EL172–EL177.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., and Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3):804–809.

Coenen, E., Zwitserlood, P., and Bölte, J. (2001). Variation and assimilation in german: Consequences for lexical access and representation. *Language and Cognitive Processes*, 16(5-6):535–564.

Coetzee, A. W., Beddor, P. S., Shedden, K., Styler, W., and Wissing, D. (2018). Plosive voicing in Afrikaans: Differential cue weighting and tonogenesis. *Journal of Phonetics*, 66:185–216.

Cole, J., Linebaugh, G., Munson, C., and McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, 38(2):167–184.

Coster, D. and Kratochvil, P. (1984). Tone and stress discrimination in normal Beijing dialect speech. *New Papers on Chinese Language Use*, pages 119–132.

Darcy, I., Ramus, F., Christophe, A., Kinzler, K., and Dupoux, E. (2009). Phonological knowledge in compensation for native and non-native assimilation. *Variation and Gradience in Phonetics and Phonology*, 14:265–309.

Davis, M. H. and Scharenborg, O. (2016). Speech perception by humans and machines. In *Speech Perception and Spoken Word Recognition*, pages 191–214. Psychology Press.

De Jong, K. (1998). Stress-related variation in the articulation of coda alveolar stops: Flapping revisited. *Journal of Phonetics*, 26(3):283–310.

Deng, D., Shi, F., and Lu, S. (2006). The contrast on tone between Putonghua and Taiwan Mandarin. *Acta Acustica*, 6.

Dilley, L. C. and Pitt, M. A. (2007). A study of regressive place assimilation in spontaneous speech and its implications for spoken word recognition. *The Journal of the Acoustical Society of America*, 122(4):2340–2353.

Duanmu, S. (2007). *The phonology of Standard Chinese*. Oxford University Press.

Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., and Dupoux, E. (2017). The Zero Resource Speech Challenge 2017. In *2017 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 323–330. IEEE.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59.

Erickson, D., Iwata, R., Endo, M., and Fujino, A. (2004). Effect of tone height on jaw and tongue articulation in Mandarin Chinese. In *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*.

Fant, G. (1970). *Acoustic theory of speech production*, volume 2. Walter de Gruyter.

Farnetani, E. and Recasens, D. (1997). Coarticulation and connected speech processes. *The Handbook of Phonetic Sciences*, 371:404.

Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4):752.

Flemming, E. and Cho, H. (2017). The phonetic specification of contour tones: Evidence from the Mandarin rising tone. *Phonology*, 34(1):1.

Francis, A. L., Kaganovich, N., and Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in english. *The Journal of the Acoustical Society of America*, 124(2):1234–1251.

Fu, Q.-J. and Zeng, F.-G. (2000). Identification of temporal envelope cues in Chinese tone recognition. *Asia Pacific Journal of Speech, Language and Hearing*, 5(1):45–57.

Fu, Q.-J., Zeng, F.-G., Shannon, R. V., and Soli, S. D. (1998). Importance of tonal envelope cues in Chinese speech recognition. *The Journal of the Acoustical Society of America*, 104(1):505–510.

Fujisaka, H. and Kunisaki, O. (1976). Analysis, recognition and perception of voiceless fricative consonants in Japanese. In *ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 158–161. IEEE.

Gandour, J. (1984). Tone dissimilarity judgements by Chinese listeners. *Journal of Chinese Linguistics*, pages 235–261.

Gao, J. (2016). Sociolinguistic motivations in sound change: On-going loss of low tone breathy voice in Shanghai Chinese. *Papers in Historical Phonology*, 1:166–186.

Gao, J., Hallé, P., and Draxler, C. (2020). Breathy voice and low-register: A case of trading relation in Shanghai Chinese tone perception? *Language and Speech*, 63(3):582–607.

Gao, J., Hallé, P., Honda, K., Maeda, S., and Toda, M. (2011). Shanghai slack voice: acoustic and EPGG data. In *Proceedings of the 17th International Congress on Phonetic Sciences*, pages 719–722.

Garellek, M. and Keating, P. (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association*, pages 185–205.

Garellek, M., Keating, P., Esposito, C. M., and Kreiman, J. (2013). Voice quality and tone identification in White Hmong. *The Journal of the Acoustical Society of America*, 133(2):1078–1089.

Gaskell, M. G. (2003). Modelling regressive and progressive effects of assimilation in speech perception. *Journal of Phonetics*, 31(3-4):447–463.

Gaskell, M. G., Hare, M., and Marslen-Wilson, W. D. (1995). A connectionist model of phonological representation in speech perception. *Cognitive Science*, 19(4):407–439.

Gaskell, M. G. and Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 22(1):144.

Gaskell, M. G. and Marslen-Wilson, W. D. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 24(2):380.

Gauthier, B., Shi, R., and Xu, Y. (2007). Learning phonetic categories by tracking movements. *Cognition*, 103(1):80–106.

Gordon, P. C., Eberhardt, J. L., and Rueckl, J. G. (1993). Attentional modulation of the phonetic significance of acoustic cues. *Cognitive Psychology*, 25(1):1–42.

Gow, D. W. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics*, 65(4):575–590.

Gow, D. W. and McMurray, B. (2007). Word recognition and phonology: The case of English coronal place assimilation. *Papers in Laboratory Phonology*, 9(173-200).

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649. Ieee.

Grósz, T., Kurimo, M., et al. (2020). Visual interpretation of DNN-based acoustic models using deep autoencoders. In *International Workshop on Machine Learning in Visualisation for Big Data*, pages 25–29.

Hallé, P. A., Chang, Y.-C., and Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, 32(3):395–421.

Harrington, J., Kleber, F., and Reubold, U. (2008). Compensation for coarticulation,/u/-fronting, and sound change in Standard Southern British: An acoustic and perceptual study. *The Journal of the Acoustical Society of America*, 123(5):2825–2835.

Hazan, V. and Rosen, S. (1991). Individual variability in the perception of cues to place contrasts in initial stops. *Perception & Psychophysics*, 49(2):187–200.

Heinz, J. M. and Stevens, K. N. (1961). On the properties of voiceless fricative consonants. *The Journal of the Acoustical Society of America*, 33(5):589–596.

Herrmann, F., Cunningham, S. P., and Whiteside, S. P. (2014). Speaker sex effects on temporal and spectro-temporal measures of speech. *Journal of the International Phonetic Association*, 44(1):59–74.

Hochreiter, S. and Schmidhuber, J. (1996). LSTM can solve hard long time lag problems. *Advances in Neural Information Processing Systems*, 9:473–479.

Hollien, H. (1974). On vocal registers. *Journal of Phonetics*, 2(2):125–143.

Holt, L. L. and Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, 119(5):3059–3071.

Honorof, D. N. and Whalen, D. H. (2005). Perception of pitch location within a speaker's F0 range. *The Journal of the Acoustical Society of America*, 117(4):2193–2200.

Howie, J. M. and Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones*. Cambridge University Press.

Huang, S., Liu, J., Wu, X., Wu, L., Yan, Y., and Qin, Z. (1998). Mandarin broadcast news speech (HUB4-NE). LDC98S73. Web Download. Philadelphia: Linguistic Data Consortium.

Jadoul, Y., Thompson, B., and de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

Jelinek, F. (1997). *Statistical methods for speech recognition*. MIT press.

Jiang, B., Dunbar, E., Sonderegger, M., Clayards, M., and Dupoux, E. (2020). Modelling perceptual effects of phonology with ASR systems. In *Proceedings of CogSci 2020*, pages 2735–2741.

Jiang, B. and Kuang, J. (2016). Consonant effects on tonal registers in Jiashan Wu. *Proceedings of the Linguistic Society of America*, 1:30–1.

Jolliffe, I. (2003). Principal component analysis. *Technometrics*, 45(3):276.

Jongman, A. and McMurray, B. (2017). On invariance: Acoustic input meets listener expectations. In *The Speech Processing Lexicon*, pages 21–51. De Gruyter Mouton.

Jongman, A., Qin, Z., Zhang, J., and Sereno, J. (2016). Just noticeable differences for pitch height and pitch contour for Chinese and American listeners. *The Journal of the Acoustical Society of America*, 140(4):3225–3225.

Jongman, A., Wang, Y., Moore, C. B., and Sereno, J. A. (2006). Perception and production of Mandarin Chinese tones.

Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., and McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance*, 43(9):1594.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. (2008). Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3933–3936. IEEE.

Kim, D. and Clayards, M. (2019). Individual differences in the link between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience*, 34(6):769–786.

Kingston, J. (2011). Tonogenesis. *The Blackwell Companion to Phonology*, pages 1–30.

Kirby, J. P. (2014). Incipient tonogenesis in Phnom Penh Khmer: Acoustic and perceptual studies. *Journal of Phonetics*, 43:69–85.

Kleinschmidt, D. F. and Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2):148.

Kleinschmidt, D. F. and Jaeger, T. F. (2016). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? *Psychonomic Bulletin & Review*, 23(3):678–691.

Kleinschmidt, D. F., Weatherholtz, K., and Florian Jaeger, T. (2018). Sociolinguistic perception as inference under uncertainty. *Topics in Cognitive Science*, 10(4):818–834.

Kong, E. J. and Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, 59:40–57.

Kong, E. J. and Lee, H. (2018). Attentional modulation and individual differences in explaining the changing role of fundamental frequency in Korean laryngeal stop perception. *Language and Speech*, 61(3):384–408.

Kong, Y.-Y. and Zeng, F.-G. (2006). Temporal and spectral cues in Mandarin tone recognition. *The Journal of the Acoustical Society of America*, 120(5):2830–2840.

Kronrod, Y., Coppess, E., and Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, 23(6):1681–1712.

Kuang, J. (2011). Production and perception of the phonation contrast in Yi. [Master's thesis]. *University of California, Los Angeles*.

Kuang, J. (2013). The tonal space of contrastive five level tones. *Phonetica*, 70(1-2):1–23.

Kuang, J. and Liberman, M. (2015). The effect of spectral slope on pitch perception. In *INTERSPEECH 2015*, pages 354–358.

Laurent, R., Barnaud, M.-L., Schwartz, J.-L., Bessière, P., and Diard, J. (2017). The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception. *Psychological Review*, 124(5):572.

Lenth, R. (2018). Package 'lsmeans'. *The American Statistician*, 34(4):216–221.

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, 69(1):1–33.

Leung, K. K. and Wang, Y. (2020). Production-perception relationship of Mandarin tones as revealed by critical perceptual cues. *The Journal of the Acoustical Society of America*, 147(4):EL301–EL306.

Li, Q. and Chen, Y. (2016). An acoustic study of contextual tonal variation in Tianjin Mandarin. *Journal of Phonetics*, 54:123–150.

Lin, M. (1988). Putong hua sheng diao de sheng xue texing he zhi jue zhengzhao [Standard Mandarin tone characteristics and percepts]. *Zhongguo Yuyan*, 3:182–193.

Lisker, L. (1978). In qualified defense of VOT. *Language and Speech*, 21(4):375–383.

Lisker, L. and Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech*, 10(1):1–28.

Luo, X. and Fu, Q.-J. (2004). Enhancing Chinese tone recognition by manipulating amplitude envelope: Implications for cochlear implants. *The Journal of the Acoustical Society of America*, 116(6):3659–3667.

Lyons, J., Wang, D. Y.-B., Gianluca, Shteingart, H., Mavrinac, E., Gaurkar, Y., Watcharawisetkul, W., Birch, S., Zhihe, L., Hölzl, J., Lesinskis, J., Almér, H., Lord, C., and Stark, A. (2020). Python_speech_features: release v0.6.1.

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabi, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., et al. (2020). Earshot: A minimal neural network model of incremental human speech recognition. *Cognitive Science*, 44(4):e12823.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman and Co., Oxford.

Marslen-Wilson, W. D. and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10(1):29–63.

Massaro, D. W., Cohen, M. M., and Tseng, C.-y. (1985). The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. *Journal of Chinese Linguistics*, pages 267–289.

Mayo, C. and Turk, A. (2004). Adult–child differences in acoustic cue weighting are influenced by segmental context: Children are not always perceptually biased toward transitions. *The Journal of the Acoustical Society of America*, 115(6):3184–3194.

Mayo, C. and Turk, A. (2005). The influence of spectral distinctiveness on acoustic cue weighting in children's and adults' speech perception. *The Journal of the Acoustical Society of America*, 118(3):1730–1741.

McAuliffe, M. and Babel, M. (2016). Stimulus-directed attention attenuates lexically-guided perceptual learning. *The Journal of the Acoustical Society of America*, 140(3):1727–1738.

McClelland, J. L. and Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1):1–86.

Mei, T. (1977). Tones and tone sandhi in 16th century Mandarin. *Journal of Chinese Linguistics*, pages 237–260.

Meng, Q., Zheng, N., Mishra, A. P., Luo, J. D., and Schnupp, J. W. (2018). Weighting pitch contour and loudness contour in Mandarin tone perception in cochlear implant listeners. In *INTERSPEECH 2018*, pages 3768–3771.

Mitterer, H. and Blomert, L. (2003). Coping with phonological assimilation in speech perception: Evidence for early compensation. *Perception & Psychophysics*, 65(6):956–969.

Moore, C. B. and Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America*, 102(3):1864–1877.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453–467.

Nagamine, T., Seltzer, M. L., and Mesgarani, N. (2015). Exploring how deep neural networks form phonemic categories. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 1912–1916.

Nakagawa, S. and Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, 18:347–373.

Nearey, T. M. (1997). Speech perception as pattern recognition. *The Journal of the Acoustical Society of America*, 101(6):3241–3254.

Nordenhake, M. and Svantesson, J.-O. (1983). Duration of standard Chinese word tones in different sentence environments. *Working Papers*, 25:105 – 111.

Norris, D. and McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, 115(2):357.

Oden, G. C. and Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, 85(3):172.

Oh, E. (2011). Effects of speaker gender on voice onset time in Korean stops. *Journal of Phonetics*, 39(1):59–67.

Ohala, J. J. et al. (1990). The phonetics and phonology of aspects of assimilation. *Papers in Laboratory Phonology*, 1:258–275.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Peng, G., Zhang, C., Zheng, H.-Y., Minett, J. W., and Wang, W. S.-Y. (2012). The effect of intertalker variations on acoustic – perceptual mapping in Cantonese and Mandarin tone systems. *Journal of Speech, Language, and Hearing Research*, 55(2):579–595.

Potisuk, S., Gandour, J., and Harper, M. P. (1997). Contextual variations in trisyllabic sequences of Thai tones. *Phonetica*, 54(1):22–42.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Przedlacka, J. and Ashby, M. (2011). Acoustic correlates of glottal articulations in Southern British English. In *Proceedings of the 17th International Congress on Phonetic Sciences*, pages 1642–1645.

Qian, N. (1992). *Studies on Modern Wu Dialect*. Shanghai: Shanghai Education Press.

Raphael, L. J. (2021). Acoustic cues to the perception of segmental phonemes. *The handbook of speech perception*, pages 603–631.

Ren, N. (1987). *An acoustic study of Shanghai stops*. PhD thesis, University of Connecticut.

Rhee, N. and Kuang, J. (2020). The different enhancement roles of covarying cues in Thai and Mandarin tones. *INTERSPEECH 2020*, pages 2407–2411.

Riehl, A. (2003). American English flapping: Evidence against paradigm uniformity with phonetic features. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 2753–2756.

Sadakata, M. and Sekiyama, K. (2011). Enhanced perception of various linguistic features by musicians: A cross-linguistic study. *Acta psychologica*, 138(1):1–10.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, 12(4):348–351.

Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49(5):336–347.

Scharenborg, O., Norris, D., Ten Bosch, L., and McQueen, J. M. (2005). How should a speech recognizer work? *Cognitive Science*, 29(6):867–918.

Scharenborg, O., Tiesmeyer, S., Hasegawa-Johnson, M., and Dehak, N. (2018). Visualizing phoneme category adaptation in deep neural networks. In *INTERSPEECH 2018*, pages 1482–1486.

Schatz, T., Bernard, M., Thiolliere, R., and Cao, X.-N. (2016). Abkhazia. [Online]. Available: https://abkhazia.readthedocs.io/en/latest/index.html.

Schatz, T., Feldman, N. H., Goldwater, S., Cao, X.-N., and Dupoux, E. (2021). Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, 118(7).

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., and Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH 2013*, pages 1–5.

Schertz, J. and Clare, E. J. (2020). Phonetic cue weighting in perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science*, 11(2):e1521.

Scobbie, J. M. (2009). Flexibility in the face incompatible english vot systems. In *Laboratory phonology 8*, pages 367–392. De Gruyter Mouton.

Shen, X. S. and Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech*, 34(2):145–156.

Shen, X. S., Lin, M., and Yan, J. (1993). F0 turning point as an F0 cue to tonal contrast: a case study of Mandarin tones 2 and 3. *The Journal of the Acoustical Society of America*, 93(4):2241–2243.

Shuai, L. and Malins, J. G. (2017). Encoding lexical tones in jTRACE: a simulation of monosyllabic spoken word recognition in Mandarin Chinese. *Behavior Research Methods*, 49(1):230–241.

Shultz, A. A., Francis, A. L., and Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America*, 132(2):EL95–EL101.

Sonderegger, M. and Yu, A. (2010). A rational account of perceptual compensation for coarticulation. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, pages 375–380.

Spille, C., Kollmeier, B., and Meyer, B. T. (2018). Comparing human and automatic speech recognition in simple and complex acoustic scenes. *Computer Speech & Language*, 52:123–140.

Stevens, K. N. and House, A. S. (1955). Development of a quantitative description of vowel articulation. *The Journal of the Acoustical Society of America*, 27(3):484–493.

Stuart-Smith, J. (1999). Glottals past and present: A study of T-glottalling in Glaswegian. *Leeds Studies in English*, pages 181–204.

Sussman, H. M. and Shore, J. (1996). Locus equations as phonetic descriptors of consonantal place of articulation. *Perception & Psychophysics*, 58(6):936–946.

Tanner, J., Sonderegger, M., and Stuart-Smith, J. (2020). Structured speaker variability in Japanese stops: Relationships within versus across cues to stop voicing. *The Journal of the Acoustical Society of America*, 148(2):793–804.

Ten Bosch, L. and Boves, L. (2018). Information encoding by deep neural networks: what can we learn? In *INTERSPEECH 2018*, pages 1457–1461.

Titze, I. R. (1994). Toward standards in acoustic analysis of voice. *Journal of Voice*, 8(1):1–7.

Toscano, J. C. and McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, 34(3):434–464.

Tupper, P., Leung, K., Wang, Y., Jongman, A., and Sereno, J. A. (2020). Characterizing the distinctive acoustic cues of Mandarin tones. *The Journal of the Acoustical Society of America*, 147(4):2570–2580.

Van der Hulst, H. (2016). Vowel harmony. In *Oxford Research Encyclopedia of Linguistics*, pages 495–534. Blackwell.

Wang, W. S. and Li, K.-P. (1967). Tone 3 in Pekinese. *Journal of Speech and Hearing Research*, 10(3):629–636.

Wang, W. S. Y. (1967). Phonological features of tone. *International Journal of American Linguistics*, 33(2):93–105.

Weber, A. (2001). Help or hindrance: How violation of different assimilation rules affects spoken-language processing. *Language and Speech*, 44(1):95–118.

Weber, P., Bai, L., Russell, M. J., Jancovic, P., and Houghton, S. M. (2016). Interpretation of low dimensional neural network bottleneck features in terms of human perception and production. In *INTERSPEECH 2016*, pages 3384 – 3388.

Weinreich, U. (2012). *Is a structural dialectology possible?* De Gruyter Mouton.

Weiss, G., Goldberg, Y., and Yahav, E. (2018). On the practical computational power of finite precision RNNs for language recognition. *arXiv preprint arXiv:1805.04908*.

Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. *The Journal of the Acoustical Society of America*, 93(4):2152–2159.

Whalen, D. H. and Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*, 49(1):25–47.

Whiteside, S. P. (1996). Temporal-based acoustic-phonetic patterns in read speech: Some evidence for speaker sex differences. *Journal of the International Phonetic Association*, 26(1):23–40.

Wiener, S. (2017). Changes in early L2 cue-weighting of non-native speech: Evidence from learners of Mandarin Chinese. In *INTERSPEECH 2017*, pages 1765–1769.

Wiener, S. and Ito, K. (2015). Do syllable-specific tonal probabilities guide lexical access? Evidence from Mandarin, Shanghai and Cantonese speakers. *Language, Cognition and Neuroscience*, 30(9):1048–1060.

Wiener, S. and Ito, K. (2016). Impoverished acoustic input triggers probability-based tone processing in mono-dialectal Mandarin listeners. *Journal of Phonetics*, 56:38–51.

Williams, D. and Escudero, P. (2014). A cross-dialectal acoustic comparison of vowels in Northern and Southern British English. *The Journal of the Acoustical Society of America*, 136(5):2751–2761.

Xu, B. and Tang, Z. (1988). A description of the urban Shanghai dialect. *Shanghai: Shanghai Educational Publishing House*.

Xu, Y. (1994). Production and perception of coarticulated tones. *The Journal of the Acoustical Society of America*, 95(4):2240–2253.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, 25(1):61–83.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*, 27(1):55–105.

Yang, B. (2015). *Perception and production of Mandarin tones by native speakers and L2 learners*. Springer.

Yang, J., Zhang, Y., Li, A., and Xu, L. (2017). On the duration of Mandarin tones. In *INTERSPEECH 2017*, pages 1407–1411.

Yang, Y. and Chen, S. (2020). Revisiting focus production in Mandarin Chinese: Some preliminary findings. *Proceedings of Speech Prosody 2020*, pages 260–264.

Yip, M. (2001). Tonal features, tonal inventories and phonetic targets. *UCL Working Papers in Linguistics*, 13:161–188.

Yip, M. (2002). *Tone*. Cambridge University Press.

Yip, M. J. (1980). *The tonal phonology of Chinese*. PhD thesis, Massachusetts Institute of Technology.

Yonezawa, T., Suzuki, N., Mase, K., and Kogure, K. (2005). Handysinger: Expressive singing voice morphing using personified hand-puppet interface. In *Proceedings of the 2005 Conference on New Interfaces for Musical Expression*, pages 121–126.

Yu, A. C. and Zellou, G. (2019). Individual differences in language processing: Phonology. *Annual Review of Linguistics*, 5:131–150.

Yu, G. (1988). Syllabary of homophones in Jiaxing dialect. *Dialect*, 3:195–208.

Yu, K. M. and Lam, H. W. (2014). The role of creaky voice in cantonese tonal perception. *The Journal of the Acoustical Society of America*, 136(3):1320–1333.

Yuan, J., Ryant, N., and Liberman, M. (2015). Mandarin Chinese phonetic segmentation and tone. LDC2015S05. Web Download. Philadelphia: Linguistic Data Consortium.

Zhang, J. (2006). *The phonology of Shaoxing Chinese*. PhD thesis, LOT Utrecht.

Zhang, J. and Liu, J. (2011). Tone sandhi and tonal coarticulation in Tianjin Chinese. *Phonetica*, 68(3):161–191.

Zhang, J. and Yan, H. (2015). Contextually dependent cue weighting for a laryngeal contrast in shanghai wu. In *Proceedings of International Congress on Phonetic Sciences*, pages 147–151.

Zhang, J.-S. and Hirose, K. (2000). Anchoring hypothesis and its application to tone recognition of Chinese continuous speech. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1419–1422.