Acoustic Distance, Acoustic Absement, and the Lexicon

by

Matthew C. Kelley

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics
University of Alberta

Examining committee:

Benjamin V. Tucker, Supervisor
Stephanie Archer, Supervisory Committee
Michael Kiefte, Supervisory Committee
Andrea MacLeod, Examiner
Michael Vitevitch, External Examiner

Abstract

It is common in linguistic analysis to treat words as strings of speech segments that are believed to be transduced from the speech signal. However, there are notable shortcomings with this approach, especially concerning word comparison. Principally, comparing speech segment strings does not directly assess the acoustic similarity of words, despite theories and evidence that words that sound similar compete for activation during spoken word recognition. The present dissertation aims to provide a perceptually-grounded method by which words can be represented acoustically and then compared with the dynamic time warping algorithm. The dissertation comprises three studies. The first study is a regression analysis to demonstrate the relationship between acoustic distance and spoken word recognition. It also investigates how to derive more abstract acoustic representations for words based on productions from multiple speakers. The second study investigates what sort of spectral distance function best reflects human perception of acoustic distance. It also examines human perceptual sensitivity to duration differences. The third study compares speech features that are learned by a neural network to mel frequency cepstral coefficients to determine which style of representation for the speech signal better reflects perception. The neural network features are an ensemble of features specific to certain regions of the speech spectrum, while the mel frequency cepstral coefficients are a summary of the entire spectrum. Together, these studies inform the processes of converting the speech signal to an acoustic representation and tuning acoustic comparisons so that they better relate to human cognition.

Preface

The three body chapters of this dissertation are intended to be published as standalone, separate research articles. Each chapter has its own introduction and conclusion. Some chapters refer to supplementary materials, and these materials are included as appendices at the end of the present dissertation. Benjamin V. Tucker supervised these projects and provided conceptual and editorial feedback throughout the writing process.

Chapter 2 has been submitted and is undergoing revision as: Kelley M. C., & Tucker, B. V. (2021). Using acoustic distance and acoustic absement to quantify lexical competition.

The study in Chapter 3 received ethics approval from the University of Alberta Research Ethics Board, Project Name "Describing human judgments of acoustic distance", No. Pro00103566, October 10, 2020.

Dedication

To my parents and brother, for their unwavering support and unconditional love. And for Jackie, a wonderful cat who will be sorely missed.

Acknowledgments

I firmly believe that large projects such as this dissertation take a village, and I would like to take this opportunity to thank the small portion of my village that I can consciously recall for supporting me throughout this endeavor. It is inevitable that I will neglect to mention some names, but if you were involved at all, I thank you.

I will begin by profusely thanking Kelly Arispe at Boise State University, who taught the first linguistics course I ever took, an introduction to Hispanic linguistics. The course was extremely engaging, and I would not have changed my major to linguistics without having taken such a wonderful course. I must also extend my heartfelt appreciation and gratitude to Michal Temkin Martinez, also at Boise State University. Without her guidance and connections, I never would have found my way up to the University of Alberta to begin my graduate career. I also owe a debt of gratitude to my family—my mother, Michelle; my father, Jeff; and my brother, Kevin—for rallying behind me when I was hesitant about moving to a new country and all that that process entailed. I have deeply enjoyed and benefited their affirmations and support over these years.

At the University of Alberta Department of Linguistics, I was met with an abundance of amazing and supportive people. I owe an enormous amount to the members

of the Alberta Phonetics Lab who immediately made me feel welcome and listened to my crazy ideas. Thank you so much to Catherine Ford, Pearl Lorentzen, Graham Tomkins Feeny, Aziz Alarifi, Yoichi Mukai, Dan Brenner, and Jae-Hyun Sung. Thank you as well to the colleagues who became a part of the lab later and who, like those previously mentioned, had many fruitful discussions with me and were just all-around good people to spend time with, Filip Nenadić, Scott Perry, Ivy Mok, Ryan Podlubny, and Annika Nijveld. I must also express my gratitude to the friends I made from other areas of the department, Katie Schmirler, Atticus Harrigan, Keely Paige, Adriana Soto-Corominas, Brian Rusk, Dalia Cristerna-Román, Regina Hert, Magda Difani, Kaidi Lõo, and everyone else whom I have interacted with. I was also met with amazing faculty who were supportive of my graduate work, Terry Nearey and Stephanie Archer, whom I owe a debt of gratitude. I also deeply thank Chelsea Fairweather and Bessie Yang for helping me survive living in university housing. I am also grateful to my examining committee (Michael Vitevitch, Andrea MacLeod, Michael Kiefte, Stephanie Archer, and Benjmain V. Tucker) and chair (Martin Guardado) for their guidance and discussion and an especially enjoyable defense.

Of course, I would be remiss if I did not mention the enormous thanks that I owe to my supervisor, Benjamin V. Tucker. He welcomed me into his lab and made me a phonetician and scientist. Thank you for encouraging me to pursue topics that interest me, pushing me to incorporate my computer science background into my work, and knowing when to reign me in when I was overreaching or being too radical. I know that there was no better place for me to end up earning my degree and no better person to be supervising me.

Finally, thank you to Jarrett Glass-Jeffrey for listening to me process my teaching and research conundrums night after night while we played video games. I don't know if I would have stayed stable and sane without you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The traditional representation of words and speech segments in phonetics and linguistics as a whole is a string of symbols. These symbols most often represent phones or phonemes, and this is often a useful representation. It is compact, discrete, and often unvarying. Whereas, recordings of speech are storage-intensive, they represent continuous events, and two recordings of the same word are virtually never identical. Yet, acoustic representations of speech are both a far more faithful representation of speech events and much more similar to the phenomenon that a human listener encounters when hearing speech. So, it is striking that there has yet to be a well-tested method to represent and compare words acoustically. This dissertation proposes the use of dynamic time warping to quantify differences between acoustic representations of speech. It analyzes various aspects of the dynamic time warping algorithm as relevant to speech perception and spoken word recognition. The present chapter serves to situate acoustic representations and comparisons in linguistic theory and provide additional information and context on dynamic time warping that will be relevant to motivating the remaining chapters.

Note that the term “representation” is used throughout this dissertation. It is common in linguistics to use “representation” with specific reference to how a linguistic object exists in the mind. However, the use of the term here is meant to invoke the more general sense of one object standing in for another. In this instance, a concrete, measurable acoustic object is used as a manipulable stand-in for the more nebulous and abstract concept of a word. That is, the acoustic representation is an instantiation of a word. This meaning of representation is rooted in the concept of representation in fields like mathematics (Goldin, 2014). In this sense, the acoustic representations are used in an instrumentalist sense more so than a realist sense.

1.1 Theoretical motivation

Within linguistics, words are commonly represented in a textual, discrete, and symbolic format for analytical purposes. Without a doubt, these formats are incredibly useful and allow for a wide range of analyses. Discrete, symbolic representations of language allow for the application of a wide range of methodology from formalized disciplines like discrete mathematics to theoretical computer science (Port & Leary, 2005). Such a treatment has been incredibly fruitful for many subdisciplines of linguistics. However, using this symbolic representation requires a transduction of the continuous speech signal into discrete elements. Often, this transduction forces the loss of some or all of the acoustic properties of the speech signal. Some approaches to phonology believe the lack of acoustic “substance” in symbolic representations is ideal and accurate for cognitive representations, as evidenced in substance-free phonology (Reiss, 2017). However, this loss of acoustic information is certainly not ideal in the context of research on speech perception and spoken word recognition as it is likely that listeners make use of this information. A wealth of evidence suggests that the sound similarity between words in the lexicon can influence the speed with which a word is recognized in speech (reviewed in Vitevitch & Luce, 2016).

Assessing sound similarity would make the most sense to do acoustically. Yet, there remains a gap in terms of representing segments and words acoustically. It may be the case that acoustic representations of words are not needed for certain aspects of cognitive modeling, but they are necessary to make acoustic comparisons between words. Local acoustic comparisons have existed in the phonetics literature for decades, including comparisons of voice onset time (Lisker & Abramson, 1964), spectral moments (Jongman et al., 2000), and formants (Peterson & Barney, 1952). No local acoustic measurement or characteristic can be applied across all speech sounds, however, so none of these methods can be simply extended to encompass word-level comparisons. Word-level acoustic comparisons have been carried out previously (Bartelds et al., 2020; Heeringa, 2004; Kirchner et al., 2010; Lewandowski, 2012; Lewandowski & Jilka, 2019; Mermelstein, 1976), but the acoustic representations and distance functions used were not cognitively validated with behavioral data from a large variety of words.

The present dissertation extends previous research on acoustic word-level comparisons by testing various components involved in these comparisons and grounding the acoustic comparisons in human cognition using behavioral data. A running theme throughout the dissertation is also what, exactly, the nature of a non-trivial acoustic representation of words might look like. Trivial representations could simply be recordings of words, but such a representation neglects the abstraction process that occurs when less relevant information is filtered out of an object. The present work focuses on abstraction regarding acoustic information and regarding exemplars of words. Abstraction over acoustics concerns determining what parts of the acoustic signal are meaningful, while abstraction over exemplars has to do with how individual tokens are categorized into larger units and what those larger units might look like. Both of these types of abstraction are needed so that acoustic comparisons between words compare relevant parts of the acoustic signal and are applicable to more than just individual recordings of words. These are common themes in the tension between exemplar-based approaches to speech perception (e.g., Goldinger, 1998; Johnson, 1997; Pierrehumbert, 2001) and abstractionist/prototype-based auditorist approaches to speech perception (e.g., Diehl & Kluender, 1989; Farrar, 1981; Kingston & Diehl, 1995; Kuhl, 1992).

1.1.1 Acoustic representation

The nature of acoustic representation that is used throughout the dissertation is continuous rather than discrete. The continuous nature of the representations is similar to that of a spectrogram, where frequency information and intensity information are represented across time. Because these representations are manipulated with digital computers, they are necessarily discretized, finite approximations of continuous phenomena. These acoustic representations are intended to be treated as physical representations of words; that is, they are simply a time-bound sequence of acoustic information. They are intended to be representations of sequences of acoustic information expected in the speech signal for a given word. In this sense, acoustic representations could be thought of as acoustic "templates" in that they specify a template of a word that the speech signal can be checked against, similar to Klatt (1981). They are not intended to be literal neurological representations of words in the brain. They are also not intended to be interpreted as individual word detectors à la the logogen model of spoken word recognition (Morton, 1969). To some degree, the acoustic representation is merely a change in perspective from a symbolic representation. A string of segments that is used to represent a word in the lexicon can be taken as an expectation of what the acoustic signal will sound like, after transducing each segment into the acoustic cues that are associated with it. However, the transduction of segments into acoustics necessarily requires some storage of the cue associations between segments and acoustics. Within the acoustic representation, that information is already present.

This style of acoustic representation has potentially far-reaching ramifications for speech production as well as speech perception. If words are treated as sequences of acoustic information, there is no longer a need to posit a transduction between segments and articulation during speech planning. Instead, there would be a stored relationship between acoustic qualities or targets and the articulations that achieve those acoustic qualities. The fact that an acoustic representation is being used here quite clearly does not imply that humans use acoustic representations in this fashion nor that speech production is only a mapping between acoustic and articulatory information without the need for segments as an intermediary. These questions are not explicitly addressed in the dissertation, but they do frame the possible future consequences of acoustic representations if they prove analytically useful.

1.1.2 Acoustic comparisons

Being that the acoustic representations are time-bound, there is a natural difficulty in comparing words to each other. Virtually no two words or even two pronunciations of the same word will be identical in duration. This problem is modestly alleviated by a finite representation due to the pigeonhole principle, which basically states that if you have n items but fewer than n boxes in which to put the items, some of the boxes must contain more than one item (consult Keller & Trotter, 2017, Chapter

4). However, even though the word-level duration between two recordings may occasionally coincide, segment-level duration is unlikely to coincide between the words. Naturally, these temporal differences between segments (and often words) mean that more care must be given to comparing the acoustic representations than simply finding the Euclidean distance between them; in many cases, it will be impossible since the discretized representations of the words are of different lengths, and in other cases, non-similar portions of the words may end up compared to each other. The latter problem is severe when it comes to cognitive modeling since humans permit some elasticity in terms of temporally matching an input to a lexical template; were this not the case, every uttered exemplar of a segment category would need to be identical in length to all other exemplars of that category, which is obviously false.

A family of algorithms exists that by default allow for distance-over-time calculation with elastic correspondences between time steps. These algorithms are generally referred to as elastic matching algorithms (Uchida & Sakoe, 2005) or elastic alignment algorithms (Abanda et al., 2019). Because these algorithms compare sequences across time, they naturally fit in with the mathematical notions of dynamical systems and differential equations. Indeed, they sum distance values across the time dimension, a quantity which is referred to as "absement" (Mann et al., 2018; Mann et al., 2006). This situation is analogous to how summing velocity values across time yields the distance quantity. By nature of using such algorithms, the dynamics—that is, patterns of changes over time—of acoustic differences between words are centered. While many of these algorithms may potentially be useful, one particular algorithm within this family stands out for use in speech due to its history of use in automatic speech recognition: dynamic time warping. Using dynamic time warping to compare words is a direct extension of work dating back at least to Mermelstein (1976), who used a temporal alignment algorithm to calculate a form of distance with Mel frequency cepstral coefficients (MFCCs) as the acoustic representation.

1.2 Dynamic time warping in detail

Dynamic time warping is a dynamic programming method used to compute the following recurrence relation (similar to a recursive phrase-structure rule or syntactic schema) for two sequences S_1 and S_2 of length M and N, respectively:

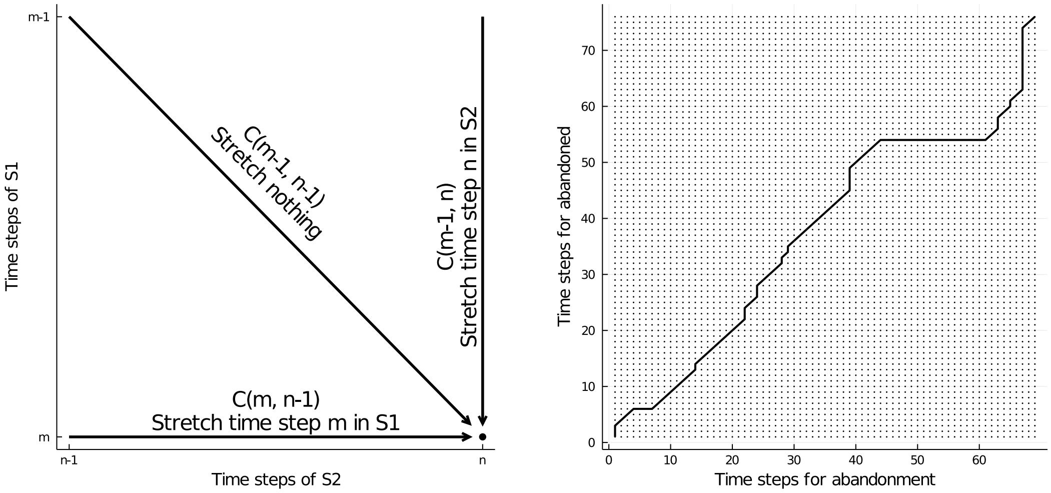
$$C(m,n) = d(S_1^m, S_2^n) + \min(C(m-1,n), C(m,n-1), C(m-1,n-1)) \tag{1.1}$$

$$C(m,1) = d(S_1^m, S_2^1) + C(m-1,1) \tag{1.2}$$

$$C(1,n) = d(S_1^1, S_2^n) + C(1,n-1) \tag{1.3}$$

$$C(0,0) = C(0,n) = C(m,0) = \infty\,, \tag{1.4}$$

where $C(m,n)$ indicates the cost function evaluated at time step m in S_1 and time step n in S_2, $d(\cdot,\cdot)$ indicates a distance function like Euclidean distance, and the notation S_1^m indicates the m-th time step of S_1, and S_2^n indicates the n-th time step of N. When the min function in Equation 1.1 chooses $C(m-1,n)$, the n-th time step of S_2 is being stretched over an additional time step in S_1. When the min function chooses $C(m,n-1))$, the m-th time step of S_1 is being stretched over an additional time step in S_2. And, when the min function chooses $C(m-1,n-1)$, neither sequence is being stretched in time at the corresponding time steps. These possibilities are displayed in Figure 1.1a. It may appear as though S_1 and S_2 are mislabeled on the axis labels or in the text annotations, but they are not. Rather, it is what is being held constant with each path option that must be interpreted. For example, in the horizontal line at the bottom, it is the time step of S_1^m that is being held constant across the time steps of S_2. This means that S_1^m is being stretched across two time steps in S_2.



(a) Possible paths checked in the min function to calculate $C(m, n)$.

(b) Optimal path through dynamic programming trellis for the words *abandonment* and *abandoned*, which are on the x- and y-axes, respectively.

Figure 1.1: Dynamic time warping process shown for an individual frame and for comparing two different words.

This recurrence is often (though not always) computed by filling in a dynamic programming matrix, from which the lowest cost alignment path can be determined via backtracking from the value of $C(M, N)$. An example of this path for the words *abandonment* and *abandoned* from the Massive Auditory Lexical Decision database (Tucker et al., 2019) is given in Figure 1.1b. The overall dynamic time warping cost between the two words is equivalent to summing the distance between each point along the black warping path. Note that the warping path is not a straight line, indicating that some time steps in each word are compared to multiple time steps in the other. A straight line is also impossible because the words have a different number of discrete time steps.

Dynamic time warping was historically used in automatic speech recognition since Sakoe and Chiba (1970), though it has since fallen out of popularity in favor of deep learning models and hidden Markov models. It is still popular when comparing time sequences in data mining, such as in Petitjean et al. (2014) and Rakthanmanon et al. (2012). In fact, it often provides state-of-the-art results when comparing the similarity of time series (Ding et al., 2008). Given its popularity in comparing time-series and the frequent analysis of time-series data in phonetics and linguistics, it is conspicuous that dynamic time warping has not been taken up significantly in our field yet, outside of a handful of studies such as Mermelstein (1976), Kirchner et al. (2010), Mielke (2012), McCloy (2013), and Bartelds et al. (2020). Perhaps this is because it is not obviously relevant to the sorts of questions that researchers ask. Indeed, it may be the case that the focus on segmental analysis of speech has obscured the potential utility in dynamic time warping as a research tool. It does, after all, require numerical distances between the objects being compared, and methods for computing the distance between a character and a sound are not obvious.

Previous research has pointed out that dynamic time warping and the Viterbi algorithm bear considerable similarities. Unlike dynamic time warping, the Viterbi algorithm uses a more symbolic approach for speech recognition that involves phoneme recognition in tandem with word recognition. These algorithms are so similar, in fact, that they can be said to optimize the same quantity (Oates et al., 2000). It is worth considering some aspects of this similarity in greater detail, in addition to potential differences.

There are a number of algorithms that are related to the Viterbi algorithm that are also used in speech recognition, like the token passing algorithm and the beam search algorithm, such as given in Graves (2012). They are all interrelated by virtue of being dynamic programming algorithms. All of these algorithms are designed to perform a search over all possible symbol strings to find the one that best matches the acoustic input. For automatic speech recognition, it is typical to search both phoneme strings and word strings in parallel, one coming from the other via dictionary lookup. Although, newer models map directly to graphemes (e.g., Amodei et al., 2016). These systems are trained to perform categorization, providing probability distributions over phonemes or graphemes at each time step of audio being processed. In so doing, these models provide an alternative solution to the difficulty in measuring the similarity between symbols and acoustic input.

It is worth considering some of the assumptions that using dynamic time warping makes and how those assumptions relate back to phonetics and speech perception. The first assumption has to do with what sorts of objects are compared with each other. Dynamic time warping is often used to compare like-to-like. That is, the representation of both objects is often the same, like comparing two different formant tracks to each other. The underlying assumption in systems using dynamic time warping for speech recognition is that it is tenable to represent words acoustically for speech recognition systems. By using dynamic time warping as a form of acoustic comparison between words, the same assumption is made in this dissertation. But, the assumption is narrower in that representing words acoustically is not just useful for engineering purposes, but also for cognitive modeling. These assumptions necessitate experimental validation. At the word level, experimental results are needed that the results of dynamic time warping relate to human perception. And, at the subsegmental level, experimental results are needed that distance comparisons made in dynamic time warping relate to human distance perception, as well as that the acoustic representation itself relates to human perception. The studies in this dissertation are designed to gather experimental evidence related to those assumptions.

Contrast that sort of like-to-like comparison with neural networks used for speech recognition. The neural networks are trained to produce similarity scores in the form of probabilities between slices of acoustic information and symbols. There is an underlying assumption that words are best represented as phoneme or grapheme strings. Mapping acoustics onto symbols reliably is a remarkably difficult problem, as anyone who has ever performed spectrographic analysis of speech is well aware (see the history in Shankweiler & Fowler, 2015). This is doubly true in the face of results like Ladefoged and Broadbent (1957), where the same acoustic input can be assigned to different phoneme categories based on previous formant context. These results suggest that context-free phoneme categorization is a one-to-many relationship. That is, a single input could have multiple possible outputs associated with it. Magnuson et al. (2020) similarly suggested that the relationship is many-to-many, that is, that multiple inputs pair with multiple outputs. These types of relationships are not learnable by neural networks and can only be approximated. This is because neural networks learn continuous functions (Cybenko, 1989), which can only represent one-to-one and many-to-one relationships. Dynamic time warping on acoustic lexical templates, then, may present the opportunity to study spoken word recognition with processes that are more computationally tractable than some sort of transduction from the acoustic signal to phoneme categories *per se*.

In dynamic time warping, the non-linear mapping produced between two words allows for the accumulation of acoustic distance across time. This accumulation represents the sustained distance between two words in time, that is, absement. In some sense, absement can be thought of as being inversely related to activation during spoken word recognition, in that higher values indicate poorer matches between the incoming signal and an item in the lexicon. This quantity is central to Chapter 2, which attempts to create a new measure of lexical competition based on acoustics. Absement values are calculated between each pair of the 26,793 English words in the Massive Auditory Lexical Decision data set (Tucker et al., 2019), and the mean value from a given word to all words in the data set is taken to be its "acoustic distinctiveness." This value is related to auditory lexical decision behavior as a measure of competition, with the hypothesis that words with low values of acoustic distinctiveness—that is, words that sound like many other words—will take longer to respond to, and vice-versa. This hypothesis mirrors results using phonological neighborhood density as an index of lexical competition during spoken word recognition (Luce, 1986; Luce & Pisoni, 1998; Vitevitch & Luce, 2016).

By using dynamic time warping, another assumption is made about how slices of acoustic information should be compared to each other. It is common to use Euclidean distance or squared Euclidean distance when computing dynamic time warping (Rakthanmanon et al., 2012). It is not clear that this is necessarily the best distance function to use to compare slices of acoustic information. That is, Euclidean distance may not be the best reflection of human cognition. In a more abstract sense, it is also not clear what sorts of acoustic features are best suited to use as the acoustic representation of a word. The notion of the feature set should not be overlooked; indeed, the feature set is what will ultimately determine what sounds are close to each other and what sounds are far apart from each other, and the distance function is merely a choice about how to quantify that distance. To more concretely exemplify this relationship between feature sets and distance functions, consider a feature set that consists only of F_1 values for vowels. It would be the case, then, that [i] and [u] would be very close together, regardless of whether the distance is quantified with Euclidean distance, Manhattan distance, etc. Taking the mean formant values by category for the male speakers in the Hillenbrand et al. (1995) data provided in the `phonTools` R package (v0.2-2.1 Barreda, 2015), [i] and [u] are only 36.98

Hz apart, compared to [i] and [ɪ] being 86.67 Hz apart. However, on adding F_2 to the feature set, [i] and [u] are suddenly much further apart at 1331.05 Hz using Euclidean distance, compared to [i] and [ɪ] being 301.61 Hz apart. This example underscores the importance of choosing a set of features; no distance function can make up for a feature set that does not appropriately separate objects. Both the feature set selection and the distance function selection must be considered together.

1.3 The present dissertation

At its core, the present dissertation seeks to provide experimental and empirical validation for using dynamic time warping in phonetics and linguistics. Chapter 2 relates the cost value derived from dynamic time warping to lexical competition, in comparison with phonological neighborhood density. Chapter 3 compares possible distance functions to human judgments of acoustic distance in perceptual tasks. It additionally derives temporal bounds to impose on the dynamic time warping process based on perceptual sensitivity to duration differences in vowels. Chapter 4 compares representing words as hand-crafted Mel frequency cepstral coefficient (MFCC) sequences to representing words with acoustic features learned by a neural network trained to perform formant and vocal tract resonance tracking. The chapters are interrelated, but each one has been written to serve as a standalone paper. As such, there are separate abstracts and introductions for each chapter to provide a more in-depth contextualization of the relevant topics and previous literature. Chapter 5 connects the results of each chapter together and offers a speculative hypothesis about speech communication as an acoustically-driven, goal-oriented process as informed by these results.

Chapter 3 and Chapter 4 together address the issue of choosing a distance function and choosing a feature set, respectively. The research question for Chapter 3 is what distance function should be used to compare audio slices to each other, in addition to how many other audio slices in the future and in the past a particular slice should be able to be compared to. These questions are addressed by modeling the results from a rating task and a duration discrimination task. In this way, the selected distance function can be related to human perception, as can the temporal elasticity, so to speak, of the audio slices in acoustically represented words. The results from both experiments are then used to re-analyze the results from Chapter 2 to situate the distance function and the elasticity in the context of spoken word recognition. The results speak to how duration differences might be accounted for cognitively during spoken word recognition and how acoustic information might be compared to a lexical template in the listener's mind.

Chapter 4 addresses the research question of what sort of features should be used to represent speech. The different features compared are MFCCs and features learned by a neural network to do formant tracking. The neural network features are interpreted in the context of determining what frequency components in a sound cause features to have high values. These features are compared with MFCCs by re-modeling the analyses from previous chapters and comparing results. The results from this analysis relate to what style of acoustic abstraction might be at play as regards spoken word recognition. The MFCCs represent summary style features that summarize the entire acoustic spectrum, whereas the neural network features each relate to the presence or absence of a combination of frequency components.

Together, these studies investigate the interplay between acoustic representations and mental acoustic comparisons. It is difficult to study one without the other, or indeed for one to exist without the other. By extension, these studies also evidence the interrelated nature of speech perception and speech production. Production data is used to explain perceptual phenomena that occur during spoken word recognition, and the acoustic representation for production data is designed to satisfy cognitive and perceptual desiderata.

Chapter 2

Using acoustic distance and acoustic absement to quantify lexical competition

This chapter has been submitted as: Kelley, M. C., & Tucker, B. V. (2021). Using acoustic distance and acoustic absement to quantify lexical competition.

Abstract

Phonological neighborhood density has been a common method to quantify lexical competition. It is useful and convenient, but it has shortcomings that are worth reconsidering. The present study quantifies the effects of lexical competition during spoken word recognition using acoustic distance and acoustic absement, rather than phonological neighborhood density. The indication of a word's lexical competition is given by what is termed its acoustic distinctiveness, which is taken as its average acoustic absement to all words in the lexicon. A variety of acoustic representations for items in the lexicon are analyzed. Statistical modeling shows that

acoustic distinctiveness has a similar effect trend as phonological neighborhood density. Additionally, acoustic distinctiveness consistently increases model fitness more than phonological neighborhood density, regardless of which kind of acoustic representation is used. Acoustic distinctiveness does not seem to explain all the same things as phonological neighborhood density, however. The different areas that these two predictors explain are discussed, in addition to potential theoretical implications of acoustic distinctiveness's usefulness in models. The paper concludes with reasons why a researcher may want to use acoustic distinctiveness over phonological neighborhood density in future experiments.

2.1 Introduction

In spoken word recognition, a listener must discriminate or recognize the word contained in an audio signal from among other potential candidates. One predominant metaphor used to describe this process is the activation/competition metaphor. Under this metaphor, potential matches for the word in the audio signal receive activation based on how well the acoustic information in the signal matches the listener's expectations for each word. A group of words that sound similar and are expected to compete have been called phonological neighborhoods (Luce, 1986; Luce & Pisoni, 1998). In Luce (1986), words are defined as neighbors on the basis of being one edit (phoneme addition, deletion, or substitution) away from each other. For example, some of the phonological neighbors of /kɪt/ are /skɪt/, /ɪt/, and /sɪt/. In this sense, the sound similarity between words is assessed using text in the form of phoneme strings. Competition is then quantified by counting the number of words in the lexicon that are neighbors with a given word. This count is defined as the word's phonological neighborhood density. Phonological neighborhood density has been found to be predictive of participant behavior in many psycholinguistic tasks. In auditory lexical decision, for example, high phonological neighborhood density values have been found to have inhibitory effects in English (Goldinger et al., 1989; Luce & Pisoni, 1998). However, facilitatory effects were found for Spanish (Vitevitch & Rodríguez, 2005) and Japanese (Yoneyama, 2002). See Vitevitch and Luce (2016) for a review of other tasks that this measure has been used for.

Yet, when the notion of phonological neighbors based on the one-edit rule was introduced, Luce (1986) remarked that a more sophisticated method of assessing sound similarity should eventually be used. He noted that the one-edit definition of neighbors applies equal weight to segmental substitutions wherever they occur in the word and does not reflect phonetic differences that would occur. For example, /kɪt/ would be considered as similar to /sɪt/ as it is to /kɪs/. What's more, equal weight is also assigned to any possible segmental change, so /pɪt/ would be considered to be as close to /bɪt/ as it is to /nɪt/, which does not reflect how the word or phrase position of a segment influences its production. This is in spite of the fact that not all speech sounds are equally similar to each other, which is readily apparent whether considering the sounds from an articulatory, auditory, or acoustic perspective.

While researchers have learned a lot about spoken word recognition and competition from the one-edit rule phonological neighborhood density, it is time to address these shortcomings. In the present study, we used acoustic distance comparisons to quantify the sound similarity between words. We then used those comparisons to operationalize lexical competition to model responses in an auditory lexical decision task and compare the results to phonological neighborhood density.

Previous research has not left Levenshtein distance or the one-edit rule unquestioned. Luce (1986) detailed more sophisticated methods of quantifying competition, ending on the frequency-weighted neighborhood probability rule. It incorporates lexical frequency, neighborhood density, and phoneme confusability. Neighbors are still detected based on the one-edit rule. However, despite the additional explanatory value of the frequency-weighted neighborhood probability rule, most studies that use, analyze, or control for phonological neighborhood effects have used the one-edit rule and classical phonological neighborhood density Vitevitch and Luce (2016). A modification to the one-edit definition of neighbors was proposed by Kapatsinski (2005) where neighbors are defined by having at least two thirds of their segments in common, as assessed by Levenshtein distance. However, this modification still does not address the original concerns about the type or position of change. In production, Nelson and Wedel (2017) suggested that the presence of minimal pairs was a better predictor than phonological neighborhood density for lexical competition during production. Switching to using the presence of a minimal pair does not resolve the concerns about the timing or type of change to the phonetic signal when assessing sound similarity, though.

It seems, then, that a method with more gradience than binary same/different comparisons is needed to assess the similarity of sounds. Comparisons between segments date back at least to Saporta (1955), who used distinctive features from English (Jakobson et al., 1952) and Spanish (Llorach, 1950) to calculate a sort of distance between segment pairs for each language. This style of assessing the similarity of sounds with distinctive features has found use in many other studies (Albright & Hayes, 2006; B. Allen & Becker, 2015; Frisch et al., 2004; Mohr & Wang, 1968). Other feature sets have also been used (Heeringa, 2004; Kondrak, 2000; Peterson & Harary, 1961; Sanders & Chin, 2009). Featural comparisons may very well be analytically useful, but it cannot be assumed *a priori* that similarity measures based on them will be relevant in acoustico-perceptual studies. From a perceptual perspective, Iverson et al. (1998) used confusion data to calculate used the phi-square coefficient, which is equivalent to the squared Pearson correlation between binary variables (Howell, 2008). This method found later adoption in Gahl and Strand (2016). However, phoneme confusion data is difficult to extend to the word level, and confusion in simple syllables may not relate well to confusion in longer words due to context effects.

Other researchers comparing linguistic units have instead focused on using acoustic data. Heeringa (2004) compared formant tracks using Euclidean distance in a dynamic programming paradigm with a speech rate normalization to ensure a consistent duration for every segment. A shortcoming of this method for use in perceptual work is the speech rate normalization, since speaking rate is ever-present in speech. Lewandowski and Jilka (2019) calculated acoustic similarity based on the amplitude envelopes of specific frequency bands of the signals in question using cross-correlation. Cross-correlation, though, does not deal with temporal distortions between two signals such as might occur between different productions of the same vowel.

Johnson (1997) and Yoneyama (2002) created acoustically-derived exemplar models of words. The exemplars used a vector-quantization technique on sequences of spectra to create the exemplars, and the vector-quantized exemplars were compared with an exponential function of Euclidean distance based on the quantized spectra. When exemplars were of different lengths, an alignment algorithm was used. As well, these representations do not truly resolve the highlighted issues for phonemic representations. The quantized spectra themselves—that is, the internal representations of the words—are discrete and effectively symbols.

Mielke (2012) introduced a method of calculating phonetic similarity between phone or phoneme categories. It works with Mel-frequency cepstral coefficient (MFCC) and delta coefficient representations of two audio signals, which is a form of time-frequency representation for sound. The distance is taken as the average distance between each pairing produced by the dynamic time warping algorithm, which finds the set of pairings between two signals that minimizes the accumulated distance between them, while maintaining temporal order. This method found later adoption in Bennett et al. (2018), who summed the distances instead of averaging them. Bartelds et al. (2020) also used dynamic time warping on MFCCs, delta coefficients, and delta-delta coefficients as a measure of pronunciation distance between words and also used a temporal normalization technique that is similar to averaging. Dynamic time warping has also been used in McCloy (2013) to align pitch and intensity contours and Kirchner et al. (2010) to create averages of speech exemplars.

2.1.1 The present study

Demonstrably, myriad methods have been used to quantify differences between words and sounds. However, fewer of these methods have been directly compared against the one-edit rule used to calculate phonological neighborhood density. Gahl and Strand (2016) found that some aspects of phonological neighborhood density did not reflect perceptual similarity, and Yoneyama (2002) reported better performance using more acoustic comparisons. There has yet to be a large-scale comparison between phonological neighborhood density and more acoustically-grounded methods, however.

The remainder of the paper describes a measure of lexical competition based on acoustic comparisons between words and analyzes auditory lexical decision data. This paper extends the methods in Mielke (2012), using dynamic time warping at the word level and in the realm of lexical competition. The most direct acoustic notion of a word having many or few phonological neighbors is whether a word is acoustically similar or distinct from many words. We refer to this as a word's "acoustic distinctiveness" and calculate this variable over a large lexicon of speech data by using dynamic time warping.

Being that dynamic time warping calculates distance at various time points in the signal and can handle temporal distortion, it seems a good candidate for assessing the similarity of sounds, as long as the format of the input captures the acoustic characteristics of the signal well. MFCCs are a good starting place to represent speech since they are the industry standard for speech recognition. Dynamic time warping also addresses the concerns about the type and position of different segmental changes, to the extent that they are present in the acoustic signal.

While some previous work on dynamic time warping has referred to its accumulated cost value output as a distance metric (e.g., Bennett et al., 2018), in the strict sense of a mathematical distance metric, this is inaccurate because dynamic time warping's output does not meet all the criteria necessary to be a distance metric. Bartelds et al. (2020) and Mielke (2012) avoided this problem by finding average or approximately average distances between aligned MFCC vectors in the dynamic time warping output. However, durational differences between otherwise acoustically similar segments will not be penalized in the output due to the nature of the alignment in vanilla dynamic time warping. Such differences may actually result in a lower average value due to a higher prevalence of small difference values in the set of numbers over which the average is calculated. Whereas, for spoken word recognition research, it is desirable for such durational mismatches to be penalized because duration is a cue for a variety of speech sounds like vowels and geminates. We believe, however, that there is an elegant solution at hand that also has a strong connection with kinematics. Specifically, acoustic distance forms the “interior” of dynamic time warping, so to speak, when a distance is computed between two chunks of audio. The accumulated distance that is output, then, is the absement between the two sequences being compared in dynamic time warping. In kinematics, absement is the time-integral of displacement or distance, and it is indeed the case that dynamic time warping sums distance over time. Absement has found use in fields such as musical instrument design (Mann et al., 2006) and kinesiological feedback (Mann et al., 2018). For vanilla dynamic time warping, absement would be the lowest accumulated mismatch between the two signals.

Our first analysis is a proof-of-concept where approximately 26,000 real word stimuli from an auditory lexical decision experiment are compared with each other to determine an overall acoustic distinctiveness value for each word using the concept of acoustic absement. The acoustic distinctiveness measure is then used as a statistical variable to predict the response latency of the participants in auditory lexical decision. The second analysis builds on the first but compares different ways of representing the words in the experiment, including using recordings from speakers that aren’t used in the auditory lexical decision stimuli and applying a sequence averaging technique to multiple recordings to create prototype acoustic representations. These results are compared with a statistical model that uses neighborhood density instead of acoustic distinctiveness to predict participant response latency. The third analysis investigates the extent to which acoustic distinctiveness and phonological neighborhood density overlap in the models. These analyses are followed by a general discussion of the results and why a researcher might choose to use acoustic distinctiveness over phonological neighborhood density.

2.2 Analyses and results

The data that are used in the analysis come from the freely available Massive Auditory Lexical Decision (MALD) data set (Tucker et al., 2019). MALD is an auditory lexical decision megastudy, with about 28,000 real words recorded by a young male speaker of western Canadian English. Each word was responded to in auditory lexical decision at least 4 times from among 231 unique participants who were also native speakers of western Canadian English, for a total of 227,129 data points (including responses to both real words and pseudowords). Stimuli sets were also recorded for two other speakers: a young female and an older male, both of whom are native speakers of western Canadian English. These other recording sets will be crucial for further development and testing of the acoustically-based measures of competition detailed later on in the present study. As such, only words that are common between these three speakers will be used, so that no particular word is left incomparable in the different representations developed herein. In total, there were 26,005 words in common between the speakers.

Further details are available in Tucker et al. (2019) on the recording process for the young male speaker, the auditory lexical decision task, and the variables included in the data set. The young female and older male speakers were recorded in a similar environment and with similar methods and equipment as the young male speaker.

2.2.1 Analysis 1

The first analysis used the stimuli from the auditory lexical decision task itself as templates to compare against each word. In this sense, the frequency information in the recordings was taken as an acoustic representation of the word.

Calculating acoustic distinctiveness

Each word was first converted to a Mel-frequency cepstral coefficient (MFCC) representation, similar to Mielke (2012). At a high level, this process converts the waveform of the audio into a transform of the frequency representation, similar in some ways to a spectrogram. More specifically, this process involves multiplying frames of the signal with a window function like a Hamming window, calculating Mel filterbanks for each windowed frame, and determining the cepstral coefficients for each filterbank with a discrete cosine transform. In the present analysis, a typical format used in speech recognition was selected, where the window length was 25 ms, and the step size for the windows was 10 ms. 13 coefficients were calculated, and the zeroth coefficient was replaced with the log energy of the frames.

Delta and delta-delta coefficients were not calculated, unlike standard practice in speech recognition and also unlike Bartelds et al. (2020) and Mielke (2012). The choice not to calculate them in the present paper was made on the grounds that the goal is to calculate the distance between time slices in the signals, and derivatives do not make sense to use in such calculations. For example, if you have two points in space and want to know the distance between them, only their current positions matter; how quickly they are moving in space does not matter.

Once the words were converted to an MFCC-by-time representation using the MFCC.jl package (v0.3.1 van Leeuwen, 2019) in the Julia programming language (v1.4.2, Bezanson et al., 2017), each individual word was acoustically compared to all other words and itself using the dynamic time warping algorithm. There was one instance of each word in the data set. After comparing each word to all the words, the mean of its absement to all words was calculated. This mean value was taken as an indicator of the word's acoustic distinctiveness, or how distinct it is on average from all the words in the lexicon. In terms of graph-theoretic (Vitevitch, 2008) and network scientific approaches to modeling connections between words in the lexicon (Vitevitch, 2008, 2021), the connections are modeled as a complete graph with the addition of a word being connected to itself. The weight on each connection is the acoustic absement. The acoustic distinctiveness value would then be a word's average connection weight. These calculations were performed using the Phonetics.jl (v0.1, Kelley, 2020) package and the DynamicAxisWarping.jl (v0.2.5, Bagge Carlson, 2020) packages.

Additionally, some words were recorded but not used in the experiment because there were not enough to fill an additional experimental list. As such, these words were used in calculating the acoustic distinctiveness for other words, but those words' acoustic distinctiveness values themselves were not used in the modeling process.

Statistical analysis

The acoustic distinctiveness values correlated highly with the duration of the stimuli ($r = .89$, $p < .001$). This is to be expected, however. The interval over which the acoustic distances are summed to calculate the absement between word pairs is linearly related to the duration of the stimuli (modulo some zero padding for the final window on which the MFCCs are calculated). And, absement increases monotonically over time in this case. The high correlation does not mean that these variables are the same, however. Consider that $f(x) = x^2$ and $g(x) = x$ also have a very high correlation when x is strictly positive, yet it is clear that x^2 and x are not equivalent.

What the correlation between duration and acoustic distinctiveness means practically is that they should not both be in the model at the same time if the results are meant to be interpretable. We also believe that absement—and, by extension, acoustic distinctiveness—provide a characterization of the role that duration plays in the modeling. That is, absement describes what is happening over the duration of the stimulus, and as a result, it more clearly represents speech processing than duration. To draw a more concrete example, consider trying to model the fuel efficiency of a car. It is standard to quantify fuel efficiency as the ratio of distance to volume of gasoline used, such as in miles per gallon or liters per 100 kilometers. However, one could also model the ratio between time spent driving and the amount of gasoline used, which would also index a car's fuel efficiency. The ratio of time to volume of gasoline is related but not equivalent to the ratio of distance driven and volume of gas. Yet, measuring fuel efficiency with time does not capture the crucial relationship between gasoline consumption and speed of travel, where faster speeds use more gasoline and reduce travel time. As such, without appealing to other factors, time spent driving obviously does not afford the same potential for explanation in a model of fuel efficiency as the actual distance driven does. The same holds for the relationship between stimulus duration and absement/acoustic distinctiveness: The time it takes to hear a word does not give the same amount of information regarding perception as the accumulated acoustic differences between a word and other words in a language.

Theoretically, the general relationship between phonological neighborhood density and acoustic distinctiveness is inverse. Where phonological neighborhood density is high, acoustic distinctiveness is low, and vice-versa. The reason for this relationship is that acoustic distinctiveness is a measure of how acoustically unique a word is in the lexicon, whereas phonological neighborhood density is a measure of how similar a word is to other words. This relationship is reflected in the linear correlation value of -0.30 between these two variables in the data used for modeling.

Acoustic distinctiveness values were used as a predictor of response latency in generalized additive mixed models (GAMMs) using the MGCV (v1.8.3, Wood, 2011) and ITSADUG (v2.3, van Rij et al., 2017) packages in the R programming language (v3.6.3, R Core Team, 2020). GAMMs were chosen to model nonlinear relationships between the variables. We feel that modeling possible nonlinear responses is especially important when introducing a new variable. Response time was measured from stimulus offset to help factor stimulus duration out of the response latency values themselves. These response times were then logged. Only correct responses to real words made after stimulus offset were retained. This restriction leaves 96,001 responses for the modeling process.

Model fitting consisted of a forward-fitting process for the random structure, where complexity was gradually added based on the f restricted maximum likelihood score (fREML), as suggested in van Rij et al. (2017). The fixed-effect structure was fit analogously but gradually removing complexity. This backward-fitting process resulted in a smooth term for age, a smooth term for education level, and a parametric term for sex being removed from the model for not contributing to the overall fitness of the model. The final model had fixed smooth terms for trial number, log COCA frequency+1, acoustic distinctiveness, phonological uniqueness point, and log moving average response latency. Phonological uniqueness point is the point at which a word can be uniquely identified from among all other competitors, and it has been found to be predictive of participant behavior in spoken word recognition (Tucker et al., 2019; William Marslen-Wilson & Pienie Zwisterlood, 1989). Log moving average response latency is a decaying average of a participants' previous responses. It was calculated using the algorithm from ten Bosch et al. (2018), with the α variable set to 0.1 globally. Both phonological uniqueness point and log moving average response latency are included in the model as control variables.

The random effect structure consisted solely of random intercepts by subject. Adding random slopes did not significantly improve the model fit. By-item random intercepts were not included in the model because the models took a prohibitively long time and amount of RAM to run. Additionally, most items had four or fewer responses after subsetting, so the explanatory power added by having the by-item random intercepts is small, and the potential for overfitting increases.

The best model from the model-fitting process was then subjected to model criticism, as outlined in Baayen and Milin (2010). There was a left skew in the distribution of the residuals, so the observations associated with residuals that were greater than 2.5 standard deviations from the mean residual value were dropped ($n = 2386$ or 2.49% of the data used for model fitting), and the model was re-fit. The table of coefficients for the fixed smooth terms in this model can be seen in Table 2.1.

Table 2.1: Table of coefficients for the GAMM after model criticism.

Predictor	edf	Ref.df	F	p-value
Trial number	3.32	4.12	23.22	$< .001$
Log COCA frequency+1	5.76	6.72	333.01	$< .001$
Acoustic distinctiveness	5.39	6.59	1154.60	$< .001$
Phonological uniqueness point	5.62	6.53	441.76	$< .001$
Log moving average RT	8.41	8.91	540.72	$< .001$

The smooths for the control variables were as expected. And, a plot of the smooth effect of acoustic distinctiveness can be seen in Figure 2.1. Smooth effect plots for the other effects can be are provided in the supplementary material.[1] The relationship is monotonically decreasing, with the amount of decrease leveling off at the higher values of distinctiveness. That is, words that are acoustically similar

[1] See supplementary materials at [URL will be insert by AIP] for additional smooth effect plots from the first model.

to many words are responded to slower. Analogously, words that are acoustically distinct from many words are responded to faster. In the frame of competition, words with many potential competitors (words that are acoustically similar to many words) are responded to more slowly, and words with few potential competitors (words that are more acoustically distinct) are responded to more quickly. This is the same general trend as reported for phonological neighborhood density, at least for English (Luce & Pisoni, 1998). In terms of speech perception, these results suggest that it takes longer for the competition process to play out in the mind when hearing a word that sounds like many words.

Concurvity was also calculated for this model. The results are reported in Table 2.2. Concurvity is a generalization of collinearity for nonlinear trends (Wood, 2011). Since GAMMs model nonlinear trends, it is appropriate to use concurvity here. The measures of concurvity from MGCV use a similar scale as correlation, where a value of 0 means no concurvity and a value of 1 means indiscernibility from other smooths, though intermediate values cannot necessarily be mapped onto standard correlation thresholds.

Of the three indices that MGCV provides, we choose to interpret the "observed" index. While the documentation suggests that this measure is possibly optimistic (underestimates) about how much concurvity is in the model (Wood, 2020), it is close to the worst-case for concurvity in our data. We also prefer that it measures the concurvity present in the data given the GAMM coefficients that the fitting process determined. For a given smooth term, the index can be thought of as the proportion of its effect that can be explained using other smooth terms. We provide a deeper explanation of these indices in our supplementary materials.[2]

[2] See supplementary materials at [URL will be insert by AIP] for these explanations of the concurvity indices.

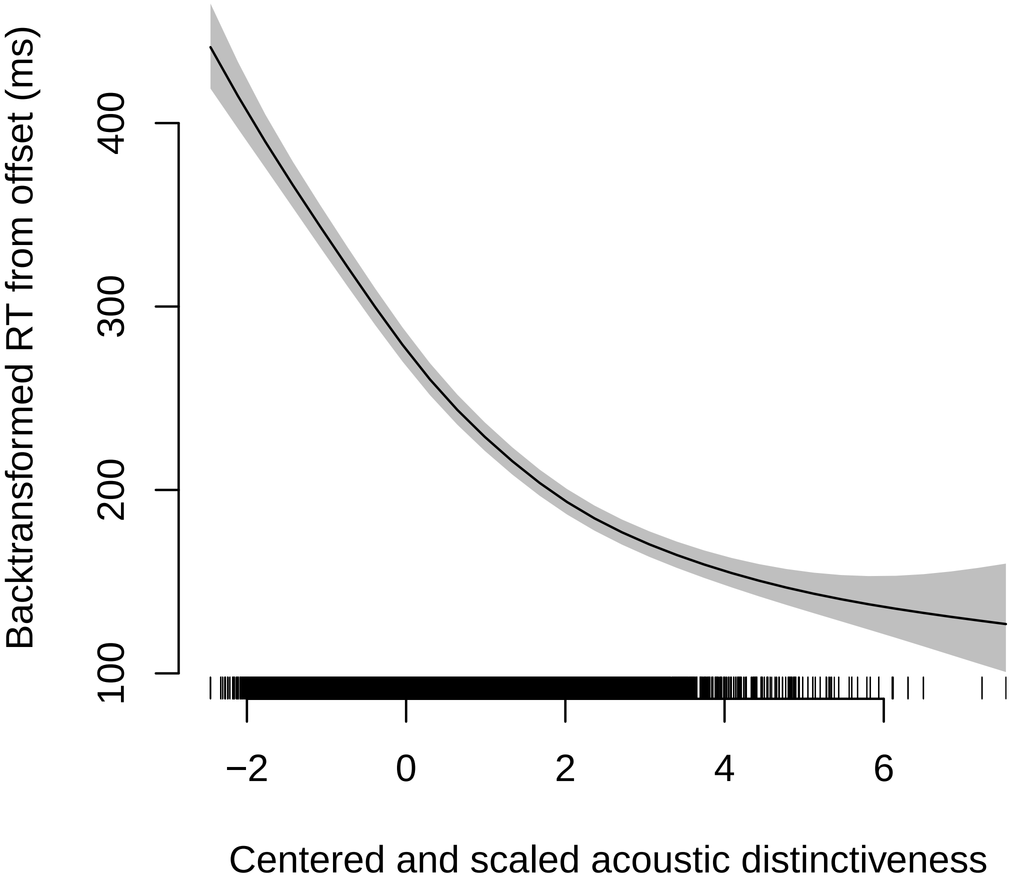


Figure 2.1: Smooth effect for the acoustic distinctiveness value, where all other predictors are held constant. The y-axis is the response latency after backtransforming from log scale. The x-axis is the centered and scaled acoustic distinctiveness. Each point in the function represents how much additional time it would take to respond to a word with that particular acoustic distinctiveness value.

There seems to have yet to be a calibration of these measures of concurvity against different kinds of statistical errors, nor is there a consensus on when the values begin to become concerning. Johnston et al. (2019) used a provisional cutoff of 0.3 in the indices as an indication of a potential problem regarding which variables to include. For our purposes, however, such a calibration is not strictly necessary because we are not using concurvity measures as a method of determining which variables to include in a model. Rather, we are interested in determining the extent to which an effect, such as phonological neighborhood density, is explained by all the other predictors in the model. In this case, we believe that a cutoff of 0.5 is appropriate. The interpretation of this cutoff is that a concurvity measure above 0.5 suggests that a majority of a predictor's effect can be explained by other terms in the model.

Table 2.2: Estimate concurvity table for smooths in the GAMM model. A value of 0 indicates no concurvity and a value of 1 indicates indiscernability of the effect among other smooths.

	Concurvity index		
Predictor	worst	observed	estimate
Trial number	0.21	0.14	0.10
Log frequency+1	0.16	0.12	0.13
Acoustic distinctiveness	0.31	0.30	0.23
Uniqueness point	0.24	0.21	0.20
Log moving average RT	0.57	0.57	0.47
Subject	1.00	0.24	0.01

There is one predictor for which the measure crosses our threshold, that of log moving average reaction time. It is a control predictor that does not really relate to the research questions, so is not really a concern for the interpretation of acoustic distinctiveness in the model. Still, an examination of the pairwise measures of concurvity from the CONCURVITY function shows that much of the high concurvity value is due to the random effect for subject, where the value of the observed index was 0.49. The concurvity between log moving average reaction time and the random effect for subject is to be expected, though, since log moving average reaction time is calculated on a by-subject basis.

There are some implications for speech processing to be gleaned from the effect of acoustic distinctiveness in the model presented here. First, it would seem that competition effects can be modeled using data directly derived from physical measurements of the acoustic signal. The MFCC templates used for calculating acoustic distinctiveness are based on the acoustic production of the speaker, and each coefficient in each frame of the template indicates frequency information. If competition were to first arise at an abstract, symbolic level—like that of phonemes—acoustic distinctiveness should not have had a great effect in modeling the response latencies because it would not connect directly to the cognitive information that is producing the competition effect. However, since acoustic distinctiveness produced a competition-style effect, its effect in this model challenges the idea that word-level competition plays out among candidates represented as symbol strings (e.g., phonemes or diphones) and not acoustics, such as suggested by the networks in TRACE (McClelland & Elman, 1986) and TISK (You & Magnuson, 2018).

Overall, these results show that calculating acoustic distinctiveness by comparing sequences of MFCC values produces a useful predictor for response latencies in the auditory lexical decision task. Due to its high correlation (and, likely, high concurvity) with item duration, acoustic distinctiveness may account for roughly the same portion of variance in the data that duration does. However, acoustic distinctiveness has a clearer relationship to the signal and other items in the lexicon than does duration. This is a particularly important point because duration is often included in models as a control predictor for nuisance variance, while that same variance can be more easily related to competition when using acoustic distinctiveness as a predictor. Additionally, in our data, phonological neighborhood density is more correlated with duration ($r = -0.46$) than with acoustic distinctiveness ($r = -0.30$). From a modeling perspective, the effects of lexical predictors in the model may be more easily interpreted when using acoustic distinctiveness than duration due to lower amounts of multicollinearity or concurvity. Acoustic distinctiveness may thus be preferable over duration in this scenario.

However, there is a potential shortcoming of using the stimuli themselves as the template against which the stimuli are compared to find their acoustic distinctiveness. Namely, it is not very ecological to the prior experience of a listener. Regardless of what the structure of the lexicon may be or what the mechanisms of speech processing are, an adult listener will have experience with a wide variety of speakers. New stimuli will be compared against this sum total experience, rather than just the experience relating to the speaker themselves. As such, the next analysis focuses on comparing templates created from different and multiple speakers and assessing how well they match the lexical decision data, with attention also paid to how they compare to phonological neighborhood density.

2.2.2 Analysis 2

To answer the question of how using different and multiple speakers to create the templates for calculating acoustic distinctiveness and how these compare to neighborhood density, acoustic distinctiveness values were calculated similarly to those in Analysis 1. This time, additional speakers' recordings were used. These were the previously mentioned young female and older male speakers. Both of these speakers' recordings were used as template sets for determining the acoustic distinctiveness of the stimuli used in the lexical decision task. Additionally, values were calculated using each possible combination of speakers as templates by using a sequence averaging technique of the words. Each of these instantiations of acoustic distinctiveness was also compared against phonological neighborhood density. The motivating hypotheses were that (1) if the acoustic representation is abstracted enough away from the raw signal, using a different speaker's recordings as the templates should also provide an indication of lexical competition, and (2) that since a listener has multiple experiences with a given word's acoustic characteristics, using an average of multiple speakers' recordings should produce a template that is closer to a listener's cognitive representation, providing a better index than a single speaker would. The different templates compared were all possible subsets of the three speakers: (1) the young male speaker, (2) the young female speaker, (3) the older male speaker, (4) the average of the young male speaker and the young female speaker (5) the average of the young male speaker and the older male speaker, (6) the average of the young female speaker and the older male speaker, (7) and the average of all three speakers.

Calculating average sequences

The averaging process comes from Petitjean et al. (2011) and Petitjean et al. (2014), which was designed for time series data generally. We started with MFCC-by-time representations as described before. Next, the medoid of the sequence was found. The medoid is a central tendency—like the mean and median—for a set of data. It is the element in the data set which is closest to all the other elements in the set, given a cost function. In this case, the absement between sequences (dynamic time warping cost) was used as the cost function to minimize. Here, the medoid is found by computing all pairwise absement values and choosing the recording that has the lowest summed absement to the other recordings.

The medoid is taken to be the time series that will be modified to find the average sequence. Subsequently, the medoid is mapped onto each time-series with dynamic time warping. In doing so, each frame of the current average sequence is mapped onto relevant frames in the other time series. Each frame in the current average sequence is then replaced with the average (or barycenter) of all the frames mapped to it from dynamic time warping. The process is repeated iteratively until a convergence criterion is met, and the resulting sequence is taken as the average. This process was carried out using the AVGSEQ function in the PHONETICS.JL package.

Conceptually, this averaging process is similar to Kirchner et al. (2010), who also used dynamic time warping to create a type of average of exemplars, though the algorithm and representation were different.

Statistical analysis

To compare the effect of each of the different methods of calculating the acoustic distinctiveness had on the model, the same model from Analysis 1 without the acoustic distinctiveness variable was taken as a baseline model. The acoustic distinctiveness values from different calculation methods were then added to the model separately, and the changes in the fREML values were observed. The change was also observed for adding phonological neighborhood density. When comparing to the baseline model, there was a decrease in fREML for each method used to calculate acoustic distinctiveness, as well as for phonological neighborhood density. The magnitudes of these decreases are presented in Figure 2.2. The decreases in fREML support both hypotheses outlined for this analysis. The second hypothesis was not fully supported, though, since using the young male speaker's recordings as the templates produced the greatest increase to model fitness. This is not completely unexpected since his recordings are naturally going to be closer to each other than they are to other speakers' recordings.

By a large margin, neighborhood density provided the least improved model fit when compared to the baseline model. However, based on the fREML value, there is a significant increase in fitness from the baseline model. Generally, all templates that included the speaker of the stimuli for the lexical decision task increased the model fitness the most.

What is more striking is that using the older male speaker's productions as templates to compare the experimental stimuli against does not improve model fitness to the same degree as the other acoustic distinctiveness values. It suggests that older male speaker's speech is not a good model of the younger male speaker's due to the greater acoustic differences. Conversely, the larger increases to model fitness from the other acoustic templates could be taken to indicate more acoustic similarity between the templates and the stimuli. Support for this idea is also found in that using the younger male speaker's recordings as the templates produces the greatest increase to model fitness. These results also suggest that age differences produce greater acoustic differences in production than do sex differences. The results also suggest that acoustic representations based on single speakers run the risk of creating idiosyncratic models of speech that may not effectively capture the important acoustic aspects of words.

Concurvity was also checked for each model, and the results were similar to those in Analysis 1, with the exception that the model that used neighborhood density instead of acoustic distinctiveness, the observed concurvity index for neighborhood density was 0.51. In the pairwise observed concurvity indices, phonological neighborhood density was most concurved with uniqueness point at a value of 0.39 and

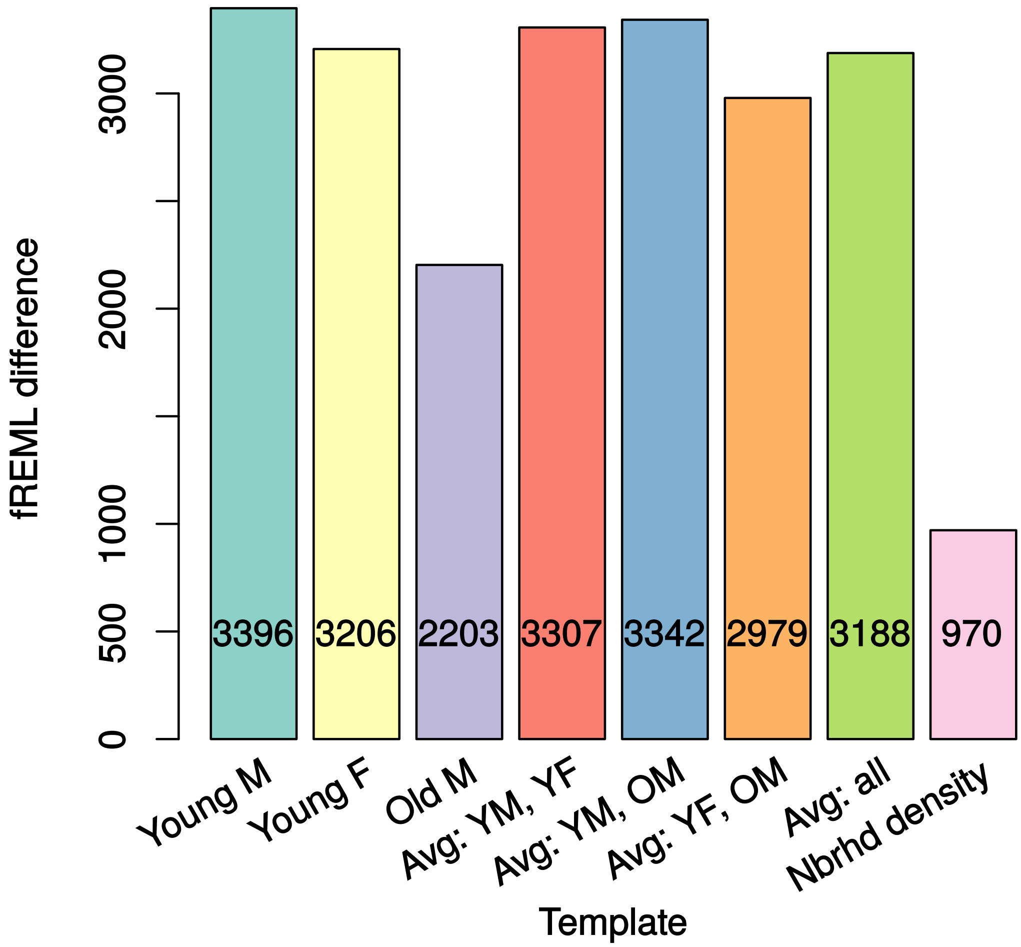


Figure 2.2: (Color online) fREML differences between acoustic distinctiveness calculations and neighborhood density. All the changes were decreases, indicating better model fit. Larger values indicate greater increases to model fitness. “YM” refers to the young male speaker, “YF” refers to the young female speaker, and “OM” refers to the older male speaker.

log frequency at a value of 0.23. Overall, these concurvity results suggests that a slight majority of the smooth for phonological neighborhood can be explained using the other variables in the model. Specifically, a moderate amount of the concurvity in the model is owed to uniqueness point and log frequency.

In the face of these observations, it is clear that acoustic distinctiveness increases model fitness more so than neighborhood density. Overall, this indicates that acoustic distinctiveness is a better predictor of response times in the model. Treating acoustic distinctiveness as an indicator of lexical competition, these results imply that competition is better measured using acoustic representations that are closer to the observed data than phoneme sequences. And, acoustic distinctiveness is closer than phonological neighborhood density to a literal reading of the phrase "sound similarity" that underlies the idea of phonological neighbors, i.e., words that sound similar.

What's more, the results suggest that this measure can be generalized to be used in future research that does not necessarily use the MALD stimuli. Because various speakers or combinations thereof can be used as templates for the stimuli in the experiment without destroying the effect of acoustic distinctiveness, a database could be produced that contains a large number of templates. A researcher could then input their stimuli to a program that would compare the stimuli to the items in the database and provide an acoustic distinctiveness score for the stimuli.

It is still unclear, though, if acoustic distinctiveness values represent the same kind of information as neighborhood density does. To answer this question, a third analysis was carried out that examined the degree to which neighborhood density further increased model fitness for models that already had distinctiveness values as predictors.

2.2.3 Analysis 3

To answer the question of whether acoustic distinctiveness and neighborhood density capture similar information about competition, a third analysis was performed. The motivating hypothesis is that if neighborhood density and acoustic distinctiveness are measuring the same thing and accounting for the same variance in the data, adding neighborhood density to a model that already has acoustic distinctiveness should not significantly increase the model's goodness of fit.

Statistical analysis

Phonological neighborhood density was added to each of the models with acoustic distinctiveness from Analysis 2, and the changes in the fREML values were observed. The fREML decreased for each model, and the magnitude of the decreases are presented in Figure 2.3. Overall, neighborhood density contributed significantly to improving the fitness of all the models, which is taken as evidence against the idea that acoustic distinctiveness and phonological neighborhood density represent closely related information about the lexicon.

Note that the level of fREML decrease (that is, the level of model fitness increase) was greatest for the model using the older male speaker's recordings as the template for acoustic distinctiveness. There is a parallel to the finding in Analysis 2 where using the older male speaker's recordings as the templates increased model fit the least compared to the other acoustic distinctiveness values. Together, these results imply again that using the older male speaker's productions as the templates for the younger male speaker's productions is a worse fit, potentially due to there being greater acoustic differences between the two speakers.

A similar trend to those from Analysis 2 is seen in the concurvity values for the

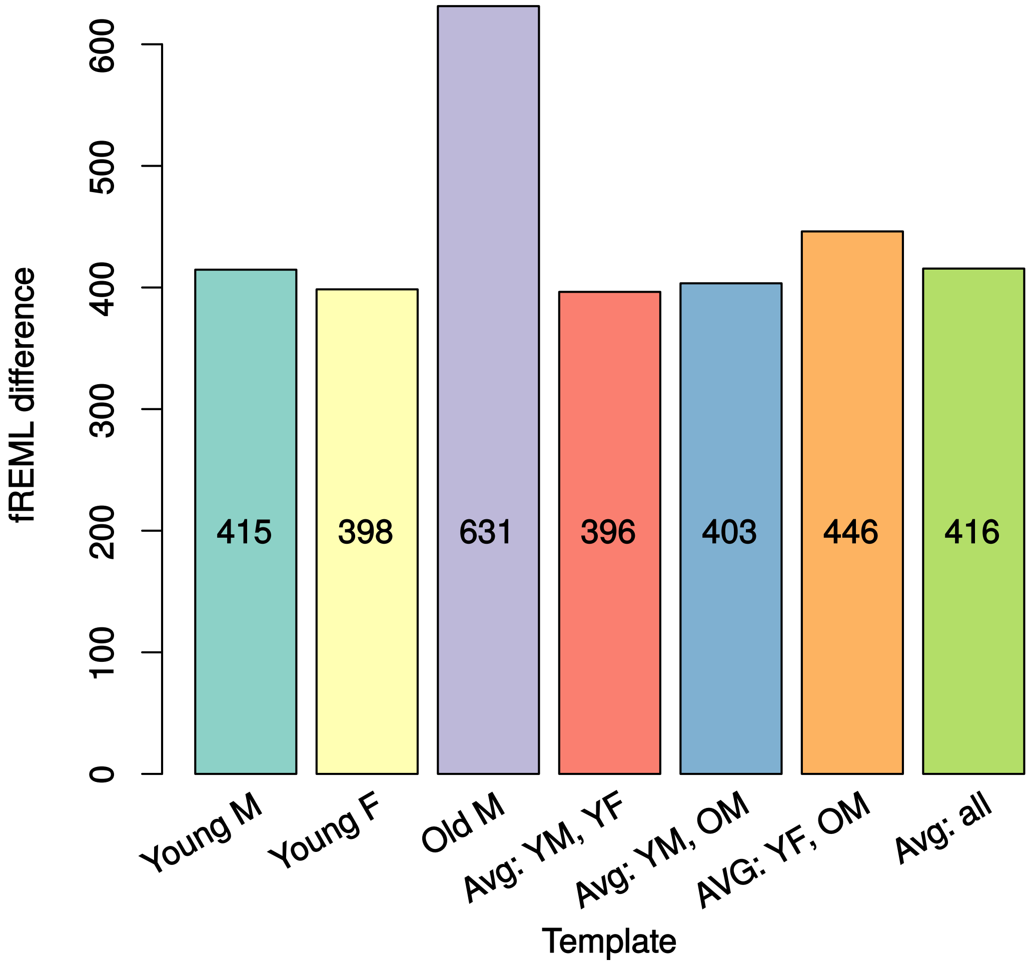


Figure 2.3: (Color online) fREML differences between acoustic distinctiveness calculations and neighborhood density. All the changes were decreases, indicating better model fit. Larger values indicate greater increases to model fitness.

models in the present analysis. The best case for phonological neighborhood density was when it was added to the model using the older male's recordings as templates.

In this case, phonological neighborhood density had an observed concurvity index of 0.53, with uniqueness point, log frequency, and the acoustic distinctiveness values being the greatest contributors in the pairwise comparisons, having values of 0.40, 0.23, and 0.12, respectively. The worst case for phonological neighborhood density was the model using the young female's recordings as templates, where neighborhood density had an observed concurvity index of 0.55, with values of 0.41, 0.23, and 0.21 for uniqueness point, log frequency, and acoustic distinctiveness, respectively. The concurvity values for acoustic distinctiveness were largely similar to those in Analysis 1. For the model with templates from the young female, the observed index on the full model was 0.37, with its largest values in the pairwise comparisons being 0.23, 0.13, and 0.25 for neighborhood density, log frequency, and uniqueness point, respectively.

In sum, the better the acoustic representation contained in the templates matched the stimuli, the more that acoustic distinctiveness explained parts of neighborhood density's effect. Further against the hypothesis motivating this analysis, it may not be possible for acoustic distinctiveness to completely subsume neighborhood density's effect since they appear to be measuring different phenomena, even if there is some overlap. There are at least three possible reasons for this difference. 1) Neighborhood density relies on phonological, phoneme-based representations that are multiple degrees removed from the observed acoustic signal, while acoustic distinctiveness does not. 2) Phonological neighborhood density's reliance on phonemes may cause it to be confounded by the effects of orthography. 3) Phonemic representations may capture some level of abstractness that is not currently captured in the way that acoustic distinctiveness is calculated. The remaining question is whether what remains of neighborhood density's effect in the presence of acoustic distinctiveness is still relevant to sound similarity.

2.3 General discussion

The overall results presented in the current study are that acoustic distinctiveness significantly predicts response latencies in auditory lexical decision, acoustic distinctiveness is more predictive than phonological neighborhood density in statistical models, and there is a degree of overlap both conceptually and statistically between what acoustic distinctiveness and phonological neighborhood density are measuring. The overlap, however, did not seem to rise to the level at which it could be said that neighborhood density and acoustic distinctiveness are measuring the exact same thing. While both measures can be interpreted as some indication of lexical competition, in reality, it should be clear that they are not the same. Acoustic distinctiveness measures an average tendency of how well a given word acoustically matches all words in the lexicon, in the form of absement. Phonological neighborhood density provides an index of approximately how many words there are that sound like a given word based on the one-edit rule.

Looking back to initial investigations using phonological neighborhood density, the focus was on examining the role of the structure of words on lexical competition (Luce, 1986). Structure was taken to be sound patterns, which can have a variety of representations. It could be a sequential string of phoneme-like units, a series of acoustically derived values, the intensity-by-time signal itself, etc. The one-edit rule was seemingly chosen simply as a tool to model lexical competition and not strictly due to theoretical motivations for how words are represented in the mind. As such, it stands to reason that what is important in any index of lexical competition is that it models trends seen in the data. As such, it does not appear that what is understood about lexical competition based on sound similarity is married to phonological neighborhood density itself.

The decision of whether to use phonological neighborhood density or acoustic distinctiveness should be based on the merits of what assumptions the measures make about lexical representation and what trends they can predict. To begin, it is informative that acoustic distinctiveness and phonological neighborhood density do not share a high level of correlation. Were this to be the case, it would suggest they could be operationalizing the same characteristics of words as each other, and are interchangeable for non-theoretical reasons. Rather, replacing phonological neighborhood density with acoustic distinctiveness must be predicated on theoretical reasons. These reasons may be on the basis of representation, in that they concern the nature of lexical representations; applicability, in that one of the measures can explain something another cannot; statistics, in that one of the measures provides a better fit to the data; or feasibility, in that the measure can be calculated easily and efficiently by researchers without being experts in high-performance computing.

Concerning representational reasons, the principal question is how a word is represented in the mind. Phonological neighborhood density relies on an assumption that lexical entries take the form of strings of phonemes. Whereas, acoustic distinctiveness makes an assumption that lexical entries contain some sort of acoustic representation. Inherently, acoustic distinctiveness is less well-defined as a concept because acoustic representations can take many forms. In the context of the present study, the acoustic representations were taken as sequences of MFCC frames, or otherwise sequences of frequency information. A representation based on acoustics is similar in spirit to approaches to phonetic and psycholinguistic analysis that do not coerce the continuous acoustic or articulatory signal to discrete symbols (Baayen et al., 2016; Goldinger & Azuma, 2003; Kohler, 1995; Pike, 1943; Port & Leary, 2005). We are not arguing for or against phonemes or abstraction more generally, but using acoustic absement and acoustic distance may form the basis of describing how sound-level contrast works on an acoustic level.

In spoken word recognition, it is definitional that the acoustic signal itself will come to bear on how words are recognized. The question is whether it is also necessary for discrete symbols like phonemes to be recognized, or if some less abstract, acoustic features suffice for representing words in the lexicon. The averaged MFCC sequences in a word represent a level of abstraction between the raw signal and phoneme strings. Discrete symbols are convenient as a representation for words because they are static. Although, provided a sufficient number of observations are available for any given word, it is likely that the average sequence would converge toward one sequence to represent that word. This average representation would be such that the addition of new observations similar to the representation does little to alter the average sequence if there is nothing particularly novel about the new exemplar. In other words, the sequence is stable and quasi-static. And, this point leads into the question of the applicability to future research since the processes of creating the acoustic specifications associated with acoustic distinctiveness are transparent and can be mapped to explaining a variety of linguistic phenomena.

One such linguistic phenomenon is when a listener adapts to an unfamiliar speaker or accent, the latter of which seems to require rapid updating of cognitive representations or processing (Adank & McQueen, 2007; Clarke & Garrett, 2004). Using the acoustically specified lexical entries, this process can be modeled as adding additional observations to the lexical entries that must be incorporated into the representation. Empirical data could be gathered from a variety of speakers to examine how the representation changes with each new speaker. This process can still be modeled when assuming phonemes as the units of lexical representation, possibly as the listener adjusting the weights they have in the connections they have between acoustic information and phonemes. However, it is unclear how this process might be simulated or modeled effectively when using phonemic strings as the representations for words instead of acoustics. The conclusion in Ohala (1996) highlights some difficulties and potential remedies to finding invariant cues for phonemes such as looking for cues to diphones or looking for different sets of features. But, to date, the constellations of cues that unvaryingly lead to the perception of phonemes are unknown, if such invariant cues exist at all.

An example of where it is not possible to use phonological neighborhood density is the analysis of perception relating to homophones. By definition, homophones will have the same phonemic representation. However, production differences in homophones have been found previously (Gahl, 2008; Lohmann, 2018; Seyfarth et al., 2018; Warner et al., 2004). Warner et al. (2004) also found that listeners are sensitive to these production differences. Any study wishing to examine the perceptual differences of homophones will not be able to use phonological neighborhood density to tease out these perceptual effects, since it will be the same for the homophone pairs. Acoustic distinctiveness, however, has the potential to be used in such studies because it allows for more granular representations of words that can be sensitive to differences in production. It would also be applicable to studies examining the effects of studies on perception, where phonological neighborhood density could not.

Turning now to statistical reasons for using one of phonological neighborhood density and acoustic distinctiveness over the other, the case for acoustic distinctiveness is stronger. The analyses presented in the current study show acoustic distinctiveness to be more predictive than neighborhood density in a variety of different methods of deriving the acoustic representation. Whether using the stimuli themselves that were being presented to the participants, recordings of the same words by different speakers, or averages of recordings, acoustic distinctiveness increased model fit more so than did neighborhood density. Phonological neighborhood density showed moderately concerning concurvity levels over 0.5 in our models, whether acoustic distinctiveness was in them or not. The parts of phonological neighborhood density that were not subsumed by acoustic distinctiveness, lexical frequency, and uniqueness point may not have to do with lexical competition, either. Since phonological neighborhood density uses letter-like units, it is possible that part of the observed effects of phonological neighborhood density is due to the effects of orthography, which has been found to have profound and varied effects on speech perception (Mukai et al., 2018; Perre & Ziegler, 2008; Taft et al., 2008; Ziegler & Ferrand, 1998). Though, demonstrating such a connection would require further research. Nevertheless, our results suggest that using acoustic distinctiveness in place of neighborhood density would reduce the chance of encountering concurvity or collinearity issues during regression modeling.

In terms of feasibility, phonological neighborhood density has some factors in its favor. It is easier to program, especially compared to the average sequencing procedure. Note, however, that the Levenshtein distance used in neighborhood density is a dynamic programming algorithm just like dynamic time warping, so the implementation differences between them are slight. Neighborhood density also uses textual data, which is easier to manipulate and gather, and it takes up less hard drive space. However, some steps can be taken for acoustic distinctiveness to make it more accessible to researchers. It can be incorporated into software packages, like Phonetics.jl, which will give researchers an accessible programmatic interface for calculating it on their stimuli. Additionally, we have made our acoustic absement comparisons and distinctiveness values available in Kelley and Tucker (2021a) for other researchers to be able to use acoustic absement in their own work.

There are, thus, various reasons to favor the use of acoustic distinctiveness over phonological neighborhood density defined using the one-edit rule and Levenshtein distance. Representationally, acoustic representations of lexical items can provide more transparent explanations of phenomena than phonemic representations. In terms of applicability, acoustic distinctiveness seems applicable to a wider variety of experiments performed in phonetic and linguistic research. Statistically, acoustic distinctiveness contributes more to model fitness than phonological neighborhood density and does not seem to have the possibility of being confounded with the effects of orthography. For those reasons, we believe the time has arrived to reconsider quantifying lexical competition with the one-edit rule and phonological neighborhood density. Recent increases in computational power and quantity of data obviate some of the technical reasons to use the one-edit rule on textual representations of words to assess sound similarity. Future research can build upon the concept of absement to measure lexical competition and sound similarity acoustically.

One specific improvement would be to ensure that the acoustic representations can account for the acoustic cues known to be relevant in speech perception. It is also crucial to develop acoustic representations based on more than just three speakers' recordings, especially so as to avoid the problem of using the experimental stimuli themselves in the acoustic template. It will also be necessary to use acoustic distinctiveness and acoustic distance in modeling spoken word recognition in non-English languages. The results presented in the present study are intended to be applicable cross-linguistically, but it cannot be determined whether these results are indeed valid across languages until future experiments are conducted. Finally, alternative representations should be explored, such as those using functional data analysis discussed in Pigoli et al. (2018) or using the encoding that an off-the-shelf automatic speech recognition system has learned. It may also be fruitful to explore the methods used in Kirchner et al. (2010).

2.4 Conclusion

We began the present paper began by discussing the activation/competition metaphor in language comprehension and discussed a common operationalization of competition, phonological neighborhood density. It was observed that acoustic distinctiveness is a stronger predictor of competition effects than phonological neighborhood density is, even if they don't completely account for the same information.

Though competition has often been reasoned about using abstract symbolic forms, acoustic distinctiveness opens the door to reasoning about competition in terms of acoustics. Lexical representations may encode acoustic information itself, rather than acoustics being a mere tool to get to abstract symbols used for representation. Additionally, the sequencing of the onset of competition effects may be earlier than once thought, beginning while acoustic information is being processed, and future models of spoken word recognition will need to be intentional in how they depict the sequencing of processing and competition.

The advent of large databases of speech and more powerful computers has ushered in the possibility of refining the notion of phonological neighborhoods. The initial concerns of Luce (1986) may finally be addressed, and characteristics of acoustic data can play a larger role in understanding the comprehension of spoken language, as well they should.

Acknowledgments

The authors would like to thank the attendees of the 11th International Conference on the Mental Lexicon, the attendees of the winter 2019 University of Alberta Department of Linguistics Generals Paper conference, and the attendees of the 179th

Meeting of the Acoustical Society of America for their thoughts and feedback on earlier versions of this project. We would also like to thank the members of the Alberta Phonetics Laboratory for insightful conversations on the topics of this paper. This project was funded in part by a Social Sciences and Humanities Research Council Grant #435-2014-0678.

Chapter 3

Perception and timing of acoustic distance

Abstract

The notion of acoustic distance figures into many aspects of phonetics such as vowel overlap and phonological neighborhoods. A measurement of word-level acoustic distance useful for cognitive modeling must account for two aspects of perception: listener sensitivity to acoustic differences and the duration discrepancies between different words. The present paper suggests the use of dynamic time warping as a way to measure how acoustic distance accumulates between words over time. The results of a distance rating task with synthesized vowels are used as a basis for selecting a mathematical function that best matches listener sensitivities. Additionally, the results of a reminder task with synthesized vowels are used to determine a just noticeable difference threshold for vowels. The results suggested that a distance function based on the 4.5-norm and using a 30 ms radius for dynamic time warping best matched human behavior. A third analysis used these new dynamic time warping

configurations to model reaction times in an auditory lexical decision task and found that Euclidean distance and no temporal constraints on dynamic time warping best matched human behavior during spoken word recognition. These ultimate results are discussed in relation to models of spoken word recognition, including how to assess the acoustic match between the speech signal and a word in the lexicon based on the perceptual results given here.

3.1 Introduction

There are various cognitive models of spoken word recognition. At a minimum, these models explain spoken word recognition conceptually. Some models, such as the cohort model (W. D. Marslen-Wilson & Welsh, 1978) and the neighborhood activation model (Luce & Pisoni, 1998) stop at this level. Still other models are computational or mathematical in nature and provide the steps necessary to use the model on a computer. Such models include TRACE (McClelland & Elman, 1986), DIANA (ten Bosch, Boves, & Ernestus, 2015; ten Bosch, Boves, Tucker, et al., 2015), the discriminative lexicon (Baayen et al., 2019), TISK (You & Magnuson, 2018), EARSHOT (Magnuson et al., 2020), Shortlist B (Norris & McQueen, 2008), Fine-Tracker (Scharenborg, 2010), and PARSYN—itself a computational implementation of the Neighborhood Activation Model that was referred to but not fully described in Luce et al. (2000) and Vitevitch et al. (1999).

Every model that has been proposed and taken up in the literature has been informed by some sort of behavioral data. These data are often of the sort that come from lexical decision and/or cross-modal priming tasks, among others. Some models also go so far as to incorporate the results of experiments that might be considered more perceptual in nature, of the sort that might be seen in discrimination, identification, and rating tasks. For example, the Neighborhood Activation Model uses experimental phoneme confusion data from a phoneme identification task as part of the frequency-weighted neighborhood probability rule, which Luce (1986) and Luce and Pisoni (1998) suggested has having more explanatory power than phonological neighborhood density. Shortlist B uses confusion data from a gating task as a basis for calculating probabilities. Fine-Tracker was tested against data on human perception of duration as a cue for word boundaries. The features learned in EARSHOT were compared to electrocorticography data, essentially establishing a relationship between the model and neurolinguistic speech perception results. Not every aspect of these models is grounded in perceptual data either, though, nor is it realistic to expect as much. The present study focuses on a specific computational aspect of spoken word recognition, how two words might be compared acoustically. Specifically, it provides data from a distance rating task for synthetic vowels and a reminder task for vowel duration that inform the process of acoustically comparing words with dynamic time warping in a cognitively informed way.

Peculiarly, there is a lack of specificity regarding what would seem to be a crucial process that goes on in the Neighborhood Activation Model. Specifically, Luce and Pisoni (1998) posited that phonetic or acoustic (mis)match between the acoustic signal and items in the lexicon mediates the accumulation of activation for a candidate during spoken word recognition. Generally, a strong match between the signal and an item would induce high activation, and vice-versa. There is not much more detail on acoustic or phonetic matching given in the model description, however. The level of similarity also figures into the definition of “phonological neighbors” as words that sound similar. Yet, the degree of acoustico-phonetic matching is assessed using

Levenshtein distance on phoneme strings and termed "sound similarity." Despite invoking the concepts of sound and acoustics, textual evaluation of phoneme strings is a decidedly non-acoustic way to assess sound similarity. There is, thus, a mismatch in the description given of lexical activation and competition in the Neighborhood Activation model and the quantitative methods employed. Vitevitch and Luce (2016) noted that Levenshtein distance is but one way to assess sound similarity between words, though it remains the most common.

More computationally explicit models perform some degree of acoustico-lexical matching, but it is difficult to compare words to each other in this sort of framework. Consider DIANA, in which words are specified as a sequence of sub-phone states (ten Bosch, Boves, Tucker, et al., 2015). The sub-phone states of [p] as in *peep*, for example, might be roughly modeled as 1) beginning of [p], 2) middle of [p], and 3) end of [p]. The acoustic signal is compared against these sub-phone states with Gaussian mixture models providing probabilities of belonging to a particular sub-phone class, while a hidden Markov model assigns a probability to transitioning to a new state or continuing in the current state. The speech signal is thus compared with the segmental symbol sequence of a word in the lexicon more probabilistically than acoustically. It can likely be inferred that words with similar probabilistic activation values in DIANA are acoustically similar if lexical frequency is ignored. However, sub-phone states have not yet been demonstrated to be acoustically separable with machine learning techniques. In a similar acoustic model in Graves (2012, Chapter 6), for example, approximately 1 in 3 phones were identified incorrectly, suggesting that the sub-phone states were not completely acoustically distinguishable from each other. It is thus somewhat unclear how well the Gaussian mixtures model acoustic differences if they are modeling acoustic categories that they can't separate well.

To address the apparent tension in the description of spoken word recognition and methods used to implement those descriptions, Kelley and Tucker (2021c) proposed using the results of dynamic time warping to assess acoustic absement between recordings of words. Acoustic absement is the accumulation of acoustic distance over time. That is, when comparing two recordings, one can find the acoustic distance between two discrete time points. When the distances computed from these discrete time points between the two recordings are summed over time, the quantity produced is absement. Using dynamic time warping provides an algorithmic-level description of the assessment of sound similarity and how acoustic mismatch (or distance) between two words may be accumulated over time. That is, using dynamic time warping quantifies the acoustico-phonetic dissimilarity between two words more appropriately than the Levenshtein distance method previously described. Because dynamic time warping directly compares acoustic information between two sound sequences, the results are more clearly interpretable acoustically than the probabilistic results that might come from a model that is using hidden Markov models like DIANA.

Dynamic time warping has its origins in speech technology and was in part designed to account for temporal variation between different exemplars of the same item in speech. It dates back at least to Sakoe and Chiba (1970, as cited in Sakoe & Chiba, 1978). It was used to calculate how different a recording of a given word was from a bank of recorded words, for the purpose of automatic speech recognition. It produces a nonlinear alignment between two sequences such that the overall dissimilarity between the two signals is minimized. The alignment process is very similar to that of forced alignment, but instead of mapping phones onto chunks of the speech signal, small spectral or cepstral chunks are mapped onto each other between two speech signals. The net result is that similar regions of the signals are compared with each other. For example, small temporal variation in the production of a vowel between two utterances of *heed* [hid] would not cause the algorithm to end up comparing [d] or [h] in one recording to [i] in the other.

An example alignment for two hypothetical recordings of *heed* is presented in Figure 3.1. Each tick on an axis represents a time step associated with a particular phone label. Note that the recording represented on the x-axis has a shorter [h] than the second recording, but it has a longer [i]. The orange line represents the comparisons that would happen if the sequences were aligned linearly. Note how at (3, 3) the [i] in the first recording is being compared to the [h] in the second recording, which is not ideal. The optimal alignment path identified by dynamic time warping avoids this problem so that, between the two recordings [h] is compared with [h], [i] is compared with [i], and [d] is compared with [d]. Note also that a linear alignment is impossible if the sounds are of different lengths, which happens often with phonetic data, especially when comparing recordings of different words. The present example is simplified, but it captures the general behavior of the alignment in the algorithm.

The output of the dynamic time warping algorithm comprises the alignment path between the two sounds and the dissimilarity score that comes from comparing the aligned portions of the sounds. There are two components of dynamic time warping that, when modified, will have an obvious impact on the calculation of acoustic absement and are related to human perception. The first component is how distance is calculated between time steps. The second component is how long a time step is allowed to be warped in the warping path.

The motivation for investigating the second component is as follows. An acoustic representation of a word should roughly approximate the characteristic sound patterns of the word. This is especially true when the acoustic representation is a mean over multiple exemplars of a word, as in Kelley and Tucker. It is reasonable to expect that these acoustic regions of a word have some degree of time-boundedness,

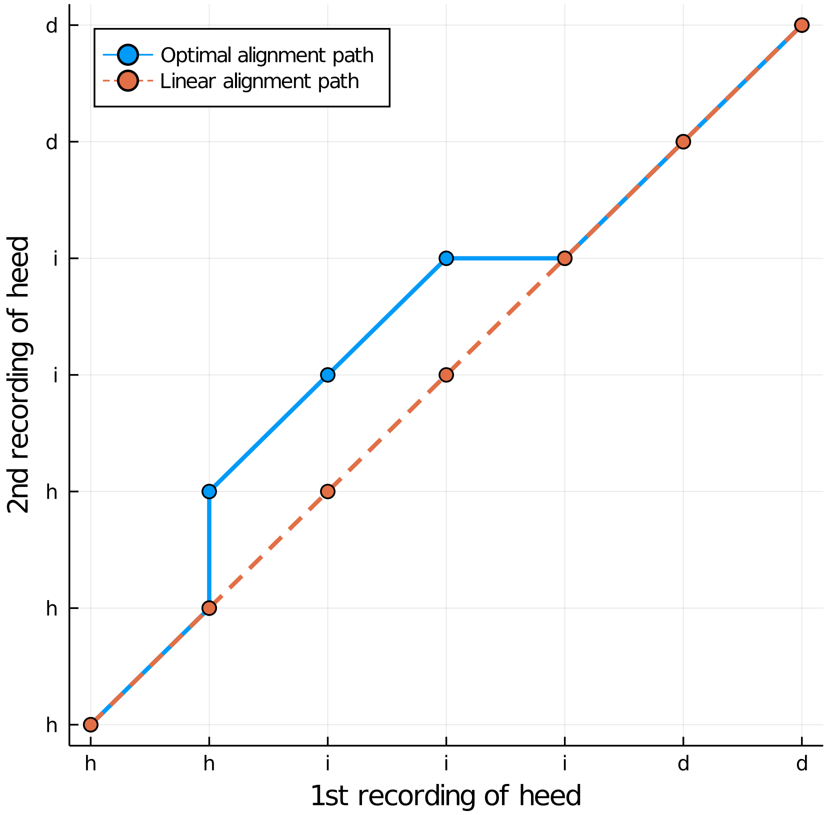


Figure 3.1: Example alignment of two recordings of *heed* being compared with dynamic time warping.

that is, that their temporal location matters. Indeed, the time-boundedness must be true at some level because *peak* /pik/ and *keep* /kip/ are not the same word. In fact, it must be true at the subphonemic level because temporal aspects like voice onset time (Lisker & Abramson, 1964) and vowel inherent spectral change (Nearey & Assmann, 1986) themselves have a time-bound nature, let alone the formant patterns of diphthongs. As such, the comparisons that dynamic time warping performs should also be time-bound in such a way as to better reflect production and perception. That is, each acoustic frame of a word should only be comparable to a small number of frames in another word, and comparisons between frames that are hundreds of milliseconds apart should be disallowed. The question is to determine what the constraint should be on dynamic time warping to reasonably limit how many previous or following frames a given acoustic frame should be allowed to be compared with.

There have been a number of studies on the psychological perception of auditory distance. Kawahara and Matsui (2003) referred to perceptual distance while developing a speech resynthesis tool. But, they did not experimentally investigate whether their resynthesis technique reflects human judgments. However, Terasawa et al. (2005) focused on non-speech sounds and designed a measure of the perception of “timbre,” which they define as any acoustic perception that is not related to pitch and loudness. They validated their measure with behavioral experiments to see if participant behavior matched the measure they designed, using a 10-point Likert scale rating of dissimilarity between stimulus pairs. They found that a perceptual timbre space based on Mel frequency cepstral coefficients (MFCCs) with comparisons based on Euclidean distance was a good match to human behavior. However, it is unclear what sort of stimuli they used, except that they were described as an additive synthesis of sine waves that were voice-like and had pitch vibrato. It cannot be assumed *a priori* that these results will extend to speech.

There has been some work looking directly at judgments of similarity between words as well. Vitz and Winkler (1973) correlated Likert-scale participant ratings of word pair similarity with what they define as the “predicted phonemic distance.” The predicted phonemic distance is a proportion between how many phoneme mismatches there are between the aligned phoneme strings and how long the aligned sequences are. Such a measurement does not actually capture the nature of the perceptual task that humans perform when hearing speech, but rather it attempts to capture some sort of linguistic processing that putatively occurs once the continuous speech signal has been discretized to abstract symbols. Similarly, W. Marslen-Wilson et al. (1996) examined how phonological distance affects participant responses in behavioral tasks. Phonological distance is calculated in terms of how many distinctive features differ between the initial segments of word pairs. No assessment of human perceptions of distance was performed, however.

While both Vitz and Winkler (1973) and W. Marslen-Wilson et al. (1996) studied similarity between words, they lacked mathematical particulars for how the acoustic signals should be compared to each other. While phonemes have some basis in acoustics, even sophisticated deep learning models of speech recognition that perform well on word recognition (e.g., Graves & Jaitly, 2014; Zeghidour et al., 2018) cannot with high accuracy map the acoustic signal to even a subset of phonemes that collapse acoustically similar segments. Such systems top out around 70 or 80%, whether they are trying to classify small sections of speech as belonging to a given phoneme (as in Graves & Schmidhuber, 2005) or merely create a phonemic transcription (as in Zhang et al., 2016). As such, it seems implausible that complete descriptions of human perceptual judgments of acoustic distance can be accomplished looking only at the phoneme level. There simply does not appear to be a strong enough relationship between phonemes and the acoustic signal. Certainly, though, automatic speech recognition systems perform much better when word-level language models are combined with phone recognition models (as in Graves & Jaitly, 2014). However, word-level language models, such as n-gram models, are based purely on frequency of occurrence and not acoustics. Thus, they should not come to bear on judgments of acoustic distance, though they may be important for judging some form of lexical distance between words.

Regarding perceptions of duration, Hirsh (1959) reported a minimum interval of approximately 20 ms between two sounds before listeners could accurately recall the order in which two stimuli were presented, perhaps suggestive of the temporal resolving power of the auditory processing system. This result and claim about temporal resolving power were reified in various studies like R. E. Pastore et al. (1977). However, Kewley-Port et al. (1988) examined perception of voice onset time in a /ba/-/pa/ continuum and found that participants could perceive differences down to 10 ms depending on the task. They also found a 50% crossover threshold of approximately 34 ms for listeners in a labeling task in a quiet environment. For vowels, Tomaschek et al. (2011) found a range of just noticeable difference values of approximately 43 ms to 65 ms when comparing German /a/ and /aː/ for native German listeners, depending on the duration of the stimuli being compared. The lowest just noticeable difference was achieved near the category boundary for /a/ and /aː/, suggesting that sensitivity to duration differences changes as category boundaries are approached. Note, however, that German has a phonological distinction between /a/ and /aː/, so phonological perception is necessarily implicated in the results.

Porretta and Tucker (2013) found that native English listeners could perceive phonetic differences in duration using a speeded AX discrimination task. The probability of discriminating the difference increased as the duration difference between consonant stimuli increased, and especially so when the instructions indicated that listeners would be hearing duration contrasts. Due to the variability in the perception of duration differences based on experimental task, task instructions, speech sound category, and native language, it seems reasonable to assume that determining some sort of parameter to limit the temporal extent of warping in dynamic time warping will be more heuristic than an exact representation of human auditory processing. Still, a heuristic will likely be more useful when calculating acoustic absement than having no explicit constraint at all.

The remainder of the present paper presents two experiments. The first experiment is designed to determine how acoustic distance should be calculated within dynamic time warping. The experiment employs a rating task similar to Terasawa et al. (2005). The second experiment is designed to determine a reasonable warping allowance for dynamic time warping when comparing words. The results of these experiments are then evaluated by calculating a predictor for lexical competition like in Kelley and Tucker (2021c). These calculations involve different configurations of dynamic time warping based on the experimental results. Then, the different configurations are evaluated based on how well the predictor of lexical competition relates to auditory lexical decision response times.

3.2 Experiment 1

The goal of Experiment 1 was to determine what sort of distance function would fit human judgments of acoustic distance the best. There are an infinite number of possible ways to calculate mathematical distance, so searching all of them is impossible. One particular class of distance functions that can be easily searched over is the one formed from the p-norms since only the value of p needs to be manipulated. The general formula for the distance function d_p for which different values of p would be tested is given in Equation 3.1. The vectors x and y are of length n, χ_i and ψ_i are the i-th elements in x and y, respectively, and $|\cdot|$ is the absolute value function.

$$d_p(x, y) = \left(\sum_{i=1}^{n} |\chi_i - \psi_i|^p \right)^{\frac{1}{p}} \tag{3.1}$$

This formula is mathematically identical to the formula for spectral distance in (Lindblom, 1978). Note how d_p becomes familiar distance function at specific values of p. When $p = 1$, it becomes Manhattan distance or summed absolute distance, which has been used in other areas, for example, formant tracking (Barreda, 2021). The formula is shown in Equation 3.2.

$$d_1(x, y) = \left(\sum_{i=1}^{n} |\chi_i - \psi_i|^1 \right)^{\frac{1}{1}} = \sum_{i=1}^{n} |\chi_i - \psi_i| \tag{3.2}$$

Similarly, when $p = 2$, d_p becomes Euclidean distance, which has featured prominently in, for example, analyses of vowel overlap (Fridland & Kendall, 2017; Kendall & Fridland, 2012; Nycz & Hall-Lew, 2015). The formula is shown in Equation 3.3.

$$d_2(x, y) = \left(\sum_{i=1}^{n} |\chi_i - \psi_i|^2 \right)^{\frac{1}{2}} = \sqrt{\sum_{i=1}^{n} (\chi_i - \psi_i)^2} \tag{3.3}$$

It is then relatively simple to iterate over a range of values of p to find which of these possible distance formulas is closest to matching human judgments. To have human judgments to compare against, an acoustic distance rating task was performed on synthetic stimuli. Due to the equivalence of vector norms (consult van de Geijn & Myers, 2020), they will all have a strong correlation with the data if any of them have a strong correlation, but it may still be important for cognitive modeling to determine how much weight to apply to outlying acoustic difference values in individual dimensions of an MFCC vector.

3.2.1 Methods

In this experiment, participants were asked to perform a distance rating task on synthetic vowels. The ratings were used as the independent variables for both a correlation analysis and a regression analysis of the different distance functions.

Participants

Participants were recruited from the University of Alberta Department of Linguistics subject pool, which consisted of undergraduate students enrolled in introductory linguistics classes. Participants were compensated for their time with credit for the course they were enrolled in. At the beginning of the experimental session, participants took a demographic questionnaire. One hundred sixteen participants took part in the experiment, and 84 of these participants' data were successfully uploaded to the servers. 39 of these participants were determined to be monolingual native speakers of English.

As part of the demographic questionnaire, participants were asked if they had been diagnosed with hearing loss. Also during the questionnaire, participants were informed that it was strongly preferred that they participate in a quiet room and that they wear wired headphones. They responded to a yes/no question with whether or not they were in a quiet environment and whether they were wearing wired headphones. One participant was excluded for reporting being diagnosed with hearing loss, and one more participant was excluded for reporting not participating in a quiet room. Thirty-seven participants remained to be analyzed.

Of these 37 participants, 34 participants reported their gender as "female", 2 reported their gender as "male", and 1 participant reported their gender as "nonbinary". The demographic question asking about gender was a freeform text answer, so the reported genders were based on participants' interpretation of the word "gender" in the question. The average age was 19.62 years old ($SD = 1.53$).

Materials

The experimental stimuli were synthesized based on vowel formant measurements from the Hillenbrand et al. (1995) data set. The data set was chosen because the formant values were hand-verified and provided in a format that made subsequent modeling easier. Formant values were modeled as a multivariate Gaussian distribution, using the mean F1 and F2 measurements of the male speakers in the data set. The vowel categories used were all the monophthongs: [i], [u], [ʌ], [ɔ], [ʊ], [ɛ], [æ], [ɑ], and [ɪ]. The mean vector of the distribution was calculated as the mean of each formant for a given vowel category, and the covariance matrix was calculated as the covariance of the F1 and F2 values for a given vowel category. The pitch was determined by finding the medoid male speaker based on each speaker's F1 and F2, using Euclidean distance to compare each speaker. The medoid speaker is the one labeled as "m07" in the data, and his average pitch of approximately 99 Hz was selected as the pitch for all of the synthetic vowels.

For each vowel category, a series of paired stimuli were synthesized using `Praat` v6.1.27 (Boersma & Weenink, 2020). Both stimuli in the pair were randomly sampled from the category distribution. Each formant value then had noise added to it from a Gaussian distribution with a mean of 200 Hz and a standard deviation of 50 Hz. These parameters were determined through manual search to balance not deviating too far from a vowel category's center with not generating many formant values that would result in nearly indistinguishable vowels syntheses. Thirty-three pairs of vowels were created for each vowel category (9 vowels), for a total of 297 pairs of stimuli. The mean and standard deviation for each nominal stimulus category can be found in Table 3.2. The category labels only indicate which distribution the vowel parameters were initially generated from; the stimuli themselves are not guaranteed to still have that vowel quality. The subscript after the category label indicates whether it was the first or second sound in the stimulus pairings. While standard deviation is not the most informative statistic of spread in this case since the formant values were generated from multivariate distributions using the covariance of the formants. The covariance, however, is unwieldy to represent in a table. A table of all the generated formant values can be found in the supplementary materials accompanying the present paper.

These formant values were used to create the synthetic vowels in Praat by using the `Create KlattGrid from vowel` function. The F1 and F2 values varied as determined by the randomly sampled values, and the other values were held constant as can be seen in Table 3.2. These "KlattGrid" vowels were then converted to "Sound" objects with a sampling rate of 44,100 Hz. The `Scale intensity` function was applied to each synthesized vowel individually, with a value of 70 dB SPL (corresponding to the default interpretation of dB SPL in `Praat`). The vowel pairs were then spliced together with a 500 ms period of silence between them before being saved.

Procedure

The synthesized vowels were used to create a rating task. The COVID-19 pandemic necessitated the use of online experimental procedures, rather than in-person procedures. Running speech perception experiments online introduces more confounds and variability in the results due to a less controlled listening environment. Some of these confounds can be somewhat controlled for statistically using demographic questions, but they cannot be completely controlled for. The experiment was created using the `jsPsych` framework v6.1.0 (de Leeuw, 2015), which is used for running online

Table 3.1: Formant-wise means and standard deviations of each nominal stimulus category.

Category	Mean F1 (Hz)	Mean F2 (Hz)	SD F1 (Hz)	SD F2 (Hz)
i$_1$	538.22	2486.20	62.78	125.13
i$_2$	548.17	2529.29	45.66	118.76
u$_1$	569.24	1159.86	63.82	120.19
u$_2$	589.46	1186.45	77.59	107.50
ʌ$_1$	829.06	1383.46	57.36	70.74
ʌ$_2$	829.10	1369.29	61.94	101.23
ɔ$_1$	878.29	1238.36	57.10	83.41
ɔ$_2$	863.07	1125.58	51.62	95.82
ʊ$_1$	660.94	1330.76	68.54	85.62
ʊ$_2$	662.49	1315.02	59.57	82.62
ɛ$_1$	782.16	1995.74	69.51	133.27
ɛ$_2$	785.03	2026.62	63.51	101.14
æ$_1$	789.91	2139.28	59.05	129.34
æ$_2$	795.47	2128.45	80.73	184.48
ɑ$_1$	970.07	1494.43	77.78	12.05
ɑ$_2$	945.23	1521.43	80.87	111.48
ɪ$_1$	621.78	2226.85	49.12	109.24
ɪ$_2$	647.70	2187.97	65.97	124.67

psychological experiments. Experiments created in `jsPsych` are stored as webpages that a participant can access with a web browser. The experiment is run locally in the participant's web browser using the participant's hardware, and the results are then uploaded to a server awaiting file uploads.

The experiment was structured so that participants would listen to a block of 50 stimuli and then receive a prompt to take a break. The order of presentation for the

Table 3.2: Constant values for the `Create KlattGrid from vowel` function.

Parameter	Value
Duration	0.5 s
Pitch	99 Hz
B1	50 Hz
B2	50 Hz
F3	3000 Hz
B3	100 Hz
F4	4000 Hz
Bandwidth fraction	0.05
Formant frequency interval	1000 Hz

stimuli was randomized. It was up to each participant to determine whether or not they wanted to take a break, and if so, for how long. In each trial, a fixation cross was displayed for 500 ms. Then, the audio file containing the pair of sounds to be rated was played. Subsequently, the participant was asked to rate how acoustically different the two sounds they heard were. The rating scale ranged from 1 to 7, where a value of 1 indicated that the sounds were the same, and a value of 7 indicated that the sounds were very different. During the rating portion of the trial, the participants were textually reminded of the scale, and they responded using the number keys on their keyboard. There was no time limit for response.

Participants were instructed to use Google Chrome for the experiment because it was found to work most reliably with the online experiment and file uploading programs. After the demographic questionnaire and before the experiment began, a single synthetic vowel was played that was also scaled to the same intensity as the experimental stimuli so that participants could adjust their volume to a comfortable listening level.

At the beginning of the experiment, participants were told that the results from this task were going to be used to help evaluate a new speech synthesizer. During the experiment debrief, they were informed of the true purpose of the experiment, to better understand human judgments of acoustic distance. The deception was employed to help prevent the participants from hyper-focusing on the purpose of the task while participating. It also helped set up the participants' expectations for the stimuli since synthetic speech was used.

Thirteen participants (35.13%) reported not wearing wired headphones during the experiment. To determine the extent to which using headphones or not affected the participants' responses, the overlapping coefficient (Inman, 1984) was calculated in both a parametric and non-parametric manner. The overlapping coefficient gives an indication of how similar the underlying distributions of two groups are. Previously, it has been applied to analyze acoustic vowel overlap (Kelley & Tucker, 2020). For the parametric version, sample Gaussian distributions were determined for both the headphone-wearing and non-headphone-wearing groups with each group's sample mean and standard deviation of their rating.

Because it will be the by-item mean ratings that will be analyzed, these were the values for which the overlapping coefficient was determined. The mean of the by-item mean ratings for the headphone-wearing group was 3.02 with a standard deviation of 1.74. The mean of the by-item mean ratings for the non-headphone-wearing group was 2.89 with a standard deviation of 1.65. The overlapping coefficient between these two distributions was determined via integration using the built-in `integrate` function in the `R` programming language. The value was 0.96 with an absolute error less than 6.9×10^{-5}. The non-parametric version was calculated using the `overlapping` package (M. Pastore & Calcagnì, 2019), the result of which was a value of 0.88. These results indicate that the underlying distributions are remarkably similar. As well, a Welch two-sample t-test indicated that there was no significant difference between by-item mean ratings for each group ($t = 1.52$, $p = 0.13$). It stands to reason, then, that the differences between the groups are neither statistically nor practically different to a level of significance. As such, both of these groups will be analyzed together. Were there more participants who had reported hearing loss or not participating in a quiet room, a similar analysis could have been performed to determine whether it would impact the result to have those participants' data in the mix, but there were not enough participants who reported hearing loss or not participating in a quiet room.

Finally, the data were subset to remove implausible responses. While there was no time limit for responses, some participants reported being distracted or falling asleep during the experiment. An upper cutoff of 5 s is between the 0.98 and 0.99 quantiles, so only responses rendered in 5 seconds or less were included in the analysis, resulting in a loss of 129 responses (1.17 %). There was a mean of 36.57 responses per item, with a standard deviation of 0.66. The median rating over all stimuli was 3 ($IQR = 2$).

Analysis

There were two analyses performed in sequence. The first was to determine what the optimal value of p was for the p-norm distance function. The general procedure was to search through possible values of p and determine how well the resultant distance function correlated with the participant responses.

The participant judgments were mean-pooled over the stimuli to make the search for the value of p easier, and optimizing correlation should roughly produce a result that minimizes the distance to the mean values anyway. MFCCs were calculated for each stimulus using the `MFCC.jl` (v0.3.1, van Leeuwen, 2019) package in the Julia programming language (v1.5.3 Bezanson et al., 2017), using a window size equal to the stimulus length. The result was that a single MFCC vector was calculated for each stimulus. The motivation for this choice was that frequencies in each stimulus were approximately static over time since they were just vowels with static formant values, so there was no need to look at changes in frequency over small windows of the signal (as is done with spectrograms).

Then, a discretized linear search was performed over the possible values of p. The search ranged from 1 up to and including 100, moving up in increments of 0.01. One hundred was chosen as an upper limit because it produces results remarkably close to the infinity-norm, which itself just selects the maximum value in a vector as the norm. In terms of distance, the infinity norm would select the largest absolute difference between the elements being compared. At each step in the search, the distance function from the p-norm in question was calculated for each stimulus pair in the experiment, and the output was correlated with the mean participating ratings.

As pointed out previously, a current practice in evaluating the acoustic distance between vowels is to calculate the Euclidean distance between formant values (Fridland & Kendall, 2017; Kendall & Fridland, 2012; Nycz & Hall-Lew, 2015). To ground the p-norm on MFCC results against that practice, the second analysis compared the results of the best-performing distance functions to Euclidean distance on formants. The comparisons were evaluated both in terms of correlations and linear mixed-effect regressions (LMERs) fit with the `lme4` (v1.1-26, Bates et al., 2015) package in the `R` (v3.5.1, R Core Team, 2018) programming language. The base model had fixed effects for trial number, age, gender, education, and vowel category. The base model also had a random intercept for subject and another random intercept for item. Random slopes caused the models to have singular fits or fail to converge, so random slopes were not included in the models. Various models were then fit by adding the best-performing norms and Euclidean distances from different combinations of formant values as predictors to the base model. A categorical predictor for tense/lax was attempted to be used at one point in the modeling process, but it created a rank-deficient matrix because the tense/lax variable was not independent from the vowel category predictor. That is, the tense/lax predictor added no predictive capacity to the model, so it was not included as a predictor.

3.2.2 Results and discussion

The values for p in the p-norm with the highest correlation to the by-item mean distance ratings were clustered around 4.5. The maximum correlation was achieved at $p = 4.47$ ($r = .883$, $p < .001$). The 2-norm or Euclidean distance was also very highly correlated to the mean ratings ($r = 0.878$, $p < .001$), and the lowest correlation was found at $p = 1$ ($r = 0.864$, $p < .001$). A plot of the correlation values for different choices of p is presented in Figure 3.2. These results generally agree with results reported in Klatt (1981), where measuring the area between two spectral curves explained about 80% of the variance in distance rating data from listeners.

As a simple empirical verification of how significantly correlated the norm value is with the participant's ratings, bootstrap resampling was performed 100,000 times by creating resamples with replacement with the same number of observations as the original sample ($n = 297$) for both the norm values and the mean ratings and then calculating the correlation between these two random samples. The mean magnitude of correlation in this resampled distribution was 0.05, with a standard deviation of 0.03, and the highest achieved magnitude of correlation was 0.29. These results,

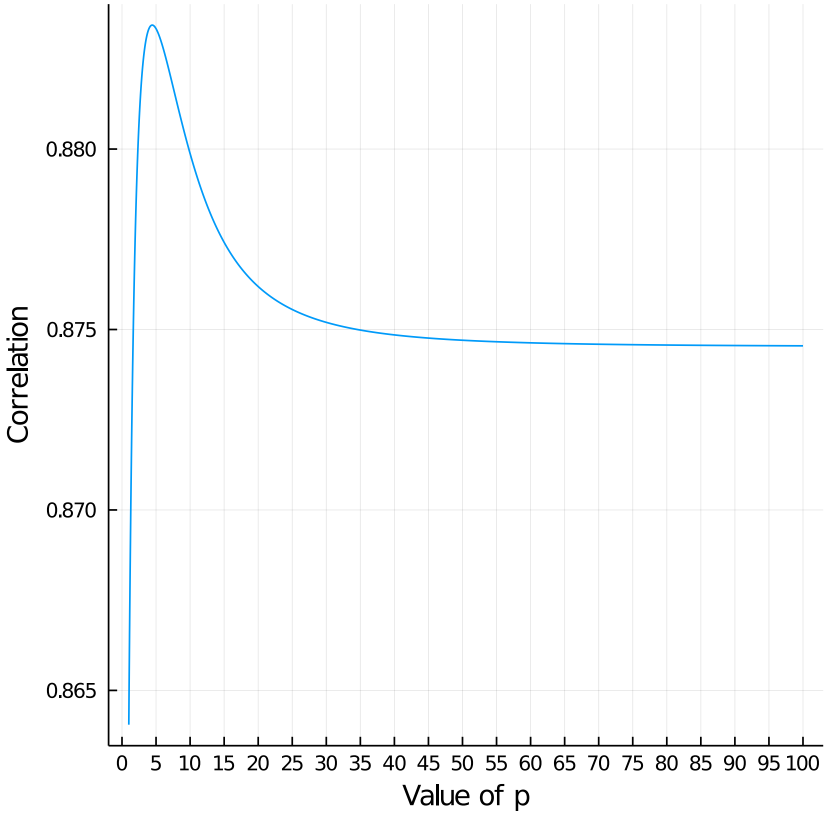


Figure 3.2: Correlation between by-item mean distance ratings and choices of p for the p-norm distance function. Note that there is no data associated with values less than 1.

along with the correlation test, demonstrate the strength of the association between a norm of the difference between MFCC vectors and the participant's ratings of the distance of the stimuli. Overall, these results suggest that performing these linear algebra computations on MFCC vectors has a strong relationship with human perception of acoustic distance between vowels.

An additional comparison was performed to see how the 4.5-norm of the MFCC vector compared to calculating the Euclidean distance (2-norm) between the F1 and F2 values used to create the synthetic vowels pairs. The achieved correlation was high ($r = 0.81$, $p < .001$). Individually, F1 differences achieved a moderately high correlation ($r = 0.66$, $p < .001$), and F2 differences achieved a slightly lower correlation ($r = 0.60$, $p < .001$). Note that the correlations between the formant-based distances and the average ratings are lower than the best correlations when using MFCC-based distances. Because MFCCs are a more complete summary of the speech spectrum than formant values, the higher correlation for the MFCC-based distances suggests that calculating acoustic distance with more full representations of the acoustic spectrum result in a closer association with participant behavior. That is, human judgments of acoustic distance seem to involve the entire spectrum of the sounds, not just the most relevant acoustic cues for speech perception. These results are buttressed by findings from Ito et al. (2001) and Nenadić et al. (2020) that, while formants are prominent contributors to vowel perception, other parts of the spectrum can be used to compensate for missing formant information. Still, not all speech sounds have strongly associated formant patterns, so representations based on the full spectrum are likely more appropriate than just formant values for the purposes of cognitively modeling acoustic distance for a broad range of speech patterns.

A visual display of these formant-based distances and the MFCC 4.5-norm is given in Figure 3.3. Note how the association between the distance metric and the by-item mean rating grows more linear as more of the acoustic spectrum is incorporated. To be clear, these results should not be construed to suggest that the human mind is literally calculating a norm on difference values between acoustic information. Rather, these results suggest that the p-norm from linear algebra provides an effective way to capture the behavioral performance of listeners.

As for the LMER models, only trial number and the distance value were signif-

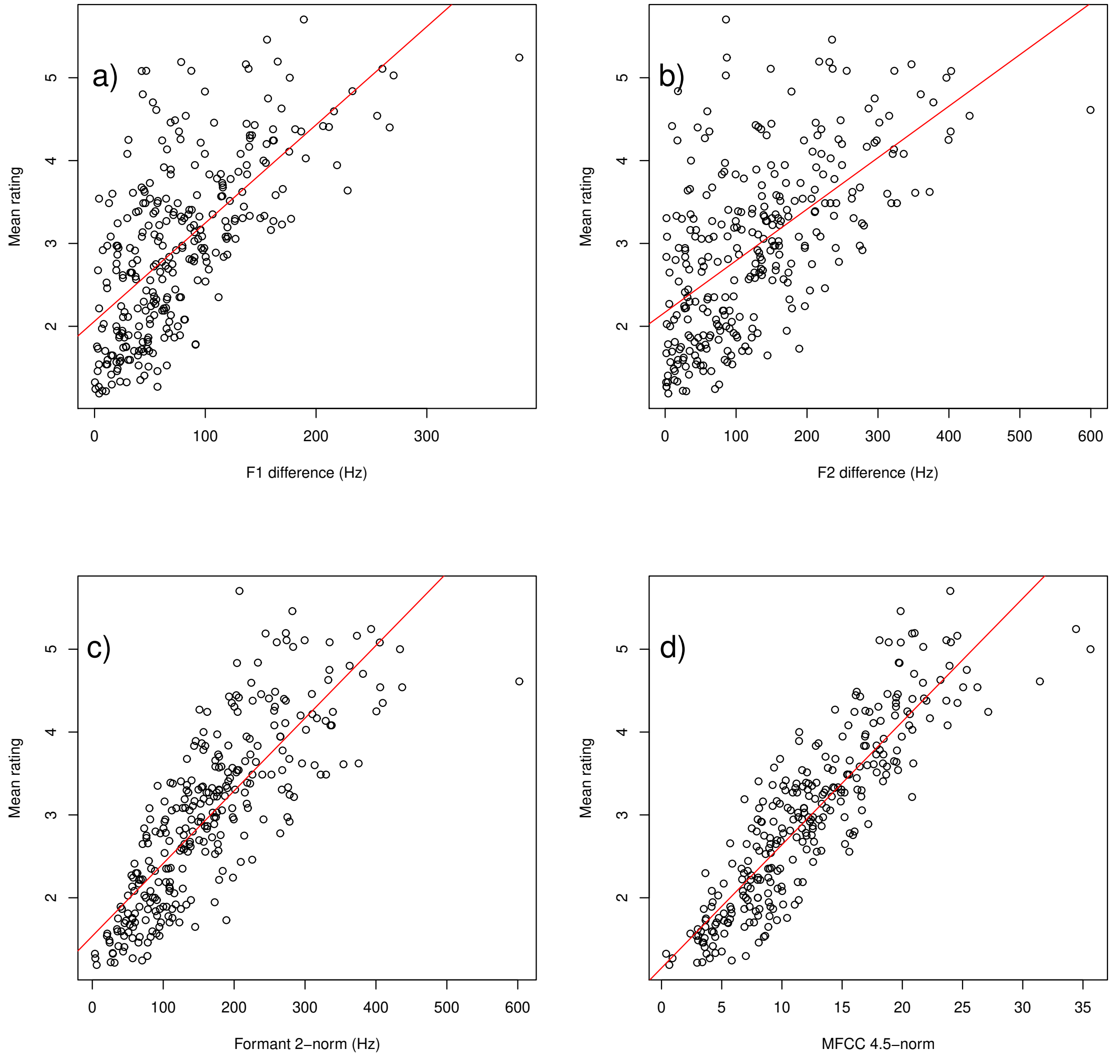


Figure 3.3: Scatter plots for mean by-item ratings over a) F1 difference ($r = 0.66$), b) F2 difference ($r = 0.60$), c) formant 2-norm (Euclidean distance, $r = 0.81$), and d) MFCC 4.5-norm ($r = 0.88$).

icant in all models. Some vowel contrasts were significant in some models, though these vowel results are not reported in detail because a vowel's significance was in-consistent across models, and vowel category was, after all, a control variable. The Akaike information criterion (AIC, Akaike, 1973, 1974) value was also calculated for each LMER model. AIC is a quantification of model fit that rewards fitting trends in the data but penalizes model complexity. Lower values indicate a better fit. The AIC values can be seen in Table 3.3. The trend of which version of acoustic distance best matched the listeners' behavior matches that of the correlation analysis, where the MFCC-based distances performed better than the formant-based distances, and the 4.5-norm performed the best.

Table 3.3: AIC values from LMER models for the different types of distance evaluated. The AIC values are sorted in ascending order, producing a best fit to worst fit ordering.

Distance type	AIC
4.5-norm	36,367.68
2-norm	36,380.24
∞-norm	36,386.98
F1 and F2	36,503.80
F1	36,648.87
F2	36,692.81

3.3 Experiment 2

The goal of Experiment 2 was to determine when a listener would be able to tell the difference in duration between two sounds. This experiment was designed to collect data relevant to the question posed in the introduction of what the temporal constraint should be on the dynamic time warping process. The experiment is designed to elicit responses that can be used to determine a just noticeable difference threshold for duration perception. The idea is that an acoustic frame in a word should be able to be compared with acoustic frames in another word to the extent that a listener would not notice the temporal mismatch. That is, when acoustically comparing two words, a given acoustic frame should only be compared with nearby frames to the extent that a listener would not notice the duration difference implied by the additional comparisons. In many ways, the MFCC-by-time representation of the words can be thought of as a stand-in for an alphabetic transcription. The difference is that each acoustic frame is representing a discretized time step in the acoustic trajectory of the word it belongs to. To the extent that the acoustic frames can be assigned to segments, they can also specify the state of a segment in time. The state specification is conceptually very similar to using multi-state models for phonemes in hidden Markov models for automatic speech recognition (consult Jurafsky & Martin, 2009, Chapter 9), whereby machine learning models are trained to detect subphonemic states of phonemes. The difference when using MFCCs and dynamic time warping is that the states are determined on a word-by-word basis in a bottom-up, empirical fashion.

Consider, for example, calculating the acoustic absement for the word pair *bad* [bæːd] and *bat* [bæt]. The MFCC representation of each word can be taken as an acoustic model of each word. Take {æː$_1$, æː$_2$, æː$_3$, æː$_4$} to be the frames representing [æː] in *bad*, and take {æ$_1$, æ$_2$} to be the frames representing [æ] in *pad*. If the unconstrained alignment were to come out as shown in Figure 3.4, and if listeners could detect duration differences greater two time steps, the alignment would be unrealistic. The reason that it would be unrealistic is that it suggests the acoustic region and information represented by æ$_1$ corresponds to acoustic events that span four sequential frames in the model for [æː]. Thus, the æ$_1$ frame would need to contain all the acoustic information in all the frames of [æː], or the acoustic information in $æ_1$ would need to be spread over more frames than currently exist. More concretely, the onset formant frequencies might be roughly represented in $æ_1$, while the offset formant frequencies are contained in $æ_2$. Whereas, the onset for [æː] might be represented in $æː_1$, with inherent spectral change (Nearey & Assmann, 1986) occurring over $æː_2$ and $æː_3$, ending with the offset formant frequencies in $æː_4$. While the alignment should allow some elasticity for duration, it should not allow a particular frame to be mapped beyond a reasonable temporal extent. In the present example, the onset formant frequencies for [æ] should not be mapped across the entire extent of [æː], spectral change and all. Choosing the correct radius constraint for dynamic time warping can prevent such unrealistic alignments from occurring.

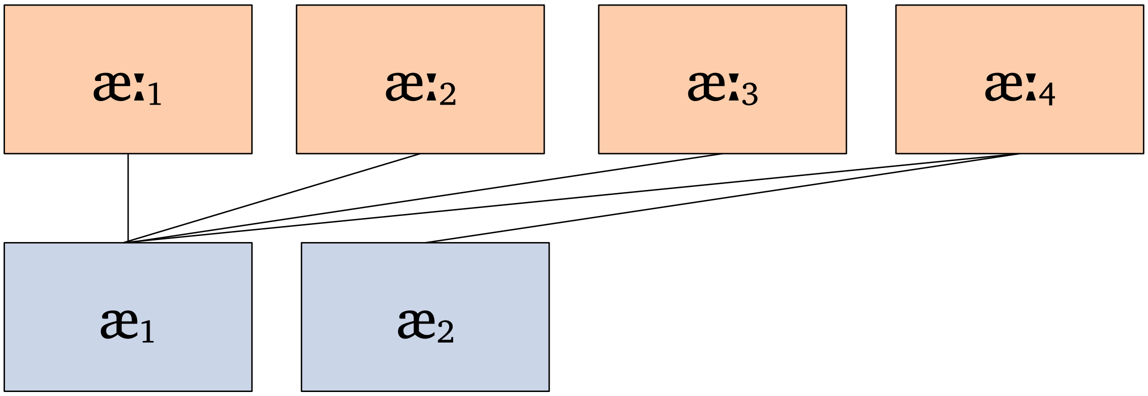


Figure 3.4: An example of an unrealistic alignment. The 1_b frame is mapped to all of the frames of [p] even though listeners would be able to discern this duration difference. The 2_b frame is only mapped to 4_p.

3.3.1 Methods

In this experiment, participants were asked to take part in a reminder task designed to elicit responses related to the discrimination of vowel duration. The ratings were used as the independent variables for a regression analysis to determine a just noticeable difference threshold.

Participants

As in Experiment 1, participants were recruited from the University of Alberta Department of Linguistics subject pool, which consisted of undergraduate students enrolled in introductory linguistics classes. Participants were again compensated for their time with credit for the course they were enrolled in. The same demographic questionnaire from Experiment 1 was used for this experiment.

One hundred twenty-five participants took part in the experiment, and data from all 125 participants who participated were successfully uploaded to the server. Of these 125 participants, 49 were determined to be monolingual native English speakers during childhood on the basis of only speaking English in the home before the age of 5. Of these 49 participants, 1 was removed from the analysis for reporting not listening in a quiet room. Among the remaining 48 participants, 31 reported their gender as “female”, and 17 reported their gender as “male”. The average age was 20.6 years old ($SD = 3.34$).

Materials

As in Experiment 1, the stimuli were based on the values from the Hillenbrand et al. (1995) data set. This time, only three vowel categories were synthesized: /i/, /ɑ/ and /u/. The F1, F2, and pitch values were taken from the medoid male speaker from before, speaker “m07.” Of note, the pitch was kept constant at 99 Hz (the speaker’s mean pitch across all vowels) so that pitch differences between vowel categories would not influence the results. Whereas, the F1 and F2 values were taken as the reported mean F1 and F2 values across the speaker’s vowel productions for each category.

For each category, a continuum of vowels was created along the duration dimension, ranging from 100 ms below the reported vowel duration to 100 ms above, in steps of 20 ms. The result was 11 steps for each category. A table containing the acoustic information for each of the vowel categories can be seen in Table 3.4.

Table 3.4: Acoustic variables for KlattGrid synthesized vowels.

Vowel category	F1 (Hz)	F2 (Hz)	Duration range (ms)
/ɑ/	642	949	165 – 365
/i/	346	2379	155 – 355
/u/	326	997	135 – 335

Due to a combination of the sampling rate, the period of the vowel sounds, and the desired duration, the synthesized vowels did not end at a zero-crossing. To avoid introducing pops that might affect the listener's perception of the duration, a 15 ms amplitude ramp down was applied during the synthesis process by using the voicing amplitude track. An amplitude ramp-up was not needed because the synthesis procedure in `Praat` already forced a sort of ramp up onto the signal. An amplitude point was set at 90 dB at 15 ms before the offset of the sound, and then a second amplitude point was set at 0 dB at the end of the sound. In so doing, clicks and pops were avoided that would otherwise occur due to the sounds not ending at or near a zero-crossing. The 15 ms value was chosen by manual search as the smallest ramp down duration that eliminated the clicks and pops in the stimuli. Each sound was subsequently scaled to 70 dB using the `Scale intensity...` function in Praat.

The sounds were then spliced together to form a reminder task, which was informed by Lapid et al. (2008). The median sound in terms of duration was used as the reminder sound for each category. Then for each sound in a vowel category continuum, the stimuli were created by concatenating the reminder sound, a 750 ms silence, and the target sound within the vowel category duration continuum. The result was 11 stimuli per vowel category, for a total of 33 unique stimuli.

Procedure

As in Experiment 1, the experiment was run online using `jsPsych`. This time the task was a reminder task, where participants were instructed to listen to pairs of sounds and indicate whether they believed the second sound in a pair was shorter or longer than the first sound. They responded by pressing "S" on their keyboard to indicate shorter or "L" to indicate longer. A fixation cross was displayed for 500 ms before each trial. The same demographic questions as in Experiment 1 were used. Each participant heard each stimulus a total of 10 times, for a total of 330 trials in the experiment. The presentation order of the stimuli was randomized at experiment startup, and the participants were prompted to take breaks after every 50 stimuli, as in Experiment 1. Each trial was untimed, and the same deception as used in Experiment 1 was used in this experiment.

Of the 48 monolingual participants, 16 indicated that they were not wearing wired headphones while participating in the experiment. A chi-square goodness-of-fit test indicated that the participants wearing wired headphones responded significantly different than those who were not ($\chi^2 = 12.02$, $p < .001$), so those further 16 participants were excluded from the analysis. In the end, 32 participants (65.31% of the original 49) remained to be analyzed. Within this subset, 18 of the participants reported their gender as "female", and 14 reported their gender as "male". The average age within this subset was 20.7 years old ($SD = 2.93$).

There were 10,560 responses before removing implausible responses. For the same motivations as in Experiment 1, 91 responses (0.86%) with a reaction time greater than 5 s were dropped. There remained 10,469 data points to analyze.

Analysis

A psychometric function was fit to the data as a mixed-effects logistic regression using `lme4` in the `R` programming language. The fitting process involved a forward-fitting process for random structure and a backward-fitting process for fixed structure. Intermediate models were compared using the `anova` function from `lme4`. The logistic regression model can then be used to calculate the just noticeable difference. Some previous studies have used a single-variable logistic regression to determine the just noticeable difference (Lapid et al., 2008; Tomaschek, 2013). Mixed-effects regression presents an advantage from a modeling perspective because it allows for additional nuisance variation to be accounted for in order to draw out a less noisy estimate of what the effect of the duration difference is on stimulus discrimination.

The location of the just noticeable difference or difference limen has been defined as half the interquartile range (e.g., Lapid et al., 2008). This quantity is also known as the semi interquartile range. It presents the average distance from the second quartile to the first and third quartiles. It is important to give this quantity some consideration. First, the second quartile of the function should represent the region where a listener's response is most uncertain, that is, where they are responding at chance. The question is, then, how far from that second quartile point must the duration difference be before the listener can reliably notice it? One quartile away from the median is, admittedly, a somewhat arbitrary cutoff. Indeed, Boring (1939) also gave examples of using the standard deviation, the probable error, or the modulus of a normal distribution (which Gorroochurn, 2016, noted is a historical term denoting $\sigma\sqrt{2}$). Levitt (1971) used slightly different values of the roughly 0.29 and 0.71 quantiles as an estimate of the standard deviation, but for a transformed up-down procedure that was not used in the present experiment. Ultimately, some measurement of statistical spread from a central value is used. Lapid et al. (2008) described the first and third quantile cutoffs as the most typical in the literature, so those are the points that will be used in the present study for maximal compatibility with previous results. It is also easier to calculate the semi interquartile range with algebra from the logistic regression coefficients as compared to other statistics like the standard deviation.

A single-variable logistic regression can be expressed in the form given in Equation 3.4:

$$\hat{y} = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x\,, \tag{3.4}$$

where p is a probability, β_0 is the model intercept, β_1 is the slope associated with the duration, and x is a duration value. To find the duration value associated with the first quartile, Equation 3.4 must be solved for x when $p = 0.25$. To find the duration value associated with the third quartile, Equation 3.4 must be solved for x when $p = 0.75$. From these values, the just noticeable difference can be calculated.

3.3.2 Results and discussion

The final model consisted of fixed effects for stimulus duration and vowel category, with a reference level of /ɑ/. There were by-subject and by-item random intercepts. Fixed effects that were dropped were gender, education level, and trial number. A random slope for vowel category on the by-subject random intercept was fit, but it did not significantly increase model fit, so it wasn't included in the final model.

Additionally, a random slope for vowel category was attempted to be fit for the by-item random intercept, but it resulted in a singular fit. The table of coefficients for the final model can be seen in Table 3.5. Like Experiment 1, a t-value of 2 or more in magnitude was taken as an indication of statistical significance.

Table 3.5: Table of coefficients for fixed effects of final model.

Predictor	Estimate	Std. Error	z value
Intercept	0.12	0.13	0.95
Duration difference	0.04	0.00	41.51
Scaled trial number	0.10	0.03	3.10
/i/	-0.21	0.12	-1.70
/u/	-0.83	0.15	-5.43

As in Lapid et al. (2008), to assess how well the model fit the data and matched the proportions of response types to each stimulus, the estimated values were plotted against the raw proportions and can be seen in Figure 3.5. There is relatively little difference between the estimated values and the raw probabilities. Some of the probability values are contained within the confidence interval of the estimates. The values that are outside of the confidence interval are only slightly outside, and the differences would be unlikely to make a practical difference on the results. As such, the model fit seems to match the data well.

The effect of the duration variable was then used to find the just noticeable difference for the perception of duration differences. After solving for x in Equation 3.4, the first quartile is associated with a duration value of -22.46 ms and the third quartile is associated with a duration value of 28.08 for an interquartile range of 50.54 ms, which is illustrated with the black lines in Figure 3.5. The semi interquartile range, and thus the just noticeable difference threshold, is 25.27 ms. To use this value

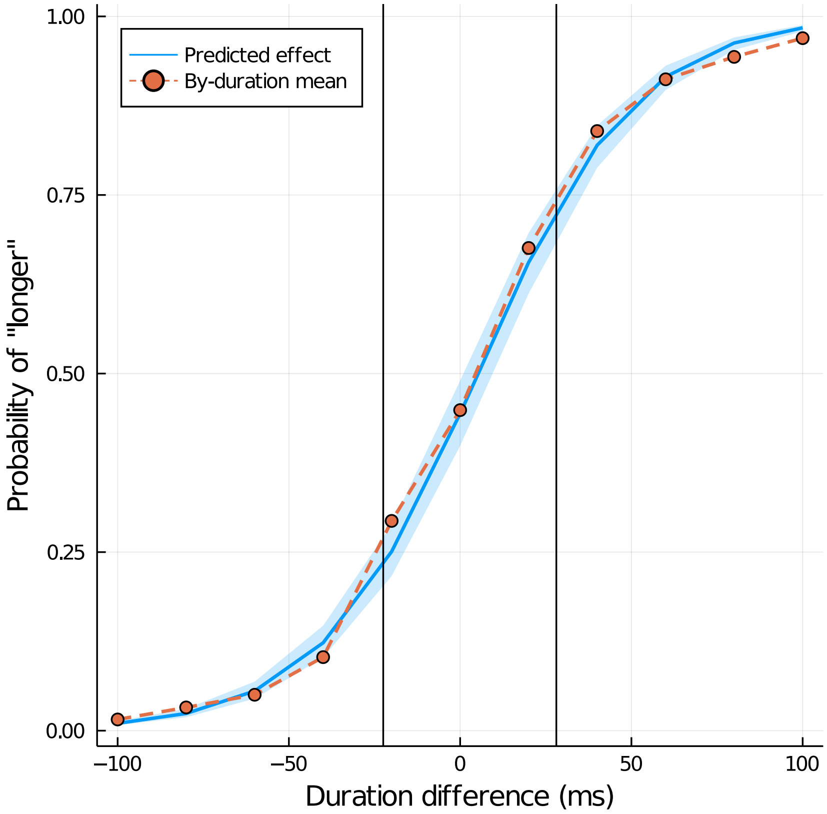


Figure 3.5: Comparison between mean probability of responding "longer" and the predicted probability from the logistic model. A 95% confidence interval based on evaluating the fixed-effect portion of the model is presented for the predicted values in blue. The confidence interval and the predicted values were calculated using the `effects` R package (v4.1-4, Fox & Weisberg, 2019). The first quartile and third quartile are indicated with the vertical black lines.

as a guideline for acoustic absement calculations, the radius must be at least this value. So, in the case of computing acoustic absement with dynamic time warping and using 10 ms between steps of acoustic information, 30 ms (or 3 steps) would be an appropriate radius to choose as a limit for how far forward or backward in time any given time step could be matched to other time steps. In effect, this choice means that a particular time step would not be allowed to be compared with time steps more than 30 ms before or after it.

The 25.27 ms just noticeable difference threshold is close to approximately 31 ms threshold that Lapid et al. (2008) found for discriminating between white noise sounds of different durations. Additionally, the 25.57 ms threshold is similar to the 32.9 ms threshold Henry (1948) reported for for pure tones of 175 ms in duration. (This last threshold is calculated by multiplying the reported Weber ratio of 0.188 by the duration of 175 ms to get the just noticeable difference threshold of 32.9 ms.)

3.4 Re-analyzing previous models

(Kelley & Tucker, 2021c) fit a series of generalized additive mixed models (GAMMs) were fit to auditory lexical decision data. The goal was to determine if lexical competition could be quantified acoustically by using dynamic time warping. This would be as opposed to quantifying lexical competition through textual means as is done when calculating phonological neighborhood density. A new variable called acoustic distinctiveness was created by calculating the acoustic absement from one word to all words in the Massive Auditory Lexical Decision (MALD) database (Tucker et al., 2019). Acoustic distinctiveness was found to produce a similar effect in the data as phonological neighborhood density. It was reasoned that the absement value that comes from dynamic time warping was relevant to cognitive processes, especially those used during spoken word recognition. The experiments in the present study were designed to fine-tune certain aspects of the dynamic time warping algorithm to be more related to human perception. As such, it makes sense to re-fit the GAMM models from Kelley and Tucker (2021c) after calculating acoustic distinctiveness using the results of the previous experiments in the dynamic time warping algorithm.

By so doing, the relevance of the experimental results to determining the acoustic absement between words can be assessed.

3.4.1 Data

The data used came from the MALD database. The data to be modeled in the GAMMs was the response latency in the lexical decision task. MALD is an auditory lexical decision megastudy, with about 28,000 real words recorded by a young male speaker of western Canadian English. Each word was responded to in auditory lexical decision at least 4 times from among 231 unique participants who were also native speakers of western Canadian English, for a total of 227,129 data points (including responses to both real words and pseudowords). Stimuli sets were also recorded for two other speakers: a young female and an older male, both of whom were native speakers of western Canadian English. Further details are available in Tucker et al. (2019) on the recording process for the young male speaker, the auditory lexical decision task, and the variables included in the data set.

There were a total of 113,675 responses to real words in the data set. Responses faster than 500 ms were removed for being implausible responses. For modeling purposes, response time was measured from stimulus offset to help factor stimulus duration out of the response latency values themselves. These response times were then logged. Only correct responses to real words made after stimulus offset were retained. This restriction leaves 96,001 responses (84.45% of the original number) for the modeling process.

3.4.2 Methods

Acoustic distinctiveness was calculated using the stimuli from the young male speaker in the experiment. To keep the models in line with those in Kelley and Tucker, only words that are common between the three speakers will be used. In total, there were 26,005 words used in the calculation of acoustic distinctiveness. Each word in the database was first transformed into a series MFCC vectors using the `MFCC.jl` package in the `Julia` programming language. The window length was set to 25 ms, and the step size for the windows was set to 10 ms. 13 coefficients were calculated, and the zeroth coefficient was replaced with the log energy of the frame.

Then, to calculate the acoustic distinctiveness for a given word, its absement to all words in the data set was calculated with dynamic time warping. Three particular configurations of dynamic time warping were compared. The first is the standard formulation, which uses the 2-norm (that is, Euclidean distance) to compare MFCC slices of the words with no constraints beyond the default dynamic time warping ones on what frames can be compared to each other. The second configuration incorporates the results from Experiment 1 and uses the 4.5-norm to compare MFCC slices of the words, still with no additional constraints on what frames can be compared. And, the third configuration incorporates the results from Experiment 2 and sets a radius constraint on the dynamic time warping process so that a given acoustic frame in a word can only be compared with frames up to three time steps before or after it in the other word, while still using the 2-norm to compare frames. An example alignment for two hypothetical words with the radius set to 3 is shown in Figure 3.6. Note how there is some degree of compensation enforced in the warping path when a given time step is stretched across various time points. One such example in the given figure is how time step 4 in word two would only be allowed to be stretched by 1 time step due to how long time step 3 was stretched in word 2.

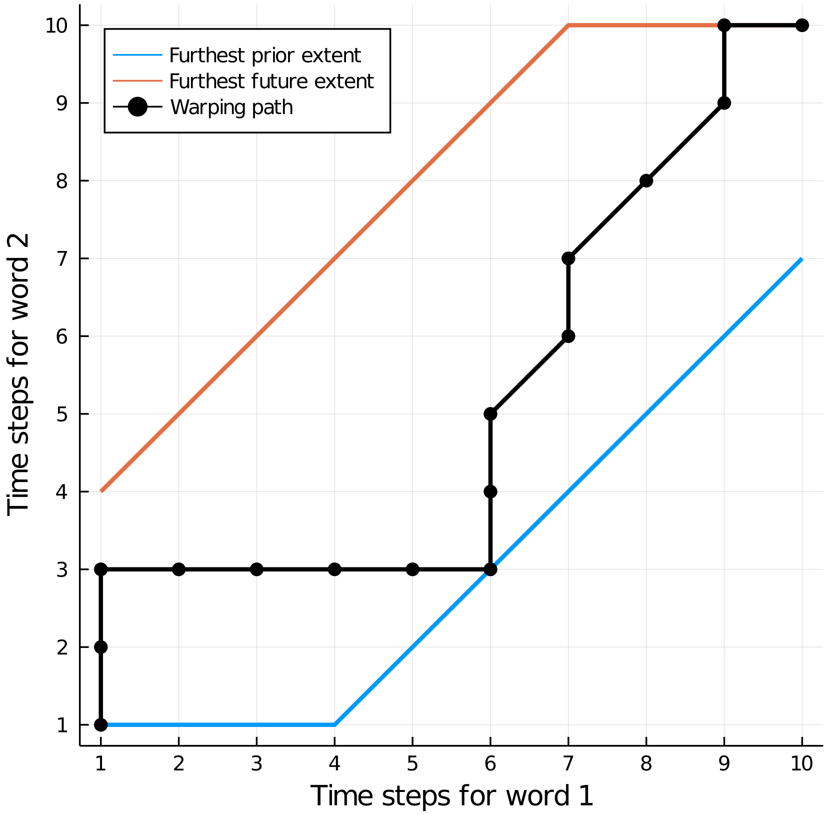


Figure 3.6: Hypothetical warping path using a radius of 3 time steps. The order of the axes (with word 1 on the x-axis and word 2 on the y-axis) indicates that we are looking at word 1 being stretched onto word 2. The blue line on the bottom represents the earliest time point in word 2 that a given time step in word 1 can be mapped onto. Similarly, the orange line on the top represents the latest time point in word 2 that a given time point in word 1 can be mapped onto.

The baseline model from the second analysis in Kelley and Tucker (2021c) was used as a baseline model. It consisted of fixed-effect smooths for trial number, log frequency from the Corpus of Contemporary American English (Davies, 2008) from the MALD data set, phonological uniqueness point, and log moving average response latency (as defined in ten Bosch et al., 2018). There was also a smooth term for a by-subject random intercept. The different forms of calculating absement were assessed by adding them individually to the base model and comparing how much the fast restricted maximum likelihood (fREML) improved. The fREML statistic is GAMM analog to AIC, where lower values indicate better fit. Models were fit using the `mgcv` (v1.8-36, Wood, 2011) package in `R` version 3.6.3. Model criticism was not applied to the models because it could not be guaranteed that each model would be fit on the same data.

3.4.3 Results & discussion

The statistics indicating model fitness and explanatory power can be found in Table 3.6. There was slightly worse performance in terms of fREML when using the 4.5-norm or the 3-step radius. As an implementation note, because some words were too short to only have a 3-step radius, the last time step was used for all comparisons that would otherwise be out of range when calculating the absement using the 3-step radius constraint.

Table 3.6: Indicators of model fit for the GAMMs. The "fREML improvement" column indicates the magnitude of the decrease in fREML from the baseline model, so higher numbers are better.

Variant	fREML improvement	Adj. R^2	Deviance explained
Standard	3,395.95	0.28	27.75%
4.5 norm	3,309.87	0.28	27.43%
Radius of 3	3,359.67	0.28	27.70%
Both	3,275.99	0.27	27.60%

Overall, there were numerical differences in the fREML scores, with the standard variant performing the best. However, the R^2 and deviance explained statistics suggested that the differences may not have meant much in practice, whether or not they are statistically significant. It may be the case that the 4.5-norm distance function relates better to more static signals, like simple vowels. Whereas, the differences that it would highlight between static signals do not matter much or at all when comparing dynamic signals like full words. Note, too, that the 4.5-norm distance is more computationally expensive to compute than the 2-norm/Euclidean distance because there are techniques that speed up Euclidean distance. For comparison purposes, computing the acoustic distinctiveness values using Euclidean distance for all the words in MALD could be performed in about six hours using parallel multiprocessing with 15 cores with a clock speed of 2.93 GHz. On the same hardware, computing acoustic distinctiveness using the 4.5-norm distance took around one week. As such, any modest accuracy benefit that may ostensibly be achieved using the 4.5-norm distance is likely not worth the significant time burden that comes with it. To drive that point home even further, consider that calculating acoustic distinctiveness on the same data set would take around 15 weeks to run on a single core of the same power when using the 4.5 norm distance. Some minor improvements are likely possible. The speed difference between the 2-norm distance the 4.5-norm distance will likely remain prohibitive, though.

The results of using the 3-step radius are more felicitous, though. There was a difference in fREML between the model using standard dynamic time warping and the model using the 3-step radius. But, this difference was smaller in magnitude than when comparing the 4.5 norm distance to the 2-norm distance. Using the type of temporal constraint reduces the amount of computation needed to be performed. As such, it would be reasonable to use this parameter setting even if just for speedup reasons since it does not seem to significantly disturb the relationship between acoustic absement and human behavior.

3.5 General discussion

There are several factors to discuss regarding the components of dynamic time warping used to calculate acoustic absement. When comparing MFCC vectors to each other, it seems that many of the p-norm possibilities for distance functions are useful. And, in fact, this is something that should well be true anyway if any single p-norm based distance function is useful. As previously mentioned, this result is predicted by the equivalence of vector norms, where when one norm is high for a vector, all the other norms will be high for that same vector, and the analog is true when a norm is low (see, for example van de Geijn & Myers, 2020). This is clearly obviously connected to the notion of correlation. As such, while p-norms where $p \approx 4.5$ provided the highest correlation with human judgments in these data, virtually any p-norm would perform well too. For reasons of familiarity and computational efficiency, it is likely better to choose to use the Euclidean norm (2-norm) or the squared version of it because there already exist algorithms that can compute pairwise Euclidean norms/distances far more efficiently than other p-norms by taking advantage of matrix-matrix multiplication.

In terms of modeling human perception, these results suggest that the choice of a distance function is likely to be largely inconsequential if deciding between distance functions that are based on p-norms. Other forms of distance functions exist, however, and they may have larger consequences for modeling purposes. Extending the results of the distance function search further, it may be the case that it is the acoustic features that matter most for modeling human perceptions of acoustic distance. Simply put, the acoustic features must have enough capacity to allow for speech sounds to be differentiated, that is, discriminated. So long as the chosen acoustic features maintain the discriminability of different speech sounds or components of the speech signal, the choice of distance function does not matter very much because much of the comparison has already been implicitly performed via the feature set. It can be postulated, then, that there is some sort of cognitive acoustic feature space (one example of which is the F1-by-F2 vowel space) in which dissimilar parts of the speech signal are naturally separated and a distance function merely serves as a way to quantify and rank the similarity between segments of acoustic information in the speech stream. This concept of space is similar to how semantic vector spaces are used in other areas of linguistics to group words with similar meanings and separate words with disparate meanings (e.g., Baayen et al., 2019; Foltz et al., 1998), though a physical interpretation of acoustic features is more straightforward than for semantic features. The present study used MFCCs as the features of that space, but many other acoustic properties may also be used as the features. The resultant processing mechanism by which words receive activation based on acoustic similarity may then be algorithmically similar to those used in models like Shortlist B, Fine-Tracker, and DIANA.

As for the duration perception, the use of a radius has a clearer motivation. The results of calculating acoustic distinctiveness did not substantially change between standard dynamic time warping and the use of the radius. However, using the radius can be leveraged to decrease the amount of calculation that is needed. This is a common technique to increase computational efficiency in data mining (Ratanamahatana & Keogh, 2004; Wu & Keogh, 2020), and it dates back to original uses of dynamic time warping by Sakoe and Chiba (1978) for speech recognition. However, Ratanamahatana and Keogh (2004) and Wu and Keogh (2020) pointed out that the increased computational efficiency is merely a fortunate coincidence of what this sort of constraint should actually be used for. In fact, they noted that such constraints can prevent what is known as “pathological warping,” where part of a time series gets mapped over a section that is unrealistically long in another time series. A phonetic example would be one phone in a word being mapped over five phones in another word. Such warpings are undesirable, and having some sort of radius on the temporal extent that a frame in one word can be mapped onto another prevents such a situation. This is all the more important given the goal to model cognitive processing, in which it is implausible that a listener would map one short time step of acoustic information across many time steps in another sound.

Different models of spoken word recognition can be approximated with dynamic time warping by tuning the duration constraint as well. With a short duration radius, the results of the dynamic time warping comparisons will resemble the predictions of the cohort model. The cohort model’s behavior is such that words will be pruned from the list of competitors when phonetic mismatches between onsets occur. With a short duration radius in dynamic time warping, words where this sort of onset mismatch occurs will have a high absement value because the radius constraint will prevent the algorithm from accommodating the mismatch. Similarly, a wider radius can allow for roughly one phone to be skipped, allowing for results similar to those in the neighborhood activation model and phonological neighborhood density, where words that are one phone different will be competitors. The wider radius will allow the algorithm to accommodate longer-term mismatches in acoustic comparisons of words, which will allow for words that are one-off in terms of phones to accrue less absement than they would with a stricter duration radius. Of course, these examples are merely possible approximations since dynamic time warping as used in the present study is working on the acoustic, continuous level, not on the segmental, discrete level that the cohort and neighborhood activation models work on.

Insofar as recommendations for acoustic absement calculations using dynamic time warping, the evidence suggests that using Euclidean distance between acoustic slices and using a 3-step radius will provide the best result in terms of both computational efficiency and matching participant behavior. However, adding a sort of radius to the dynamic time warping process broaches the issue of how to handle sequences of wildly different lengths. For example, how should *a* be compared to *crepuscular*? It is likely that that *crepuscular* will be more than three times longer than *a*, in which case it is undefined what to do once the last time step in *a* is reached and has been matched with three time steps in *crepuscular*. Ratanamahatana and Keogh (2004) summarized a common technique for resolving length or duration differences wherein the shorter sequence is reinterpolated so that it is the same length as the longer sequence. Yet, this technique cannot reasonably be applied to speech data because the reinterpolation process would put one of the utterances on a different time scale than the other. It is unreasonable to cast cognitive processing as operating at a variable time scale that depends on the relative length difference of the input and a lexical template.

In models of spoken word recognition like the cohort model, TRACE, TISK, EARSHOT, and DIANA where the signal is processed sequentially, it might be reasonable to expect that the search would be abandoned in this sort of scenario. This could certainly be written algorithmically into a model of spoken word recognition without too much trouble. However, it is less clear what should be done when a word needs to have some sort of comparison to every other word in the lexicon, as when calculating acoustic distinctiveness. In Analysis 2, the choice was made to force all comparisons that would otherwise be outside the radius to be compared to the last slice of the shorter word. This will effectively penalize the score for the short word in this sort of scenario by introducing many comparisons that may not be between acoustically similar regions. Ultimately, there is no perfect solution because the sort of global comparison where every candidate must be fully evaluated for a given input signal is not ecologically valid due to the nature of competition and activation. This is especially true given descriptions of competition (Luce & Pisoni, 1998) where activation of competitors is mediated by how well the incoming acoustic signal matches the potential candidates for recognition. In sum, the choice of how to penalize this sort of durational mismatch for global lexical measures will necessarily be non-ecological and such choices must be predicated on other desiderata, like having high absement values when the duration mismatch is great.

Future work may wish to explore different methods of computing distance or restricting how long a particular frame of audio can be stretched in time. For example, Bagge Carlson and Chitre (2020) have recently identified Wasserstein distance (among other spectral distances) as a candidate to quantify spectral distances that may be reflective of how humans are sensitive to frequency differences. However, the spectrum of the speech signal changes constantly, sometimes to a drastic extent such as from a vowel to a stop or fricative. As such, the Wasserstein distance would need to be used as the method to compute the distance between small slices of acoustic information. Using it globally to compare the spectra for entire words is not ecological, either, because the order of the changes in the acoustic spectra is an important differentiating feature for words made up of the same phones in different order, such as *apt* [æpt], *pat* [pæt], and *tap* [tæp]. Similarly, borrowing concepts of differential equations and dynamical systems from planning systems involving algorithms like dynamic time warping (e.g., Lavalle, 2006) may help express the calculation of acoustic absement and distance in a way that better interfaces with fields such as control theory and artificial intelligence.

Additional work must also be done to build upon the duration discrimination results in Experiment 2. It is particularly important to determine whether duration differences in other speech sounds are also perceptible at the same threshold as the vowels tested here. For example, in conjunction with the results from the present study, the finding from Kewley-Port et al. (1988) that voice onset time discrimination had a 50% crossover point around 34 ms in good listening conditions suggests that listeners have different temporal sensitivities to different categories of speech sounds. Of course, it may well be the case that testing sensitivities that relate to phonological contrasts (such as voice onset time) will result in different sensitivity thresholds than testing for the ability to perceive phonetic differences between sounds. If dynamic time warping is to be used as some sort of algorithmic representation of auditory processing when computing acoustic absement, these sorts of differences must be accounted for. Within the algorithm, it may be possible to determine variable boundaries with which to compare the frames of audio based on what type of sound is represented. However, the most straightforward way of doing this would be to build some sort of speech sound category recognizer, which itself brings assumptions and marked complications to the algorithm. It may also be possible to make the larger acoustic comparison process account for differences between speech sounds simply by ensuring that the acoustic representation allows for these differences to be salient. If the acoustic differences between different categories of speech sounds are apparent in the chosen representation, the very process of performing dynamic time warping may prevent the implausible temporal extension of a single frame or short sequences of frames from being mapped far beyond what the human auditory system would do.

3.6 Conclusion

There are two easily tunable components of dynamic time warping: the distance function used to compare time steps and the radius in which a time step is allowed to match with other time steps. By choosing settings that are informed by behavioral experiments, dynamic time warping as used to calculate acoustic absement can be made to better reflect human auditory processing. The experiments from the present study suggest that Euclidean distance is acceptable as a distance function, and its computational efficiency will likely outweigh its slightly lower correlation with human judgments than p-norms where $p \approx 4.5$. Using a radius of 3 time steps (when the time steps are 10 ms apart) will afford a close match to the reported duration discrimination results. Non-vowel speech sounds have different durational discriminability properties, and indeed different individual vowels may as well. However, the 3 time step window should suffice as a starting point, especially when the features for acoustic representation represent salient frequency differences well.

Overall, the results from these experiments serve as a perceptual grounding for dynamic time warping as an algorithmic representation of auditory processing, at least as regards words. As is typical in linguistics, this algorithmic representation is not meant to serve as a literal neuroscientific description of auditory processing. In fact, it is the mapping between acoustic frames that is important, not the trellis-filling process induced through dynamic programming. This mapping is intended to serve as a high-level description of the ongoing processing, and various implementations are assumed to be possible. This is analogous to the way that most programming languages work, where the language itself is used to specify a program at a high level, and a separate compiler or interpreter actually creates the machine code needed for program execution.

The overall results are suggestive of a model of acoustic absement between words that is grounded in dynamics and dynamical modeling like in the dynamical hypothesis of cognition (van Gelder, 1995, 1997). There is a history of dynamical models being used in phonetics and phonology, such as in articulatory phonology (Browman & Goldstein, 1992, 1995) and task dynamic models of speech production (Saltzman & Munhall, 1989). Dynamical approaches are, of course, a natural fit for obviously continuous phenomena like the movement of articulators during speech production. It is less immediately obvious that a concept such as "distance" (or some related quantity) between words should be thought of in a more continuous fashion. And, perhaps this is due to a predisposition to think of words as discrete objects, so a logical conclusion would be that the distance between words should be static and discrete. Yet, distance is itself the result of the accumulation of velocity over time, and distance is an indication of how absement is changing over time. Segmental or acoustic representations of words do also, however, specify some sort of state that changes over time, whether the state is segments or acoustic spectra. The temporal nature of both distance and words suggests that, in point of fact, dynamical modeling is an appropriate tool for words, distance, and absement. Dynamic time warping—customized based on experimental results such as those from the present study—is a step forward in modeling the nature of acoustic differences between words.

Chapter 4

Features for acoustic distance

Abstract

A coustically comparing words requires choosing a set of features by which to make the comparison. It is common to represent words as sequences of mel frequency cepstral coefficients (MFCC), though this set of features has not truly been compared to other features for human judgments of acoustic distance. The present study seeks to compare MFCCs to a set of features learned by a neural network that does tracks vocal tract resonances. A neural network was trained to track vocal tract resonances. The features it learned were probed with a simulated behavioral experiment and interpreted as being similar to resonant filters localized to specific regions of the spectrum. These features were then compared to MFCCs in quantifying lexical competition as part of regression modeling of response latencies in an auditory lexical decision task. MFCCs better quantified lexical competition in a way that matched human behavior. Subsequently, the features were compared as predictors of human judgments of acoustic distance between synthetic vowels. MFCCs correlated more highly with the ratings than did the neural network features. These results

overall suggested MFCCs—which summarize the entire spectrum—better modeled the acoustic information human listeners use during speech processing.

4.1 Introduction

Acoustic distance between words has a relationship to spoken word recognition. This concept is often invoked through the notion of sound similarity (Luce, 1986; Luce & Pisoni, 1998). Kelley and Tucker (2021c) used acoustic distance as the basis for a new way to quantify lexical competition. Words were converted to an acoustic format using mel frequency cepstral coefficients (MFCCs). Acoustic distance between two words' MFCC sequences was accumulated over time using dynamic time warping, resulting in a quantity known as absement. A word's average absement to all words in the lexicon—termed "acoustic distinctiveness"—was found to bear a strong relationship to response latencies in auditory lexical decision. Kelley and Tucker (2021b) delved further into the specifics of dynamic time warping to determine what sort of distance formula most closely matched human judgments of distance in a rating task, finding that Euclidean distance offered the best compromise between correlation to human judgments, familiarity of calculation, and computational efficiency.

However, another challenge when cognitively modeling the acoustic distance of speech sounds and spoken words is determining the best set of features to use. Kelley and Tucker (2021c) and Kelley and Tucker (2021b) used MFCCs due to convenience and familiarity. MFCCs have been used in a variety of other perceptual work (Bartelds et al., 2020; Mermelstein, 1976; Mielke, 2012). These features are commonly used because they are traditional and common features in automatic speech recognition research and development (Graves, 2012; Graves & Schmidhuber, 2005; Hinton et al., 2012; Jurafsky & Martin, 2009). Several other feature sets could also be used. A similar feature set is a (log) mel filterbank, as in Zhang et al. (2016) and Graves and Jaitly (2014) or perceptual linear predictive coding coefficients as in Hendriks et al. (2004). These options for representing an acoustic signal are all hand-crafted to mimic certain aspects of human perception. However, machine learning techniques also present an interesting possibility for creating features that are learned to be relevant for a given task. For a task that involves human speech, it could be possible to learn a set of features that is optimally useful for representing speech. The present paper seeks to investigate this possibility and compare features learned by a neural network to MFCCs as a basis for cognitive modeling.

The present study compares the acoustic features a neural network learns to MFCCs for cognitive modeling of phonetic phenomena. Necessarily, it will involve training a neural network. These features will be compared to previous cognitive modeling results using MFCCs in Kelley and Tucker (2021c) and Kelley and Tucker (2021b). Because it is important to understand what these new features represent, an attempt will be made to interpret them using a form of the previously discussed simulated experiment technique. These results will be discussed with specific reference to whether these features are a plausible alternative to MFCCs for cognitive modeling. Special attention will also be given to the process of neural network feature interpretation as a part of cognitive modeling and how such interpretations must be situated in a context that makes the numbers in the features meaningful.

4.1.1 Learning features

Machine learning techniques have a history of use in feature engineering (Hastie et al., 2009, Chapter 5). Perhaps the most prominent technique of recent years is deep neural networks. An attractive feature of deep neural networks is that they can learn useful representations and features from minimally processed data (Goodfellow et al., 2016, Chapter 1). "Usefulness" in this respect is defined relative to the task that the machine learning model is being asked to perform and not necessarily for the purpose of scientific description, explanation, and experimentation. The major problem for using a set of features developed from machine learning is that a researcher must determine what the machine learning model should learn. This task is not as straightforward as it seems. Beyond choosing what should be learned—such as a mapping from a waveform to phoneme labels—there are choices that must be made about how those concepts or mapping should be learned, such as what should be optimized (e.g., how much does the relative magnitude of an error matter?).

A common approach for learning useful features is to use an autoencoder neural network (Goodfellow et al., 2016, Chapter 14). In this type of neural network, the model is trained to take its input, compress it to a smaller number of features, and then reconstruct the input based on the compressed features. For psycholinguistic experiments, autoencoders have been used to learn useful features related to word embeddings (Jones & Brandt, 2019; Vitevitch & Storkel, 2013). However, an autoencoder may not prove to be the best type of network to use for the question of learning features that may reflect human cognition. Indeed, because an autoencoder is learning how to reconstruct its input from its learned features, it will learn structure that is inherent in the input data, as is relevant for reconstructing it and not necessarily as relevant for human perception.

A different approach would be to train a neural network to perform a task that is related to what humans do with speech and language. In doing so, the network would be forced to learn features that are useful for completing the task. This is not to say that the learned features will be the same features that humans use during cognition, but rather, they are features that can be used to complete tasks related to human cognition. Determining the cognitive validity of a given set of features requires experimentation. This is much in the same way that determining whether acoustic correlates of a given speech segment are also perceptual cues requires controlled experimentation. Consider, for example, how $F4$ lowering is an occasional acoustic correlate of flaps but not a reliable perceptual cue in English (Warner et al., 2009; Warner & Tucker, 2017). The features that a neural network learns are analogous to acoustic correlates.

Within the realm of speech perception, three tasks stand out as potentially relevant for a neural network to learn. The first two are phoneme and grapheme recognition. Phoneme and grapheme recognition are common tasks for neural networks used in speech recognition systems (Graves & Jaitly, 2014; Graves & Schmidhuber, 2005; Hinton et al., 2012; Zhang et al., 2016). As potential tasks related to acoustic distance, phoneme and grapheme recognition are not particularly good choices, though. Both phonemes and graphemes are alphabetic targets, not acoustic targets. For that reason, they are not good targets to predict if the goal is to find acoustic features that are relevant for describing acoustic distance. That is, there are non-acoustic components to phoneme and grapheme recognition, such as phonotactics and spelling conventions. It might be argued that, for example, phonotactics do manifest in the acoustic signal at some level like having no English utterances begin with [ŋ] or end with [h]. However, this itself is not an acoustic property of the speech signal, but rather, a statement about the patterning of certain aspects in the acoustic signal and thus not actually acoustic. An analogous argument applies to spelling conventions. The features that the neural network learns could then not reliably be said to wholly relate to the acoustic properties of the speech signal. Thus, the features could not be used to measure acoustic distance proper in the same way that distances measured with a feature set for speech that includes spatial coordinates of the tongue and lips cannot be said to be completely acoustic. While it is likely true that these sorts of information can be useful during the cognitive processing of speech, they are unrelated to determining the acoustic distance between words or sounds in speech.

A third potential task is tracking what Deng and O'Shaughnessy (2003, Chapter 10) referred to as vocal tract resonances. Vocal tract resonances are related to formants, but they are more general and apply to all types of speech sounds. They are the frequencies that would resonate in the vocal tract based on the articulatory configuration for a given segment. For sounds with formant structure like vowels and approximants, they coincide with formants. Vocal tract resonance tracking is less popular than phoneme or grapheme recognition, though Dissen and Keshet (2016) and Dissen et al. (2019) presented some recent deep learning results for formant tracking. Such projects are often described as formant tracking projects, but they are, in reality, vocal tract resonance tracking projects. The advantage of having the neural network learn to predict vocal tract resonance values is this: Because the network is predicting acoustic phenomena, the features it learns should be relevant to acoustic processing. That is, formants are defined acoustically as peaks in the spectrum, so the neural network should be processing its input into more-or-less acoustic features. And, vocal tract resonances are trackable (Deng & O'Shaughnessy, 2003). Vocal tract resonance tracking is, then, a reasonable task to ask a neural network to learn so that it learns acoustic features. It is not necessarily the case that humans literally track formants in speech, to say nothing of all vocal tract resonances. But, human listeners have shown sensitivity to relatively small changes in formant frequencies (Kewley-Port & Watson, 1994), addition to sensitivity to formant trajectories (Nearey, 1989; Nearey & Assmann, 1986). As a machine learning task that is more or less strictly acoustic, then, vocal tract resonance tracking seems to bear some relation to what humans do during perception of vowels and other segments that involve formants, even if formant tracking and vocal tract resonance tracking is not exactly and explicitly what humans do.

4.1.2 Formant tracking

While tracking vocal tract resonances is more involved than tracking formants, it is still related to tracking formants and uses a lot of the same methodology. For this reason, it is instructive to discuss formant tracking approaches, which are more prevalent and well-documented than tracking vocal tract resonances. Detecting formant values requires analyzing the acoustic spectrum for resonant frequencies. Given the bell shape of a resonant filter, determining the central frequency of a formant requires choosing the frequency where the peak of the filter is. That is, a local maximum must be found. This general procedure is known as “peak-picking,” and it is common throughout many fields of science. It is obvious to a human where the peaks are in a spectrum, but it is difficult to write a program that will automatically perform peak picking. In classical models of computing, peak picking might be expressed as a calculus problem of finding the zeros of the derivative of a function, or else of performing a search for values that are higher than their surrounding values (as in the `find_peaks` function in SciPy, Virtanen et al., 2020). For simple spectra, these sorts of algorithms might work well. Speech signals, however, do not have a simple spectrum; rather, a typical speech spectrum has many peaks, and correctly determining the formant frequencies requires choosing the correct peaks.

Historical approaches to formant tracking have relied primarily on analysis using linear predictive coding (LPC). In speech signal processing, LPC analysis is a method that estimates the vocal tract filter as a polynomial. The poles or roots of the polynomial (where the polynomial is equal to zero) can be used to determine the formant frequencies and the formant bandwidths, bearing in mind that the LPC filter was an estimate of the vocal tract filter. Effectively, this method leverages linear algebra to find the peaks in the spectrum.

As is typically the case with signal processing techniques, one must choose the length of the window of analysis in order to track formants over time. This is the case because many signal processing techniques are most amenable to stationary—roughly, “static-state”—signals, and static states in the speech signal exist only over small increments of time (to the extent that they exist at all). As all phoneticians are aware, the choice of this window length incurs a tradeoff between frequency resolution and temporal resolution. Choosing a short time window will allow for greater temporal resolution, but lower frequency resolution, as represented in a broadband spectrogram. Conversely, choosing a longer time window will allow for a greater frequency resolution, but a lower temporal resolution, as represented in a narrowband spectrogram. This tradeoff is part of the Gabor uncertainty principle (Gabor, 1946), itself based on the Heisenburg uncertainty principle Heisenberg (1927). The Gabor uncertainty principle has been discussed directly for time-frequency analysis (inter alia, Benedetto et al., 1992; Hsieh & Saberi, 2016; Parhizkar et al., 2015), among many others. It has also been discussed indirectly in phonetics when describing the differences between broadband and narrowband spectrograms (Johnson, 2012; Zue & Cole, 1979, inter alia,).

Formant tracking techniques usually choose temporal resolution over frequency resolution. In part, frequency resolution is not of extreme importance in formant analysis for traditionally male voices because formant frequencies do not often have such narrow bandwidths that the entire resonant filter could be missed by the coarser sampling. And, harmonics for male voices occur at more frequent intervals due to an, on average, lower fundamental frequency. For traditionally female and children's voices, greater spacing between harmonics can be a source of error when estimating formant frequencies (Kent & Vorperian, 2018) since there are fewer opportunities for a harmonic to be amplified through the resonant filter. Logically, a more granular frequency resolution may help ameliorate this difficulty by at least capturing the general shape of the resonant filter if not the peak, but too coarse a frequency resolution runs the risk of smearing the shape of the resonant filter in much the same way that too low a sampling frequency will not appropriately capture the peaks and valleys of higher-frequency sounds.

This tradeoff leads to a natural difficulty in casting formant tracking as a deep learning problem. Formant tracking can be thought of as a localization problem, where the peaks in a spectrum must be found and localized, and those localized peaks will be connected to a regression model for $F1$, $F2$, and $F3$. However, due to differences in formant spacing for typical male, female, and children's voices, the network may have a difficult time associating a particular peak location in frequency space to a given formant when the region is ambiguous. A potential alternative is to make the uncertainty tradeoff at variable points throughout the frequency spectrum by using wavelets instead of the Fourier transform. The outcome is a multiresolution analysis, which typically results in greater resolution at lower frequencies and less resolution at higher frequencies. Conversely, there is less temporal resolution at low frequencies and more temporal resolution at high frequencies. In effect, each time step in the output of such a representation should contain the fundamental frequency and perhaps some of the next few harmonics, followed by more spaced out frequency components that should result in formant peaks being detected. The network can use this information to determine which formant to assign a localized peak to.

Historical approaches to formant tracking have been more computational than connectionist in nature. For example, the Burg technique as described by Press et al. (1992) is a commonly used method in software like Praat Boersma and Weenink (2020). The Burg method computes linear predictive coding coefficients on a small chunk of audio and then performs pole picking from among the coefficients to determine the local maxima associated with formant frequencies. While this process may involve some steps similar to those that are performed by humans during the process of audition, the algorithm was not designed to reflect certain aspects of human cognition. The input representation is a particular standout in that regard, where the spectrum is represented linearly even though human perception of frequency tends to be more logarithmic in nature as in the mel scale (S. S. Stevens et al., 1937) and the bark scale (Zwicker, 1961). More recent speech features have tried to more closely represent human cognition such as perceptual linear predictive coding coefficients (Hendriks et al., 2004), though they do not appear to be used prominently in the methods that phoneticians currently use for formant tracking. It is also worth noting that the approach from Dissen and Keshet (2016) and Dissen et al. (2019) used features based on linear predictive coding coefficients as input for some models, though they did also use spectrograms as input for convolutional neural network models.

4.1.3 Interpreting neural networks

One problem facing the notion of using acoustic features learned by a neural network is that neural networks are generally considered black boxes. That is, whatever they have learned is difficult to access and describe. It would not be advisable—to the point of being unscientific—to recommend a black-box feature set for scientific analysis and cognitive modeling in place of features that are generally easier to understand like MFCCs. As such, features learned by the neural network would need to be interpreted to be usable. For this reason, there is a significant focus in the present paper on potential methods for understanding neural networks and the features they have learned. In this way, the features that the neural network learns can be more meaningfully compared with MFCCs or other features for phonetic and cognitive modeling purposes. This focus requires a discussion of why exactly it is difficult to interpret a neural network, which will motivate the interpretation methods used in the present paper.

Two principal reasons that neural networks are difficult to interpret are that they are semantically non-compositional, and they represent non-linear functions. Concerning non-compositionality, consider the perceptron given in Figure 4.1 and Equation 4.1. Both of these formats are equivalent approximations of the logical AND function. This function takes in two TRUE/FALSE variables and returns TRUE if both variables are TRUE; it returns FALSE otherwise. There is no way to use the morphological and syntactic composition of these representations to understand that they compute the AND function. This situation is analogous to X KICKED THE BUCKET being a non-compositional idiom whose meaning cannot be apprehended from the constituent parts. Standard neural networks are far more complex than this simple example, so the difficulties incurred by non-compositionality are far greater than this example suggests. Regarding non-linearity, humans seem to have difficulty in intuitively understanding non-linear relationships and often treat them linearly. For example, humans have a tendency to treat exponential functions as linear when reasoning intuitively (Banerjee et al., 2021; Levy & Tasoff, 2017; Schonger & Sele, 2020; Stango & Zinman, 2009). There is also a sentiment prevalent within quantitative fields, where linear systems are said to be easier to understand than nonlinear systems (see, for example, Hastie et al., 2009, Chapter 5).

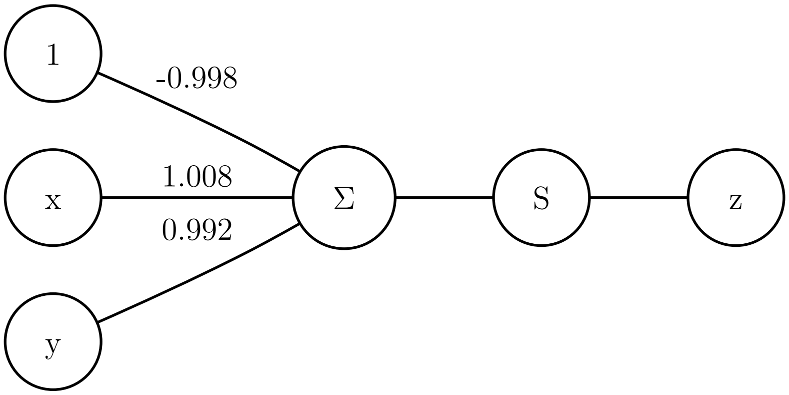


Figure 4.1: Representation of the logical AND function as a simple network with floating-point weights.

$$z = \left(1 + \exp\left(\begin{bmatrix}-1.008 & 0.998 & 0.992\end{bmatrix}\begin{bmatrix}1\\x\\y\end{bmatrix}\right)^{-1}\right)^{-1} \tag{4.1}$$

The non-compositional, nonlinear meaning of neural networks is not unlike the sorts of systems and processes that are studied in phonetics, psycholinguistics, and the social and behavioral sciences generally. As a phonetic example, neural networks are conceptually similar to spectrographic representations of speech. A spectrogram itself is well-defined mathematically, but its mathematical definition does not help a phonetician much in systematically determining what linguistic information the spectrogram represents, nor do the numbers in a spectrogram lend themselves to easy interpretation. Rather, a phonetician may be able to determine acoustic correlates of linguistic information based on knowledge that was derived through experimentation and careful analysis, such as the correspondence between formant values and vowel quality. By gathering systematic scientific knowledge, we create much-needed context surrounding the acoustic measurements that are being made. And, perhaps applying these techniques to analyzing a neural network can provide some level of explanation of what, exactly, a neural network is doing. The conjecture here is that one method of understanding neural networks is to create context by asking appropriate questions informed by domain expertise and answering them using the statistical and analytical methods employed in the social and behavioral sciences. This issue of context as an important factor in interpreting neural networks has also recently been highlighted by Sheu (2020). In a broader sense, cognition is a black box in much the same way that a neural network is.

Previous research has interpreted a variety of aspects of neural networks. Some of these aspects include optimal architectures and generalization (Tishby & Zaslavsky, 2015) and decision explanation for playing video games like *Frogger* (Ehsan et al., 2019). It is also common to visualize the parts of a network that detect features for recognition tasks (Krizhevsky et al., 2012; Zeiler & Fergus, 2014). Word embeddings have also been used to cluster semantically similar image labels to analyze image recognition errors (Dharmaretnam et al., 2021). More closely related to speech and audio processing, some studies have implicitly performed perceptual experiments on their networks by examining how the neuron activations changed according to different types of input. Krug et al. (2018) examined clustering for phoneme and grapheme categories by examining how different phoneme categories affect neuron activation in the network. Palaz et al. (2015) examined the frequency response to different stimuli in the first convolutional layer in their networks, finding behavior similar to a filterbank. Similar techniques have been applied with a more explicitly-stated connection to psycholinguistic work, such as Baayen et al. (2011) examining the activation to various linguistic stimuli in a naive discriminative learning network, and Baayen et al. (2019) examining activation diversity for linguistic stimuli in a linear discriminative network. It is these types of implicit and explicit simulated perceptual experiments that inform the analysis that was performed in the present study.

4.1.4 The present study

The present study trained a neural network to track vocal tract resonances in speech. The input features were scalograms that resulted from multiresolution wavelet analysis, for the previously mentioned possible advantages. The network was trained using the Vocal Tract Resonance Database from Deng et al. (2006). Due to the scientific importance of understanding the features used to represent an object, an attempt was made to interpret the features that the neural network learns. The feature interpretation was framed as a perceptual experiment to find the acoustic correlates of neurons' activation.

Following the feature interpretation, the models from Kelley and Tucker (2021c) and Kelley and Tucker (2021b) were re-fit using the neural network features as the acoustic representation for words. The re-analysis based on Kelley and Tucker (2021c) refit the statistical models used to analyze auditory lexical decision data from the Massive Auditory Lexical Decision data set Tucker et al. (2019). These results speak to how well the neural network's features relate to spoken word recognition, especially regarding the accumulation of acoustic distance over time. The re-analysis of Kelley and Tucker (2021b) focused on re-assessing the suitability of various distance functions to measure the acoustic distance between vectors of neural network features. This analysis involved correlating a variety of distance functions against the mean-pooled human participant responses from a distance rating task. The best correlation achieved using the neural network features was compared to the best correlation achieved with MFCCs to discern which feature set might best represent the sorts of acoustic information that matters for judgments of distance in the human mind.

These results all jointly inform the discussion of which feature set best suits the description of acoustic distance for humans.

4.2 Training the formant tracker

The neural network to be trained was roughly modeled after the convolutional neural network from Dissen et al. (2019). Unlike Dissen et al., the convolutional layers did not include recurrent connections. This choice was made because recurrent connections will make the network much more difficult to interpret. That is, recurrent connections in the network would require the output of the networks and layers to be interpreted in relation to all of the times steps before and/or after the current time step, or else the temporal aspect of the features will need to be ignored. Choosing a network that is made up of strictly convolutional and fully-connected layers avoids having to account for the temporal dimension of the predictions. However, this does mean that the network may struggle to learn how to use previous and future context to make decisions. This is a potential tradeoff sacrificing some amount of accuracy for the sake of interpretability. The network architecture will be discussed in a subsequent section.

4.2.1 Data

The data set used was the Vocal Tract Resonance Database (Deng et al., 2006). It contains extra annotations for a subset of the TIMIT corpus (Garofolo et al., 1993). These extra annotations are peculiar because they provide formant values for sounds that are not typically associated with having formant values, such as fricatives and stops. The data set itself is actually designed to track vocal tract resonances and not strictly formants, so it could be argued that it is not the most suitable data set to work with when only trying to track formant values. However, to maintain compatibility and comparability with previous results from Dissen et al. (2019), this data set was used.

When Deng et al. (2006) created the annotations, they started by using the tracking algorithm described in Deng et al. (2004). The tracking algorithm models the resonances as LPC cepstra poles. The resonance tracks are statistically smoothed, and the prediction errors are modeled with Gaussian mixtures. In effect, the model for the resonances is LPC cepstra poles that are smoothed over time and adjusted for errors based on learned distributions of the prediction error. This is similar to how formants are tracked, though the additional statistical processes adjust the values. After tracking the resonances in the subset of TIMIT, the initial annotations were hand-corrected and interpolated based on values from Deng and O'Shaughnessy (2003, Chapter 10).

While Dissen et al. (2019) used spectrograms as input to the convolutional neural network, scalograms were used instead for the present study. Scalograms are a visual representation of the continuous wavelet transform instead of windowed Fourier transforms. Scalograms provide good frequency resolution at lower frequencies and good temporal resolution at higher frequencies. In effect, there is a gradual shift from a narrowband spectrogram-style representation for lower frequencies to broadband spectrogram-style representation for higher frequencies. More technical information on wavelets and the continuous wavelet transform can be found in Mallat (2009), and a more applied explanation of the scalogram and continuous wavelet transform for speech analysis can be found in Farouk (2018). A manual examination of spectra contained in the temporal slices of the scalogram suggested they were generally less noisy than those of a spectrogram calculated for the same data, and the peaks associated with formant values were generally easier to discern. Ultimately, the scalogram representation was chosen over the spectrogram representation because the cleaner spectra, more clear formant peaks, and the inclusion of pitch and harmonic information at lower frequencies was hypothesized to be more useful to the neural network than the spectral information that a broadband spectrogram would provide.

Note that while scalograms and spectrograms may appear at first glance to be visual representations of the speech signal, this is not necessarily true and is not true in this case. Despite the fact that phoneticians often visualize spectrograms in programs like `Praat` Boersma and Weenink (2020) and interpret them visually, all values in scalograms and spectrograms are acoustic and have no visual interpretation by default. This is the same situation as for representing stereo sounds, where each channel forms either a row or column in a matrix, but that matrix is no more a visual representation than are scalograms and spectrograms. It is only when scalograms and spectrograms are attempted to be visualized that they gain some sort of visual interpretation via a mapping from the acoustic intensity values to grayscale or color representations. If the scalograms or spectrograms were first saved as, for example, a PNG image before being analyzed, they would become a visual format to represent the sound. This is to say that scalograms and spectrograms are not the same as their visualizations. Neither conversion to an image format nor visualization is not taken for this neural network, so the network is still receiving acoustic and not visual information with the scalogram, although the acoustic information is spread over axes representing frequency and time.

The scalograms were computed from the appropriate TIMIT files. The `cwt` function from the `signal` module in `SciPy` (Virtanen et al., 2020) was used to calculate the scalograms. The complex `morlet2` wavelet was used. A 200-length linearly spaced frequency range from 1 Hz to 5000 Hz was used to create the widths of each wavelet using the formula given in the function's documentation, with a width parameter of 6. The other parameter only had to deal with the output type, which was left at the default of `float64` for real-valued wavelets and `complex128` for complex-valued wavelets. Once the scalogram was calculated, it was decimated with a non-overlapping mean window such that each time step was equivalent to 1 ms, rather than 1 sample. This step was necessary in part because the full scalograms produced large files, and it would be difficult to load all of them into memory at once when training the network.

The steps of calculating the scalograms and decimating them can roughly be thought of as a pre-determined convolutional layer on the raw waveform followed by mean pooling. The continuous wavelet transform in `SciPy` is computed as a convolution, and the decimation with the mean operation simply is a mean pooling operation. These steps cannot be easily added to a convolutional neural network, however, because the kernels that a convolutional network learns are not generally large enough to represent the family of wavelets that are used in a continuous wavelet transform. While a custom implementation of the convolutional layers to specifically learn relevant wavelets could be designed, such a project is beyond the scope of the present paper.

4.2.2 Network architecture

The network consisted of a series of interleaved convolutional and pooling layers to perform spatiotemporally localized feature extraction, followed by a series of five fully-connected layers to model connections between the features. The first convolutional layer had a kernel size of 10-by-8 (time-by-frequency), a stride length of 10-by-1, 256 feature maps, and zero-padding of 4 in the temporal dimension. This layer was followed by a max-pooling layer with a size of 1-by-4. The next convolutional layer had a kernel size of 3-by-8, a stride length of 1-by-1, 256 feature maps, and zero padding of 1-by-4. This was followed by a max-pooling layer with dimensions of 1-by-4. The data was then reshaped and permuted so as to conform to the upcoming fully-connected layers so that the features in each time step would be multiplied appropriately with the layers' weights. This section of the network is visualized in Figure 4.2.

There were 5 fully-connected layers with 512, 256, 128, 64, and 3 neurons in sequence. All convolutional layers and fully-connected layers used the ReLU activation function, except for the output layer, which used the identity activation function.

The particular setup for the first convolutional section was to take each millisecond of the scalogram and decimate it temporally to have time steps that corresponded to each 10 ms in the audio to match up with the provided formant values. This is the second time that decimation was applied to the speech signal, where the first was when the scalograms were calculated. While it is possible to have decimated the signal to 10 ms increments from the start, this second decimation process allows for the network to learn what kind of convolutional kernel should be used to perform

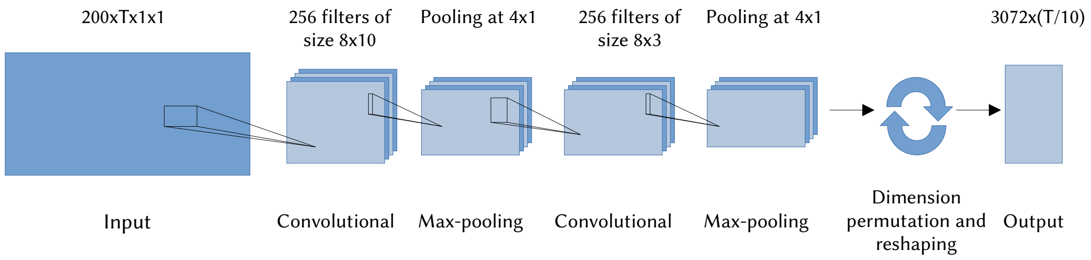


Figure 4.2: Architecture of the convolutional portion of the network. The input begins as a 200 by T by 1 by 1 matrix, ordered as height-width-channels-batch size and where T is the length of the recording in milliseconds. The rectangles in each layer indicate the relative size of the filter or pooling operation in the subsequent layer.

the second decimation, rather than a strict mean-pooling operation.

4.2.3 Network training

The network was created and trained using the `Flux` deep learning library (Innes et al., 2018; Innes, 2018). The network was trained for 100 epochs. Eighteen sentences were held out from the training data to serve as validation data. The model from the epoch with the lowest validation loss is the one that was kept. The optimizer used was the Adam optimizer (Kingma & Ba, 2015), and the parameters were kept at the default values in `Flux`.

4.2.4 Network performance

The results of the network on the test data are presented in Table 4.1, which also contains the results from the convolutional neural network trained in Dissen et al.

(2019). Each column contains the mean absolute regression error of a particular formant or vocal tract resonance, plus-or-minus the standard deviation. The vocal tract resonances are broken down by broad segment category. In every case, the results from the network trained in the present paper were worse than those from Dissen et al. The discrepancy in the results may be due to the network in the present approach not having recurrent connections, which limits its ability to handle context, especially since the convolutional section is not very deep. Whereas, the convolutional network from Dissen et al. did have recurrent connections, which would allow it to better model time-series data. It also used a size-restricted spectrogram as input, and it is possible that spectrograms are better fits to this problem domain than scalograms.

Table 4.1: Comparison between the convolutional neural network in Dissen et al. (2019) and the network trained in the present paper. All units are in Hz, and each entry is presented as mean absolute error ± the standard deviation of the absolute error. Results are separated by speech sound category. The columns for the Dissen et al. (2019) paper come from their convolutional neural network model trained on spectrograms. Whereas, the columns under the "CNN" heading are the results of the convolutional model trained in the present study. Each column is labeled as a formant for parity with previous results, though the predicted values are only truly formants for the sonorant categories and are instead vocal tract resonances for the obstruent categories.

	Dissen et al. (2019)			CNN		
Category	$F1$	$F2$	$F3$	$F1$	$F2$	$F3$
Vowels	53 ± 52	73 ± 74	108 ± 128	66 ± 96	123 ± 60	144 ± 148
Approx.	68 ± 62	111 ± 143	160 ± 187	75 ± 147	220 ± 70	222 ± 266
Nasals	69 ± 66	191 ± 208	158 ± 152	104 ± 249	195 ± 95	275 ± 198
Fricatives	139 ± 118	142 ± 143	167 ± 156	148 ± 168	195 ± 130	167 ± 172
Affricates	174 ± 146	173 ± 144	195 ± 164	166 ± 172	219 ± 147	147 ± 170
Stops	123 ± 102	135 ± 149	170 ± 168	143 ± 159	200 ± 121	182 ± 188

Regardless of the performance, though, the vowel formant tracking performance is not that far off compared to previous work. Additionally, the purpose of this network is not to be a competitive formant tracker, as useful as such a tool could be. Rather, the network is being trained to learn acoustic features useful for formant prediction. And, the network certainly must have learned something related to formants and vocal tract resonances since the results are not nonsensical. Even though the network could theoretically perform better, it should suffice for the purposes of assessing its features for cognitive modeling.

4.3 Interpretation of features

The conjecture from the introduction that understanding a neural network's features necessitates crafting context motivates the present analysis. One way to produce that context is to determine what the acoustic correlates of neural activation are. An obvious choice is frequency since frequency is one axis/dimension of the scalogram fed into the neural network. As an explanation, though, saying that frequency is an acoustic correlate of neural activation is uninformative when all of the information the neural network receives is frequency-over-time information. A more useful and specific explanation would involve determining what frequency components cause a neuron to receive activation. The resultant analysis could be structured to determine which frequency components are acoustic correlates of a specific neuron's activation.

Frequency components can be modeled as simple, pure tone sinusoids, as is well-known from the Fourier transform. For this reason, the present analysis will assess the relationship between acoustic components and neuron activation using pure tone sinusoids. Doing so requires testing the neuron activation levels over a range of frequency values. It is likely that the intensity of a frequency component has an effect on the activation of a neuron as well, but the present analysis will hold the intensity constant and use full-scale sinusoids as a simplification.

The methods and results will be described as a perceptual experiment, with the exception that the subject in the experiment will be the neural network itself. The specific features that will be examined are those in the layer just before the output. In effect, the layers that occur in between the input features and the layer in question are abstracted into an unknown function. The exact structure of the function is the network itself, but as established, the network's meaning is non-compositional regarding its structure. As such, plotting a neuron's activation levels as a function of individual frequency components models the effect of different frequency components on each neuron's activation reflects the overall relationship between the input and the features as the sum of basis functions. This approach is roughly the same as a piecewise linear analysis in functional data analysis (Ramsay & Silverman, 2005), and functional data analysis itself is used implicitly throughout speech science via the Fourier transform when performing spectral and spectrographic analysis. This approach is also similar to Beguš and Zhou (2021), although they used TIMIT sentences instead of pure tones to examine changes in activation in their network.

4.3.1 Methods

The present analysis will be described as a perceptual experiment, with the exception being that the subject in the experiment will be the neural network itself. The specific features that will be examined are those in the layer just before the output. In effect, the layers that occur in between the input features and the layer in question are abstracted into an unknown function. The exact structure of the function is the network itself, but as established, the network's meaning is non-compositional as regards its structure. As such, using plotting a neuron's activation levels as a function of individual frequency components models the effect of different frequency components on each neuron's activation. This approach is roughly the same as functional data analysis (Ramsay & Silverman, 2005), and functional data analysis itself is used implicitly throughout speech science via the Fourier transform when performing spectral and spectrographic analysis.

Overall, treating the network in this manner is very similar to how behavioral experiments are performed with humans. The actual processing and decision-making that a human does during an experiment is a black box. Careful experimentation and modeling allow a researcher to relate specific variables to the human subject's responses and determine.

Materials

Stimuli representing a frequency sweep along the same scale as the scalograms that were used as input data were synthesized. They are scalograms generated based on the frequency scale used to create the scalograms, 1 Hz to 5000 Hz in 200 equally spaced linear steps. Each stimulus was generated as a full-scale sine wave for 1 second, sampled at 16,000 Hz. The sine wave was then treated as an audio file and passed into the same function that made the scalograms used to train the neural network, using the same parameters as before. The result was 200 separate scalograms, each representing a different one-second pure tone along the 1 Hz to 5000 Hz scale.

Procedure

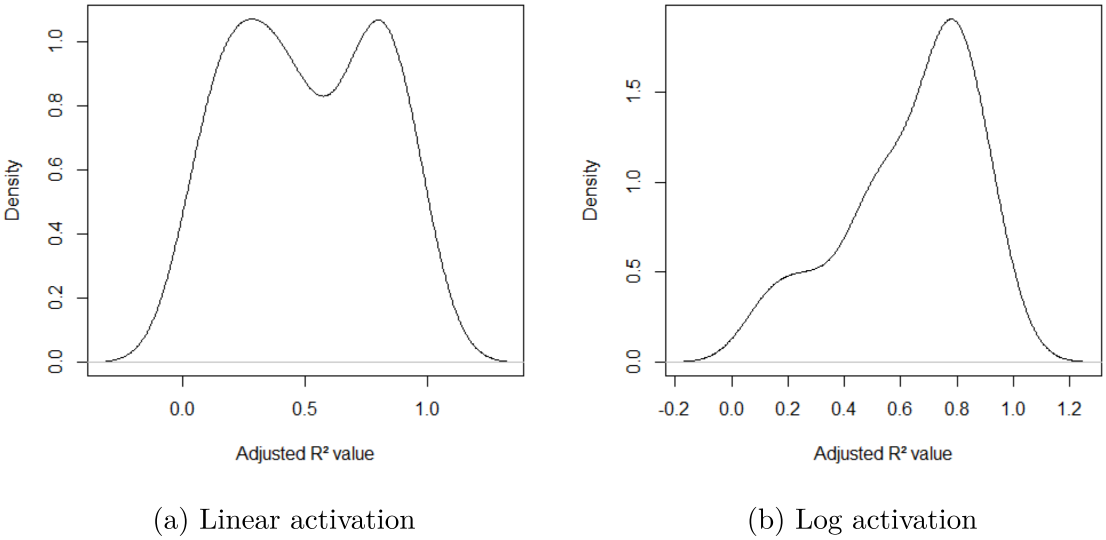
The output layer of the network that predicted formant values from the last set of features was removed, leaving the features as the output. Then, each scalogram stimulus was fed into the network. The output was the activation levels for the features in the last layer of the network. These activation levels were averaged over time for each frequency that was tested, though there was not much variation in the activation values for each time step of the pure tones (as should be the case). A matrix that consisted of the activation level of each neuron to each frequency component was created and written to disk.

When visualized, these values produce a plot that is roughly what Friedman (2001) referred to as a "partial dependence plot." Sheu (2020) classifies this method as a model-agnostic, global method of neural network interpretation. These plots have been identified as suffering from a statistical bias problem when the predictors are not independent from each other (Parr & Wilson, 2020). However, waves of different frequencies are generally orthogonal to each other and thus independent, so this potential problem should not be of great concern for the present analysis.

Each individual neuron's activation pattern was analyzed in the raw form—without smoothing—and in a smoothed form. The logged version of the activation patterns was also analyzed similarly. The smoothing was performed by fitting a generalized additive model (GAM) to the data using the `mgcv` package (v1.8.28, Wood, 2011) in R. The was used as a response variable, and the frequency values were used as the predictor.

4.3.2 Results & Discussion

The average adjusted R^2 for the GAMs fit to the linear activations was calculated (M = 0.51, SD = 0.29), and a density plot of these values can be seen in Figure 4.3a. Thirty neurons received no activation, for which no GAM model could be fit, so they were excluded. In a similar fashion, the average adjusted R^2 for the GAMs fit to the log activations was calculated for the neurons ($M = 0.64$, $SD = 0.23$), and a density plot of these values can be seen in Figure 4.3b. Again, 30 neurons received no activation (the same neurons from before), so no GAM model could be fit for their activation response pattern. These neurons likely had negative values for activation, which the ReLU activation function set to 0. It is possible that they would only respond to acoustic features that are only present with more complex inputs. Such an example might be the presence of multiple frequency components.



(a) Linear activation

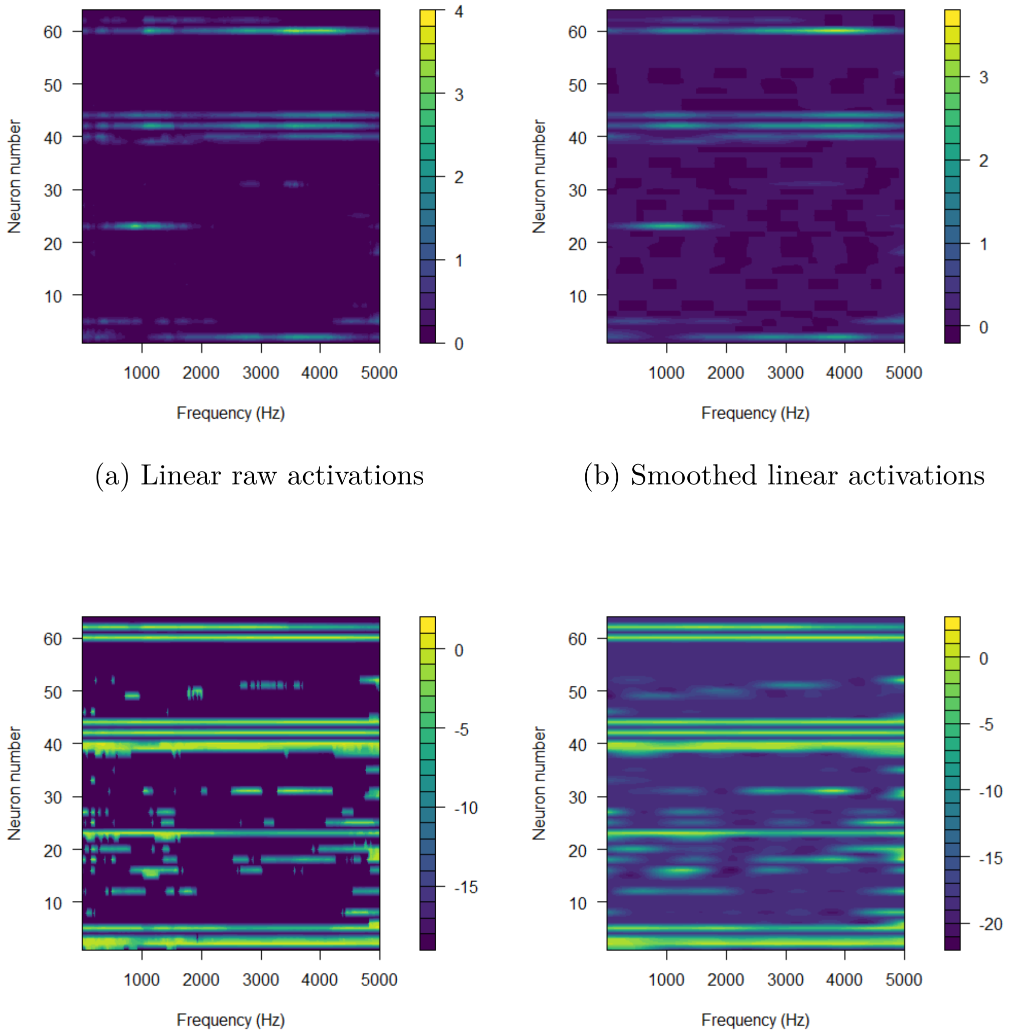
(b) Log activation

Figure 4.3: Density plots for the adjusted R^2 value for the GAMs fit to the linear and log activation values. Thirty of the neurons did not receive activation, so it was not possible to fit GAMs for them or include them in these density plots.

In general, the GAMs fit to the log activations had better explanatory power, and a paired t-test on the adjusted R^2 values indicated that they were significantly different, $t(33) = -3.83$, $p < .001$. These results generally suggest that the activation values can be modeled with reasonable accuracy using smooth functions. That is, the relationship between the frequency components and the activation levels does not have a lot of extreme discontinuities in it. This smoothness is important because it suggests that the neuron's responses are localized to particular frequency regions and are not an uninterpretable combination of frequencies that might occur if the neuron had, for example, high activation for a 1,000 Hz frequency component but low activation for a 1,001 Hz frequency component. Heatmaps of the activation of each of the neurons for each tested frequency can be seen in Figure 4.4. It is possible that some of the neurons that seem to be activated at every frequency more or less respond to the presence of sound, which might be useful as part of a binary detection of stop closures.

There is a distribution of activation for regions that may be relevant for each formant. For example, neuron 23's activation response pattern is reminiscent of a resonant filter at 1,000 Hz. This can be seen clearly in Figure 4.5, which is a plot of its linear raw activation over each frequency component. It is similar to a frequency response plot, showing how the activation levels of the neuron respond to different frequencies in the input signal. The plot depicts something like a resonant filter, which is an expected type of feature to learn when predicting vocal tract resonance values. This filter-like behavior is also observable in Figure 4.4b where the activation increases up to a peak at 1,000 Hz and then more or less gradually decreases. It could be said, then, that a frequency component near 1,000 Hz is an acoustic correlate for the activation of neuron 23.

The activation levels for neuron 42, which had high activation at a number of different frequency positions, can be seen in Figure 4.6. The log activation levels were largely similar in shape and would be redundant to examine. The activation response shows what looks similar to a complex filter made up of 3 resonant filters at roughly around 1,200 Hz, 2,900 Hz, and 3,600 Hz, based on visual inspection. These



(a) Linear raw activations

(b) Smoothed linear activations

(c) Log raw activations

(d) Log smoothed activations

Figure 4.4: Activations of each neuron in the feature set. Dark blue colors represent low activation, while yellow colors represent high amounts of activation. The activations in the smoothed plots are predicted values from a GAM fit on the data. The log space plots had an epsilon value of 1×10^{-8} added to the value before the log function to avoid taking the log of 0.

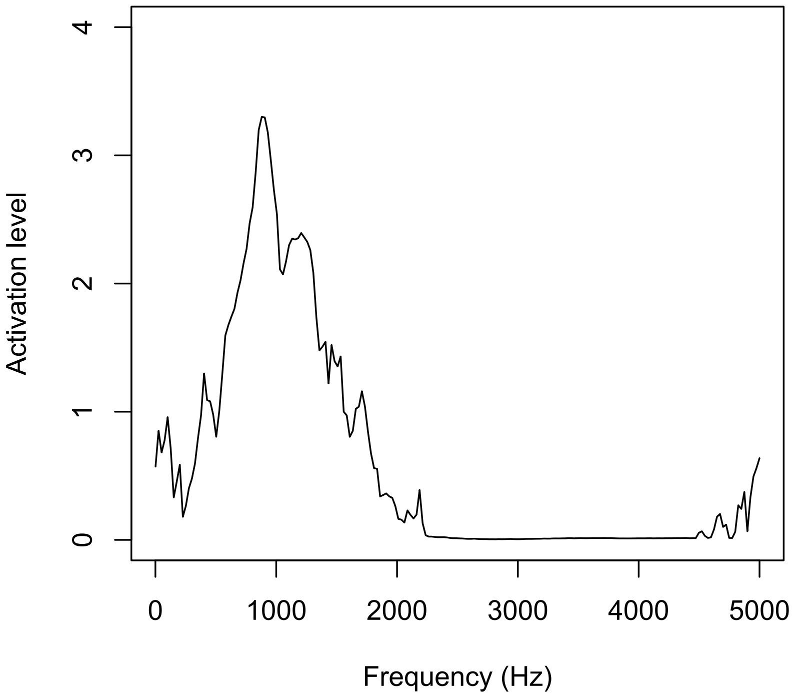


Figure 4.5: Linear raw activation response pattern for neuron 23 in the feature set.

components could be said to be acoustic correlates for the activation of neuron 42.

Some of the neurons seemed to respond in way more suggestive of a consonant than a vowel, though. Consider neuron 60, the linear activations for which are displayed in Figure 4.7. There are notable peaks in the activation pattern. The highest peak is near 3,800 Hz, which close to the spectral maximum for [ʃ] and [ʒ], which at around 3,500 Hz (Johnson, 2012; K. N. Stevens, 1998). That is to say, it seems that this neuron in particular would respond significantly to [ʃ] and [ʒ] and receive a significant amount of activation. In this sense, this acoustic correlate of neuron 60's activation matches one of the acoustic correlates of the post-alveolar

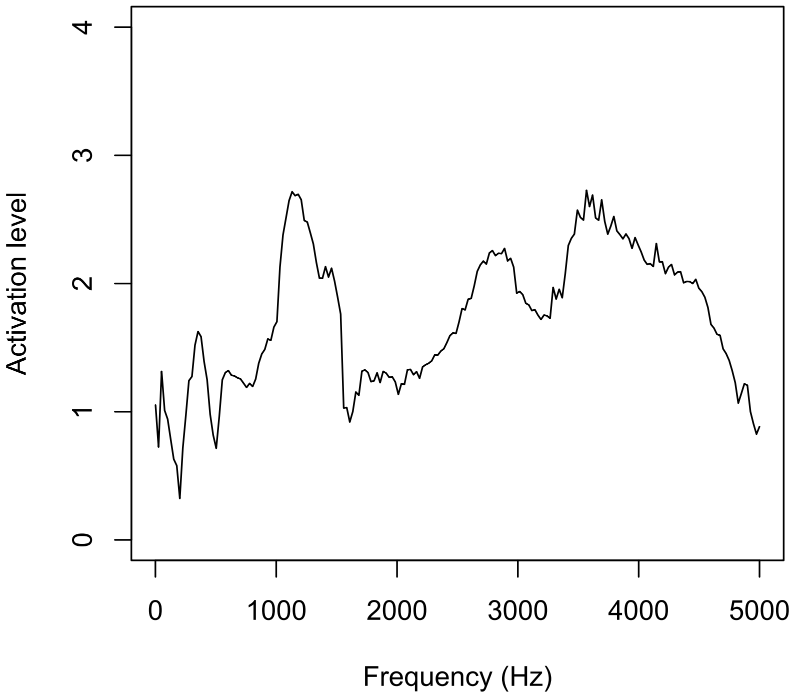


Figure 4.6: Linear raw activation for neuron 42 in the features extracted.

fricatives.

Overall, it appears that many of the neurons in the feature set respond similarly to resonant filters or combinations of resonant filters. Or rather, that they represent the response of a resonant filter on the input sound. It is difficult to pinpoint how these filter-like features might be used when predicting the formant values, especially since many of the neurons show activation for frequencies that are rather high, around 5,000 Hz. With that being said, The response to higher frequencies than might be expected for F3 could be due to the consonants contained in the data set used to train the network. It is also possible that the network uses the presence of high

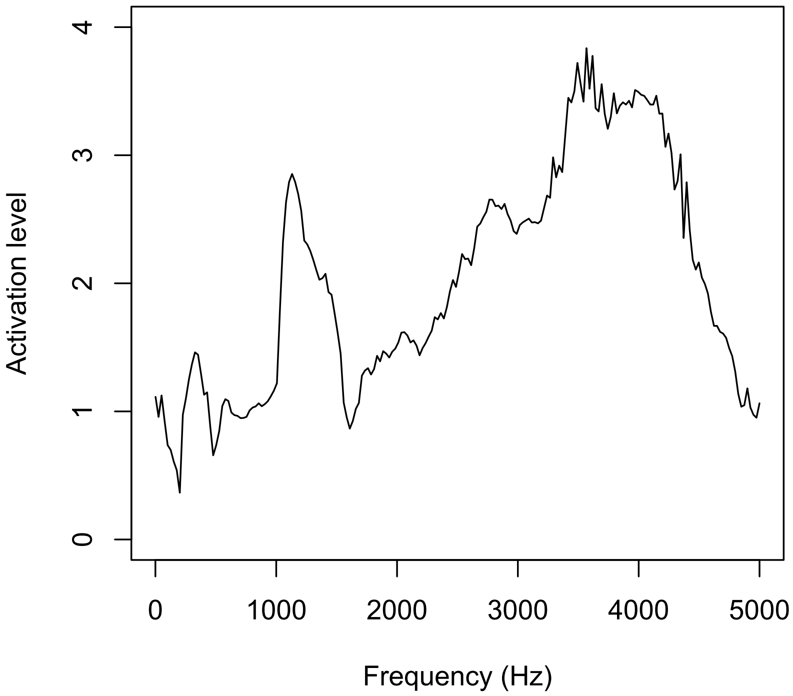


Figure 4.7: Linear, raw activation for neuron 60 in the features extracted.

energy in high-frequency regions to decrease the output for a particular formant. Further simulations would be required to determine whether or not this is the case, however. Regardless, the final set of features that the network learns thus seem to be a sort of filterbank made up of what often looks like resonant filters, and possibly some features that detect the presence of sound (as opposed to a stop closure).

4.4 Feature comparison

Having analyzed the neural network features as similar to resonant filters, the features are now ready to be compared against MFCCs in statistical modeling. The neural network features represent a summary of "important" aspects of the acoustic signal, where the regions of importance were determined in the training process. Whereas, the MFCCs represent a summary of the entire acoustic spectrum. Both of these sets of features define an acoustic space, as it were. The procedures for calculating the neural network features and the MFCCs, then, are the functions or processes by which a slice of the acoustic signal is transformed into the feature space.

Kelley and Tucker (2021b) observed that for these features to be useful in spoken word recognition, they must prove to have sufficient discriminatory power. To demonstrate this need, consider the converse scenario, where a feature set is designed to have low discriminatory power. Such a scenario might be a feature set that consists of one number, the frequency of F1 in the signal. That feature set would fail to discriminate between vowels with similar F1 values, such as [i] and [u], even though F1 is clearly very important for distinguishing vowel qualities from each other. For this reason, such a feature space would not sufficiently separate word pairs like *heap* [hip] and *hoop* [hup]. Or, when such a feature space is used to calculate acoustic absement, the absement value would not reflect differences that humans are sensitive to. An ideal feature set would represent completely and only the differences that humans are sensitive to, no more, no less. The acoustic space, then, forms the dimensions through which a word could be said to "travel" throughout its duration. The question that the present analysis seeks to answer is how much of the acoustic spectrum must be represented within a feature set to reliably distinguish words in a way that aligns with the concept of phonological neighborhoods, where words that are less distinctive in the lexicon take longer to recognize and vice-versa (Luce, 1986; Luce & Pisoni, 1998).

However, it is not enough for the acoustic spaces to merely separate acoustic slices and words. Rather, these features must also be related to cognition in some way since they are intended to explain a portion of cognition. To that end, the present section performs a comparative analysis by using the results from Kelley and Tucker (2021c) and Kelley and Tucker (2021b) to assess how much the neural network features relate to spoken word recognition and speech perception.

4.4.1 Analysis 1

Kelley and Tucker (2021c) fit a series of generalized additive mixed models (GAMMs) to model response latencies in an auditory lexical decision task from the Massive Auditory Lexical Decision (MALD) data set (Tucker et al., 2019). As previously discussed, these models used MFCCs to calculate a variable referred to as "acoustic distinctiveness", which represented how distinctive a particular word was compared to all the other real English words recorded for the experiment (approximately 28,000). Acoustic distinctiveness was calculated as a word's average acoustic absement to all recorded words in the data set using dynamic time warping. As mentioned in the introduction, acoustic absement is the accumulation of distance over time, that is, the summation of the distances between sequential acoustic frames in two separate sounds. It was found that words that had low values of acoustic distinctiveness took less time to recognize than words with high values of acoustic distinctiveness. The present analysis compares the previous results to how well the acoustic distinctiveness variable calculated with the neural network features explains the variation in the data.

Materials

The data used in this analysis came from the MALD database. The data to be modeled in the GAMMs was the response latency in the lexical decision task. MALD is an auditory lexical decision megastudy, with about 28,000 real words recorded by a young male speaker of western Canadian English. Approximately 26,800 of these words were used in the auditory lexical decision task. Each word was responded to at least 4 times from among 231 unique participants who were also native speakers of western Canadian English, for a total of 227,129 data points (including responses to both real words and pseudowords). Stimuli sets were also recorded for two other speakers: a young female and an older male, both of whom are native speakers of western Canadian English. Further details are available in Tucker et al. (2019) on the recording process for the young male speaker, the auditory lexical decision task, and the variables included in the data set.

As part of the MALD project, two other speakers recorded real word and pseudoword productions. One speaker was a young female of similar age to the young male speaker. The other speaker was an older male speaker in his 80s. Kelley and Tucker (2021c) compared using these speakers' productions to using the young male speaker's productions as the acoustic representation for items in the lexicon, as well as combinations of the three speakers' productions. Only words that were common to all of the speakers, resulting in 26,005 unique words to compare each word to. That subset of words is what Kelley and Tucker (2021c) used. For parity and compatibility with previous results, only those words were used when calculating acoustic absement in the present analysis, and only responses to those words were used when modeling the data in the present analysis.

There were a total of 113,675 responses to real words in the data set. Responses faster than 500 ms or before stimulus offset were removed for being implausible responses. For modeling purposes, response time was measured from stimulus offset to help factor stimulus duration out of the response latency values themselves. These response times were then logged. Subsequent analysis found that recordings of *automaton*, *exhalation*, *standoff*, and *sweets* were just recordings of silence, indicating that they had been extracted improperly. Responses to these stimuli were also excluded. These restrictions leave 95,992 responses (84.454% of the original number) for the modeling process.

Procedure

In Kelley and Tucker (2021c), acoustic distinctiveness using MFCCs was calculated using the stimuli from the young male speaker in the experiment. Each word in the data set was first transformed into a series of MFCC vectors using the `MFCC.jl` package in the `Julia` programming language. The window length was set to 25 ms, and the step size for the windows was set to 10 ms. Thirteen coefficients were calculated, and the zeroth coefficient was replaced with the log energy of the frame. The model that performed the best was using just the young male speaker's productions to calculate acoustic distinctiveness. As such, only the young male speaker's productions will be used for calculating acoustic distinctiveness in the present analysis as well.

Acoustic distinctiveness was re-calculated using the neural network features instead of MFCCs. Each word in the data set was pre-processed into scalograms using the same function as was used for the neural network. These scalograms were then fed into the neural network, and the neural network features were stored for each time step. These features were taken as the acoustic representation of the words, just as the sequence of MFCC vectors was taken as the acoustic representation of the words in Kelley and Tucker (2021c). Acoustic distinctiveness was then calculated by using the `distinctiveness` function in the `Phonetics.jl` package (Kelley, 2020). The interior dynamic time warping function was set to use Euclidean distance to compare time steps to each other, as recommended in Kelley and Tucker (2021c). This acoustic distinctiveness method could then be used in the GAMM.

Results

The results from the reanalysis were equivalent to those in Kelley and Tucker (2021c). The only change of real import was that the fREML values were different. The fREML value is an indication of model fitness that rewards prediction accuracy and penalizes model complexity. Lower values are better. The GAMM using the acoustic distinctiveness values from MFCCs had an fREML that was lower by 2080, which is a greater difference than between any of the models using the various methods of calculating acoustic distinctiveness in Kelley and Tucker (2021c), suggesting that it is categorically better to calculate acoustic distinctiveness using MFCCs than with the neural network features as learned here. The GAMM using the MFCC-based acoustic distinctiveness also had a higher adjusted coefficient of determination (adjusted $R^2 = .28$) compared to the model using the neural network features (adjusted $R^2 = .24$).

Additionally, the neural network features had a lower correlation with item duration ($r = .58$, $p < .001$) compared to the MFCC-based acoustic distinctiveness with a value of ($r = .89$, $p < .001$). Kelley and Tucker (2021c) acknowledged a concern about how statistically distinguishable acoustic distinctiveness was from item duration. But, the fact that the version of acoustic distinctiveness calculated with the neural network features was still a significant predictor in the model suggests that there is an important distinction between acoustic distinctiveness and item duration.

Additionally, the acoustic distinctiveness values calculated with the neural network features and with the MFCCs were highly correlated with each other ($r = .74$, $p < .001$), suggesting that they are indeed quantifying similar phenomena.

In general, these results suggest that there is a greater correspondence between spoken word recognition processes and acoustic representations of the entire spectrum than with representations of selected features.

4.4.2 Analysis 2

The next question that needs to be addressed regarding the neural network features is whether they correlate better with human judgments of acoustic distance than do MFCCs. In Kelley and Tucker (2021b), an experiment was performed where participants rated the distance of synthesized vowel pairs on a scale of 1 to 7. These ratings were subsequently mean-pooled across each item. Then, a variety of distance functions were calculated between each of the vowel pairs that participants heard, with each vowel being represented as an MFCC vector. The correlations were checked between each of the distance functions and the pooled participant ratings. Euclidean distance was found as the best option, with near-optimal performance among the tested functions and the ability to be calculated faster than most other distance functions. To better interpret how well the neural network features mimic human performance, the correlation analysis can be repeated.

Materials

The same synthetic vowels from Kelley and Tucker (2021b) were used. These vowels were created via Klatt synthesis in `Praat` (Boersma & Weenink, 2020). Each vowel pair was synthesized based on the monophthongs in the Hillenbrand et al. (1995)

data: [i], [u], [ʌ], [ɔ], [ʊ], [ɛ], [æ], [ɑ], and [ɪ]. The $F1$ and $F2$ values were randomly sampled from a multivariate Gaussian distribution. Each vowel category had its own distribution, and the mean vector and covariance matrix were determined based on the mean F1 and F2 measurements in the data for the male speakers.

A vowel pair for any given vowel category was created by first sampling two sets of $F1$ and $F2$ values from the corresponding Gaussian distribution. Four random noise values were then sampled from a univariate Gaussian distribution with a mean of 200 Hz and a standard deviation of 50 Hz, and these values were then added to each of the sampled formant values. These noise values helped to prevent the vowel pairs from being too clustered and thus indistinguishable in the rating task. The formant values were then used to perform the Klatt synthesis with the `Create KlattGrid form vowel` function in `Praat`, with other parameters being held constant. A table of these values can be seen in Table 4.2.

Table 4.2: Constant values for the `Create KlattGrid from vowel` function in `Praat`.

Parameter	Value
Duration	0.5 s
Pitch	99 Hz
B1	50 Hz
B2	50 Hz
F3	3000 Hz
B3	100 Hz
F4	4000 Hz
Bandwidth fraction	0.05
Formant frequency interval	1000 Hz

The vowels that were synthesized for Kelley and Tucker (2021b) had a sampling rate of 44,100 Hz, but the functions used to create the scalogram expected the sampling rate to be 16,000 Hz. As such, the stimuli were downsampled to 16,000 Hz using SoX (v14.4.2, Bagwell, 2015). Each sound file was then converted to a scalogram using the same functions as before, and the scalograms were then given to the neural network to extract the features from. The features were then mean-pooled across time to produce a single feature vector for each stimulus. Since the synthesized vowels were stationary signals and the neural network did not have recurrent connections, mean-pooling across time should not cause any significant information loss. These feature vectors could then be used to replicate the correlation analysis from Kelley and Tucker (2021b).

Procedure

Following Kelley and Tucker (2021b), a variety of distance functions compared with each other. These distance functions were based on the p-norm from linear algebra, and the algebraic form of the distance function for two acoustic vectors x and y is shown in Equation 4.2,

$$d_p(x, y) = \left(\sum_{i=1}^{n} |\chi_i - \psi_i|^p \right)^{\frac{1}{p}}, \tag{4.2}$$

where χ_i is the i-th element in the vector x, ψ_i is the i-th element in the vector y, and p a variable that is greater than or equal to 1 that changes how the distance function is calculated.

To find which distance function best fit the data, a variety of values of p were searched. The values ranged from 1 to 100 and increased in increments of 0.01. For each value of p, the distance values were calculated, and the correlation between the distances with the neural network features and the pooled human judgments was calculated.

Results

The results from this analysis mimic those of Kelley and Tucker (2021b), where all of the distance functions compared had a relatively high correlation, with a well-defined peak. A plot of all of the correlation values achieved is presented in Figure 4.8. The highest Pearson correlation between the human judgments and the distance calculation using the neural network features was achieved when $p = 1.6$ ($r = .701$, $p < .001$). The 2-norm was only slightly worse ($r = .700$, $p < .001$). In practice, such a small difference is unlikely to be meaningful. Note as well that the overall range of correlation values is remarkably small, ranging from 0.660 to 0.701. This reflects results from a small range of differences seen in Kelley and Tucker (2021b).

Regardless, the human judgments correlate more highly with Euclidean distance calculated on MFCC vectors ($r = .88$, $p < .001$), and with Euclidean distance on F1 and F2 values ($r = .81$, $p < .001$). However, human judgments correlate less highly with F1 ($r = .66$, $p < .001$) and F2 ($r = .60$, $p < .001$) individually compared with the neural network features.

4.5 General discussion

The neural network features performed decently when used as part of modeling human behavior in a lexical decision task. Of particular note is how they performed when used to calculate acoustic distinctiveness. Since acoustic distinctiveness did not correlate nearly as much with duration when using the neural network features, it demonstrates that the effects of acoustic distinctiveness from Kelley and Tucker (2021c) did not depend merely on the high correlation between acoustic distinctive-

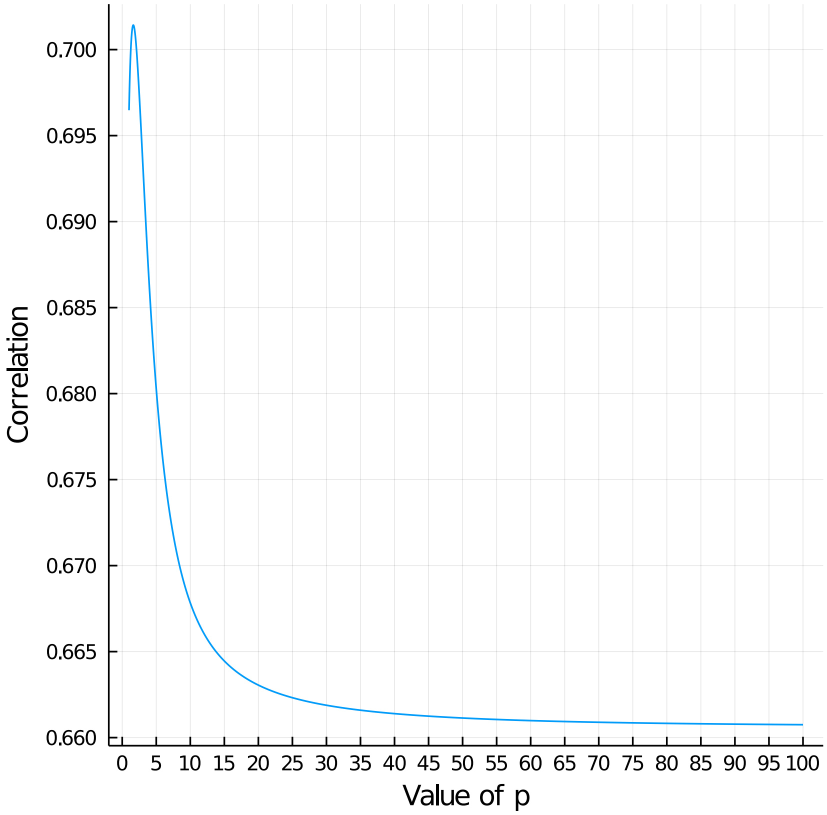


Figure 4.8: Correlation between by-item mean distance ratings and choices of p for the p-norm distance function calculated on the neural network features. Note that there is no data associated with values less than 1.

ness and duration. Rather, it suggests that while a word's acoustic distinctiveness depends in part on duration, acoustic distinctiveness cannot be completely factorized into item duration when using certain acoustic representations. However, using MFCCs still provided a better fit to performance. More work would be required before being able to recommend an alternative feature set to MFCCs.

As it appears, though, the neural-network-learned features did not fare better than MFCCs in terms of relating to human perceptions of acoustic distance. This is somewhat surprising given that the network's features were designed to predict formants (and vocal tract resonances) and the assessment of the features was based on human judgments of vowels, which themselves have strong formants. However, it merits consideration that the data set used for training contained vocal tract resonances for all types of English speech sounds.

Deng and O'Shaughnessy (2003) laid out a theoretical description of vocal tract resonances, which are more closely related to articulation than acoustics. Vocal tract resonances are the resonant frequencies that a particular vocal tract configuration would produce. In the case of formants, these resonances are clear in the acoustic signal. For other speech sounds, though, the resonances do not readily appear in the acoustic signal, and part of the analysis in Deng et al. (2004) was on calculating these hidden aspects of the signal. There has not been much uptake of this concept within phonetics and speech science. In part, this may be due to a relative lack of obvious application in discussing what the resonant frequencies of the vocal tract qua filter would be when articulating a sound such as [s]. For [s], there may be some sort of resonant-like frequency in the spectral maxima that occur, but these do not coincide with the vocal tract resonances in the data. Indeed, there are also better, more easily measured discriminating features like spectral moments (Jongman et al., 2000). This is not to say that there is no possible application of the concept of vocal tract resonances in phonetics. But, it does highlight a problem endemic to using the data set of vocal tract resonances from Deng et al. (2006) as the training data for an acoustically-driving feature set: It doesn't strictly contain observable acoustic values. As previously discussed, these values are based on vocal tract models from Deng et al. (2004) and treated as hidden or latent variables that theoretically can be deduced from the acoustic signal. Moreover, it is not an ideal data set for training a literal formant tracker since it includes more than just formant data. Machine learning systems for formant tracking such as Dissen and Keshet (2016) and Dissen et al. (2019) would need to be trained specifically on only the vocal tract resonances associated with vowels to create an actual formant tracker.

It is also worth thinking more about the resonances that were predicted for non-vocalic speech sounds being more clearly acoustico-articulatory than just acoustic properties of the vocal tract. For this reason, the features the network learned may not be strictly acoustic and may reflect some sort of articulatory information transduced from the speech signal. In this sense, the network would be learning to predict characteristics of the vocal tract filter rather than characteristics of the speech signal. The vocal tract resonances do not surface in the signal, so learning the resonances to predict for a sound like [s] has more to do with the network predicting tongue position via formants rather than characteristics of the acoustic signal itself.

Additionally, formant tracking itself, let alone vocal tract resonance tracking, is something of an odd problem for convolutional neural networks. Traditional convolutional neural networks have alternating sequences of convolutional layers and pooling layers. By using the pooling layers, the network gains a property known as translation invariance (Goodfellow et al., 2016, Chapter 9), wherein the location of a detected feature does not affect the output. This is a useful property for networks that perform tasks like face detection or image classification. This property is detrimental to a task like formant tracking, though, where the location of features—like spectral peaks—is crucial to determining the correct output. It is rather remarkable that the results with convolutional networks in Dissen et al. (2019) performed well at all considering that they used pooling layers in their network. Of note, however, is that their convolutional layers included a form of recurrence, allowing their network to incorporate previous context into their predictions. Recurrent connections were not used in the network in the present analysis so that the network could be interpreted in a more straightforward manner, but it may be the case the recurrent connections are important for accurate performance in formant or vocal tract resonance trackers.

It also seems that using the continuous wavelet transform to create scalograms instead of using a windowed Fourier transform to create spectrograms did not constitute a significant change for the neural network. While spectrograms were not tested as a possible input type for the network presented in the current work, the scalogram input type did not seem to significantly affect the task in such a way that the network had remarkably better performance than other networks. It is not possible with the data at hand to restrict the worse performance with the current network to the use of the scalograms, but it seems safe to say that they have a neutral effect on model performance at best. This result is somewhat surprising because the narrowband harmonic information contained in the lower frequency region should be useful for the network to account for formant spacing variation in different voice types. Future work on neural-network-based formant tracking that is not focused on constraining the model to be more interpretable should test scalogram representations further and against spectrographic and LPC representations.

Regarding the features, it seems that the neural network learned features that roughly corresponded to the results of resonant filters. These features are particular to the exact neural network fit here, and features learned as part of a different task, network architecture, and/or data set would very likely be different. However, it must also be considered that the perceptual question asked of the neural network's features would likely produce behavior that looks like resonant filters. In particular, by examining the activation levels for different frequency components, the output was almost assured to look like a resonant filter. As the frequency sweep approached the frequency or frequencies that a neuron was trained to respond to, the activation would naturally increase. The activation would subsequently decrease thereafter, creating the bell-like shape that is characteristic of resonant filters. Had the perceptual question been framed along the lines of what particular vowels cause excite individual neurons and presented synthetic vowels to the network, it would be reasonable to expect that certain neurons might look like vowel detectors. This possibility highlights the importance of choosing an appropriate perceptual question to ask the network.

As such, it is reasonable to expect that the results that come of interpreting a neural network are likely to be intricately tied to the context created to give its behavior meaning, including the data it was trained on and the task it was asked to perform. That is, conclusions that come from interpreting a neural network based on experiment simulation will reflect the nature of the stimuli presented to the network, analogous to task effects that are seen in behavioral experiments with humans. Future work with formant tracking networks should attempt to re-create the network features as a filterbank with classical computing and signal processing methods. The discrepancy between the network's output and the re-created filterbank could then be assessed, especially with input stimuli that are more complex than pure tones. In so doing, the filterbank analysis will be validated against the network's performance, though other possible interpretations of the features would not be ruled out. Certainly, neuron 60 showed performance that, while still visually similar to a resonant filter, is also likely better classified as responding to the spectrum of [ʃ].

Assessing neural networks' features using a simulated experimental paradigm would seem to give results that are linked specifically to the experimental task. That is, the results are conditioned on the question a researcher has. This is not necessarily bad; it may well be the case that a satisfying answer to the research question can be found using methods within the domain the network is being used—phonetics, in the present case. However, this should be borne in mind when attempting to generalize the results of an analysis since there is an enormous number of connections between neurons and possible statistical interactions in a neural network.

The relationship between the simulated experiment, the domain of inquiry, and the resultant analysis is to be expected. The experimental analysis technique is meant to create meaning for the neural network. Meaning is created with context, and the experimental task, domain, and analysis are the context created for the neural network. The present explanation and analysis may be unsatisfactory for those researchers who do not have phonetic questions about neural networks. But, the idea of using a simulated experiment should be applicable to other questions that can be asked about a neural network (as indeed it has from previous studies, e..g, Krug et al., 2018; Palaz et al., 2015; Zeiler & Fergus, 2014). For those questions about networks that cannot reasonably be addressed through simulated experimentation, the general process of creating context that affords an interpretable explanation is still applicable.

The results from the present analyses are disparate but have a unified implication for calculating acoustic distance: Differences in all regions of the acoustic spectrum contribute to the perception of acoustic distance. This particular finding is apparent in the re-analyses. The neural network features that represented a collection of more localized acoustic features fared worse as components of a cognitive model than did MFCCs, which are a global representation of the acoustic spectrum. A similar observation was made in Kelley and Tucker (2021b) when acoustic distance calculated using F1 and F2 did not correlate as highly with human judgments than acoustic distance calculated with MFCC vectors. Compared to the formant values, though, the neural network features seem to represent features that are more localized in the spectrum for specific speech sounds. Formant values are localized features, but they can be used to describe vowel qualities generally. Whereas, many of the neural network features are likely irrelevant for certain classes of speech sounds. For this reason, it seems to be the case that human listeners attend to the acoustic spectrum more generally rather than to just a select few regions of the spectrum. This conclusion dovetails with findings from Ito et al. (2001) and Nenadić et al. (2020) that vowel identification is not insurmountably hindered when one formant is missing.

The results from the two reanalyses suggest that the distance between the neural network features does not more closely match human judgments of acoustic distance than does the distance between MFCC vectors or the distance between formant pairs. This finding would seem to indicate that summaries of the entire acoustic spectrum—as represented by MFCCs—are a better representation of the acoustic signal insofar as to allow for the separation between different speech sounds to reflect cognition. That is to say that the distance between speech sounds in an acoustic space represented by MFCCs is more closely aligned with how humans perceive and rate distance than is the distance between more selective acoustic spaces like formant spaces or features learned to map to formant values.

These results suggest, then, that in the moment-to-moment process of a lexical candidate receiving activation, activation is accrued based on how well the momentary of matches the listener's expectations. It would not necessarily need to be the case that these expectations are literally acoustic, but they must have some acoustic instantiation that can be expected in the speech signal. Using Euclidean distance in this way does weight all regions of difference equally, which may not necessarily be ideal. It is worth noting that full-spectrum analysis does not decouple source characteristics from filter characteristics in the representation. Taken to its logical extreme, this could suggest that a listener's expectations of prosodic and suprasegmental information are epiphenomenal or otherwise baked into acoustic representation. Such an analysis is not supported by the results here and was not tested in the experimental design.

That a summary of the acoustic spectrum fared best in perceptual validation is suggestive of the acoustic space in which segments and words are distinguished. It would seem that the acoustic space itself is representative of the whole spectrum. In effect, listeners appear to use all of the acoustic spectrum during speech perception and spoken word recognition. While some aspects of the acoustic signal are certainly more important for distinguishing certain sounds from each other (like formants for vowels), listeners make use of all available information. Such a finding is in agreement with research on cue weighting for spoken word recognition from Redmon (2020). Within the cue weighting paradigm, unimportant components of the acoustic signal—e.g., individual cues—can be weighted less than more important components of the acoustic signal, but all components contribute to some degree to speech perception, speech processing, and spoken word recognition.

It is also worthwhile to reflect on the cognitive plausibility of MFCCs. They have fared better than the features that were learned with the neural network, and they performed better than $F1$ and $F2$ in Kelley and Tucker (2021b). It may well be the case that MFCCs provide a decent computational match to how the human mind is processing sound. The mind may not necessarily be performing all of the steps used in calculating MFCCs—such as a discrete cosine transform. But, the way that the coefficients in an MFCC vector interact with each other may indeed reflect a summary of how the mind is processing sound. This is clearly a postulation, but the constellation of the evidence in Kelley and Tucker (2021c), Kelley and Tucker (2021b), and the present research seems to point in this direction. This postulate should not be shocking, either, since the mel scale was designed to reflect logarithmic perception in humans S. S. Stevens et al. (1937). Mermelstein (1976) also found that acoustic distance between words calculated using MFCCs related well to human perception. For these reasons, acoustic representations that involve the mel scale will likely bear a considerable relation to perception by virtue of the mel scale being perceptually informed.

4.6 Conclusion

The feature set one uses to describe the speech signal must generally allow for the speech events that a researcher is interested in to be separated, to the degree that such separation is possible. It is a boon when the features used are also interpretable and/or have a demonstrated relationship to human perception. To the extent that MFCCs are an industry standard for speech recognition, it could be said that they satisfy some of these desiderata. There is a reasonable enough separation of phonemes and graphemes as categories of speech events that automatic speech recognition is possible when using MFCCs as input. They also bear some relation to human perception by being based on the mel scale. They are, however, difficult to interpret in a way that is meaningful to phonetic description, theory, and methods.

A similar remark could be made about the neural network features here, though the evidence for having met these goals is thinner. They related to human perception in the reanalyses of previous experiments. They also formed a reasonable basis on which to calculate acoustic distinctiveness, which itself speaks to the separability goal. Due to the black box-esque nature of neural networks, the features are also difficult to interpret. Via speculative experiment simulation, a coarse description of what the features represented was achieved. Yet, it would be difficult to recommend the use of these features in a broad range of scenarios due to the network not achieving performance nearly as good as those of Dissen et al. (2019), as well as the difficulty in reconciling the differences between vocal tract resonances and formants in the training data.

Nevertheless, there is still some evidence here that neural networks learn features that may be related to human cognition when modeling knowledge that humans have tacitly. Treating the neural network activation levels functionally (as in “functional programming” or “functional data analysis”) led to some level of interpretability. It was also clear though that the context provided by the analysis methods and prior domain knowledge was instrumental to understanding the numbers produced by the networks. As in all things, context is crucial. Without it, the neural network weights and activations would just be an ocean of numbers.

Chapter 5

General discussion and conclusion

The present dissertation ultimately sought to find a consistent method by which to acoustically compare words. This type of comparison is of great importance for research into spoken word recognition, which has often invoked the concept of sound similarity without analyzing actual sound. The preceding studies were designed to collect data to algorithmically summarize. The specific parts of spoken word recognition that were assessed were how well acoustic absement relates to the concept of phonological neighborhoods, how the spectra of two sounds should be compared, how sensitive humans are to duration differences between sounds, and what sorts of acoustic features are important when comparing sounds. The answers provided are, of course, preliminary, but they sketch the human behavior that must be abstracted over to create a computational method to compare sounds.

Principally, words must be compared with each other over time. While this is, in some sense, obvious, it is not at all trivial to accommodate this fact. Even when finitely approximated, few words will have a close enough duration to be directly compared with each other, let alone comparing individual pronunciations of words. It is for this reason that elastic algorithms like dynamic time warping are so useful since they can handle these types of durational discrepancies. The relationship between the cost value of dynamic time warping and perception was demonstrated in Chapter 2, where acoustic distinctiveness had a strong relationship to reaction time. It is also clear that humans must also have some sort of elastic tolerance for the duration of individual speech sounds, given how variable segment duration is during speech production. The results from the second experiment in Chapter 3 also directly speak to this fact.

But perhaps the most important aspect of all of these studies is the notion of acoustic space. The variables that are chosen to represent the acoustic spectrum are of incredible importance to the results from acoustic distance comparisons and the calculation of acoustic absement with dynamic time warping. The reason they are so important is that the choice of acoustic features dictates the distance between acoustic slices of words, which in turn dictates how acoustic distance will accumulate over time. If too specific a set of features is chosen, irrelevant information will not be filtered out and slight variations in production will cause huge increases in distance. Whereas, choosing too vague or incomplete a set of features will cause important acoustic differences to not propagate into the distance calculations. In this sense, the distance values, absement values, and alignment between words are simply measurements within the feature space. And, words are simply time-series objects that exist in the acoustic space. For these reasons, the choice of an appropriate feature space is paramount. The results from all of the chapters combined have suggested that, of the tested options, Mel frequency cepstral coefficients (MFCCs) correlated highest with human judgments of distance and were the strongest predictors of response latency in auditory lexical decision. These results are not divorced from human cognition either. To the extent that humans perform acoustic comparisons between the incoming signal and words competing for activation, the acoustic features that the mind derives from the speech signal will already dictate a large part of what words could possibly be recognized during and after audition.

The remainder of this chapter presents a summary of the results from previous chapters, the tuned dynamic time warping algorithm, possible future refinements for calculating acoustic absement, and possible applications and implications of these results for phonetics and speech communication.

5.1 Summary of results

This section presents a high-level summary of the major results from each chapter.

5.1.1 Chapter 2

Chapter 2 proposed the use of dynamic time warping as a way to measure acoustic differences between words. The output of dynamic time warping was characterized as the accumulation of acoustic distance over time (or word duration), which is the quantity of acoustic absement. Acoustic absement was calculated between all words in the Massive Auditory Lexical Decision data set (MALD, Tucker et al., 2019). A word's average absement to all words was taken as an indicator of its "acoustic distinctiveness", which was used as an index of lexical competition. High values indicated that a word was very acoustically distinct from all words in the lexicon and should have few competitors, while low values indicated that a word was acoustically less distinct from all words in the lexicon and should have many competitors. Acoustic distinctiveness was found to be highly predictive in regression modeling of response latency in the auditory lexical decision data in MALD. It also correlated highly with word duration, which was likely due to how time factors into the calculation of absement.

Subsequently, dynamic barycenter averaging (Petitjean et al., 2011) was used to compute acoustic representations of words that were averaged over multiple speakers in MALD. Acoustic distinctiveness was re-calculated using these averaged representations and the regression model was re-fit. Each time, acoustic distinctiveness was a highly significant predictor of response latency. These results suggested that acoustic absement calculated with dynamic time warping had a strong association with speech processing and spoken word recognition.

5.1.2 Chapter 3

Chapter 3 sought to refine certain aspects of the dynamic time warping process. The goals were for acoustic distance comparisons to reflect how humans judge the acoustic distance of speech sounds and for time-alignment of words to reflect how sensitive humans are to durational differences in acoustic information. Experimental data were collected in support of these goals. The first experiment was a distance rating task with synthetic vowels with varying quality but constant duration. The second experiment was a reminder discrimination task with synthetic vowels of varying duration but identical quality.

The rating task results suggested that many of the distance functions tested had a very high correlation with human judgments. It was ultimately decided that Euclidean distance provided the best compromise between reflecting the human judgments, familiarity of distance function, and computational efficiency. The reminder discrimination task suggested that the just noticeable difference threshold for vowel duration is approximately 30 ms (rounded up to the nearest 10). The distance results were factored into the dynamic time warping function via the function used to compare time steps to each other. Whereas, the duration discrimination results were incorporated by enforcing a radius in the time dimension that specified what time steps could be compared with each other.

5.1.3 Chapter 4

Chapter 4 was designed to test different styles of acoustic representation as input to the dynamic time warping function. The first style was a general summary of the acoustic spectrum, which took the form of mel frequency cepstral coefficients (MFCCs). The second style was a set of features learned by a neural network trained to predict vocal tract resonances from speech data. After interpreting the neural network features, this second style of representation was assessed as a collection of features localized to specific regions of the acoustic spectrum.

The two styles of representation were used to recalculate acoustic distinctiveness and repeat the regression analysis of Chapter 2. The summary style features fit the auditory lexical decision response latencies better than the localized features. These results suggested that acoustic distance as relates to perception is affected by the entire acoustic spectrum, rather than specific local features. It was also found that acoustic distinctiveness calculated with the neural network features correlated only moderately with word duration, in contrast to the high correlation between MFCC-based acoustic distinctiveness and word duration.

5.2 An algorithm for acoustic absement

The dynamic time warping tuning results from the previous chapters are presented algorithmically in Algorithm 1, where W_1 and W_2 are two words represented as sequences of MFCC vectors (strided at 10 ms), and $d(\cdot, \cdot)$ is the Euclidean distance function (or 2-norm). Note that arrays are indexed at 1, and ranges in for loops are inclusive of both the minimum and maximum values. The initialization of the array to be filled with ∞ and to be one column and row larger than would be needed for a comparison of W_1 and W_2 serves to obviate some nuances in indexing and bounds checking that would otherwise be needed. This approach makes the presentation of the algorithm cleaner, but it is not obligatory, and traditional bounds checking could also be used. The condition on Line 12 represents the duration constraint derived from the results of the second experiment in Chapter 3. It also forces the algorithm to compare any remaining time steps in W_2 outside the 30 ms radius to the last time step in W_1, which was discussed in Chapter 3. It is worth noting that the radius is a heuristic method of enforcing the 30 ms duration constraint. It is theoretically possible for a time step t to be stretched from $t - 3$ to $t + 3$ in the other sequence for a total of 7 time steps or 70 ms. Doing so would force other time comparisons to be shorter to compensate for the longer comparison, as discussed in Chapter 3.

The sort of heuristic with the radius is useful because it requires only a small modification to the dynamic time warping algorithm, and it is non-trivial to change the algorithm to find an optimal warping path for which a time step can only be stretched up to three times. If future research determines that more precise timing is required, a substantial modification to the dynamic time warping and backtracking algorithms would be needed. There is still an outlying empirical question, though, of what the distribution of warping lengths for time steps in word comparisons is.

Algorithm 1 Algorithm for calculating acoustic absement

```
function AA($W_1$, $W_2$)
    if $W_1$ is longer than $W_2$ then
        Swap $W_1$ and $W_2$
    end if
    $T_1$ := the number of time steps in $W_1$
    $T_2$ := the number of time steps in $W_2$
    Initialize a $(T_1 + 1)$-by-$(T_2 + 1)$ matrix $D$
    Set all values in $D$ to $\infty$
    $D[1, 1] := 0$
    for $i$ from 2 to $T_1 + 1$ do
        for $j$ from 2 to $T_2 + 1$ do
            if $|i - j| \leq 3 \vee i = T_1 + 1$ then
                $D[i, j] := d(W_1[i], W_2[j])$
                $D[i, j] := D[i, j] + \min(D[i - 1, j], D[i, j - 1], D[i - 1, j - 1])$
            end if
        end for
    end for
    Remove the first column and first row from $D$
    Calculate warping path via backtracking and store as $\pi$
    return $\pi, D[T_1, T_2]$
end function
```

That is, on average, how many time steps in W_1 is any given time step mapped to in W_2? Such distributions would give a more definitive answer as to whether the radius constraint often allows for stretching that is longer than desired. If there are many time steps in W_1 mapped to more than 3 time steps in W_2, then the warping radius is not satisfying the 30 ms constraint derived in Chapter 3, and an alternative solution would need to be pursued.

Nevertheless, the rudiments of an algorithmic method to acoustically compare words are present. The algorithm serves as an alternative method by which to compare words when the concept of “sound similarity” is invoked. Dynamic time warping in this sense is arguably a better method for evaluating the similarity of sounds than Levenshtein distance on phoneme strings, too, being that it compares sounds to each other.

5.3 Dynamic time warping and cognitive modeling

The point of using dynamic time warping in cognitive modeling is not to say that the dynamic programming aspect is directly related to cognition. Rather, it is that the absement value and the warping path specifically are related to cognition. The dynamic programming aspects are just the method by which to perform the appropriate calculations. The acoustic distinctiveness modeling results from Chapter 4 suggest that acoustic distinctiveness is not merely an artifact of duration, as the results of Chapter 2 could suggest. The correlation between duration and the acoustic distinctiveness values calculated with the neural network features was lower than when calculating acoustic distinctiveness using MFCCs. But, acoustic distinctiveness when calculated with the neural network features was still a significant predictor of response latency. Its significance suggests that acoustic distinctiveness, in general, is a significant predictor not merely because it correlates with duration, but rather, that it has a strong relationship to spoken word recognition.

The warping path should be of particular interest to models of spoken word recognition. The reason is that the warping path specifies the temporal dynamics of acoustic phonetic matching in the activation process. Individual acoustic frames within the signal and lexical template may be thought of as temporally elastic in the sense that the occurrence of the acoustic event the frame specifies is not rigidly assigned to a particular time point. It is the warping path that shows how elastic the frames were during the comparison process. There is some level of implausibility in the warping path that remains to be resolved, though. The optimal path in dynamic time warping is often determined by backtracking from the temporal endpoint of the comparisons, though it is clear that humans have more continuous and eager recognition of words in the speech signal. It is also possible to remember the optimal path by storing it as it is determined, like Jurafsky and Martin (2009, Chapter 9) do with the Viterbi algorithm. Doing so does not really resolve the issue of how the cognitive system would know when to progress to the next time step in either of the input sounds. Future research and development would be required to determine a method that solely uses a semi-elastic forward iteration to create a warping path that reasonably models the cognitive process of comparing the speech signal to a template. A key process to describe is when the comparison process knows to transition from one time step in the template or signal to the next. Perhaps a reasonable starting point would be to identify when the next time frame is more plausible than the previous one (i.e., transition eagerly, without attempting to find what is necessarily the optimal path). This sort of algorithm would be similar to the best path decoding algorithm given in Graves et al. (2006) and the beam search given in Graves and Jaitly (2014).

A more serious outlying problem is how to move from comparing isolated words to modeling the comparisons occurring during continuous speech. Inherently, it would be necessary to determine some sort of mechanism to determine when a word is recognized and when to move on to recognizing the next word. This problem is addressed in automatic speech recognition systems using algorithms like the Viterbi algorithm given in Jurafsky and Martin (2009, Chapter 9) and the token passing and beam search algorithms given in Graves (2012, Chapter 7), which constantly check the probability of beginning a new word at each assessed time step. These mechanisms are used when comparing acoustic representations of speech-to-grapheme and -phoneme models of words. One method used with dynamic time warping is searching for words starting at every frame to determine a temporal matching profile for each word at each point in the signal (Sakoe, 1979). Though, Rabiner and Schmidt (1980) pointed out that this is computationally infeasible, even when just recognizing digits, let alone all the possible words in a language. Rabiner and Schmidt instead opted to only search near the endpoints of already recognized words to shrink the amount of computation necessary. Nevertheless, work is still necessary to determine a recognition or transition mechanism that relates to cognitive processing.

Psycholinguistic models of spoken word recognition have primarily focused on isolated word recognition, and Nenadić (2020) found that few models had a mechanism that determines when a word is or should be recognized. DIANA (ten Bosch et al., 2013, 2015) and Fine-Tracker (Scharenborg, 2010) have so far been limited to isolated word recognition via modeling auditory lexical decision. EARSHOT (Magnuson et al., 2020) uses thresholds based on cosine similarity to determine when a semantic vector has received sufficient activation for a sustained amount of time that it could be recognized. And, SpeM (Scharenborg et al., 2005) and Shortlist B (Norris & McQueen, 2008) have an approach designed to be similar to the methods in automatic speech recognition. Each of these models used search algorithms that were very similar to those used in automatic speech recognition. Of these models, DIANA, SpeM, and Shortlist B all use phone or phoneme strings as the lexical model, though Shortlist B does not use acoustic data as input. DIANA in particular is not, however, restricted to using phonemic representations of words and was designed to allow for a variety of different input and representation types. Fine-Tracker, on the other hand, uses articulatory feature vectors as recognition targets, and related work has found articulatory information to be useful in computational speech processing (Espy-Wilson et al., 2019). EARSHOT uses random sparse semantic vectors. None of the representations are acoustic, though, so there is still a need to determine what the recognition mechanism would be for acoustic-to-acoustic comparisons, especially for cognitive modeling. It is possible that dynamic time warping methods from Rakthanmanon et al. (2012) could be adapted for this purpose, which have seen significant speedup for uses of dynamic time warping. Though, it would need to be possible to relate the methods or their results to cognitive processing.

Acoustic absement between words also has the opportunity to interface well with graph-theoretic models of the lexicon such as used in Vitevitch (2008) and Vitevitch (2021). These models of the lexicon represent words in the lexicon as nodes in a mathematical graph, and the nodes can be connected on the basis of a variety of factors. This graph-type of structure is overall more complex than a simple list of words as is commonly used in automatic speech recognition. It is common to connect them based on edit distance from phoneme strings. However, acoustic absement presents a different method by which to connect nodes based on their acoustic similarity. In turn, acoustic cohorts and neighborhoods of words and their associated processing effects could be more rigorously analyzed than in the previous chapters of this dissertation.

It may seem that these steps are outlined with the eventual goal of creating a new model of spoken word recognition. A new model would certainly be a natural extension of using dynamic time warping for comparing words acoustically. However, these steps are also necessary for simply extending the comparison mechanism to continuous speech. That is, a method for calculating acoustic absement for speech is incomplete if it can only handle comparisons between isolated words. For this reason, these steps are necessary to complete, whether or not a new model of spoken word recognition is pursued. One potential avenue to pursue would be to modify DIANA to use acoustic absement between the input signal and acoustic representations of words.

5.4 Invariance and speech communication

A central problem in phonetics is the lack of invariance seen in the speech signal. Appelbaum (1996) summarized the lack of invariance as a speech segment having multiple possible acoustic realizations and an acoustic parameter having multiple possible segment categories it could belong to. In mathematical terms, there is a many-to-many relationship between speech segments and acoustic parameters, which Magnuson et al. (2020) also recognized. Despite this type of relationship, listeners exhibit a consistent ability to understand the speech segments contained in a speech signal in normal listening conditions.

Appelbaum (1996) cataloged some previous attempts to resolve this problem, including the gesturalist and auditorist accounts of percepts as well as auditorist accounts of purportedly invariant acoustic cues. Ohala (1996) also cataloged some possible resolutions, such as using different units like diphones as percepts or that a different set of acoustic parameters was needed. In effect, these attempts at resolving the problem amount to trying to change the relationship between speech segment percepts and acoustic parameters to a one-to-one or many-to-one relationship (in terms of acoustic parameters mapping onto percepts) by either changing the set of parameters or changing the set of percepts. In line with Ohala's last proposed resolution (itself owing to Lindblom, 1996), I would like to conjecture that the many-to-many relationship between acoustic parameters and sets of speech segment percepts is simply a fact that we must accept and that invariance is to be found in the processing mechanisms that are employed during spoken word recognition. In other words, the relationship between speech segments and acoustic parameters is inherently variable, and invariant structures in comprehension are found through the auditory processing mechanisms that bound the possible variation in the speech signal. A variable relationship between speech segments and acoustic parameters was also acknowledged in Goldinger and Azuma (2003) when extending Adaptive Resonance Theory (Grossberg, 1982) to speech comprehension.

This speculative proposal was not directly tested in the previous chapters, but it seems a natural extension of the results and ideas. It may well be wrong, especially if an acoustically separable set of percepts can be developed. But, it is worth exploring, and in fact, some previous accounts of the relationship between speech segments or features and the speech signal have already directly and indirectly suggested invariant processing mechanisms. For example, K. N. Stevens and Blumstein (1978) suggested that there was an auditory processing mechanism whose purpose was to detect the spectral burst characteristics associated with the place of articulation of stops. Some accounts of phonology like substance-free phonology (Reiss, 2017) refer to transduction processes between distinctive features and speech, which necessarily requires a stable processing mechanism. As well, the invariance of processing mechanisms is trivially true in some sense. If this were not the case, it seems highly unlikely that speech communication would work at all since a speaker would not have any sense of guarantee that a word or utterance would be reliably processed.

Lindblom (1996) described the speech signal as inherently variable and posited that invariance in speech communication is found in the correct recognition of spoken words. He went on to describe speech as a goal-oriented process where the speaker adapts to the listener and that the listener's speech comprehension processes were designed to handle incomplete input. The perspective I am putting forth is a natural extension of these ideas, where lexical contrast is possible because of the processing mechanisms that the listener has. And, speakers are tacitly aware of these processing mechanisms and their tolerance for variability. Under this lens, speech communication is more actively collaborative because the speaker and listener are constantly working to accommodate each other. Reduction and enhancement are also better situated as related to the speaker's expectations of the listener's tolerance for signal variance, similar to the adaptive processes Lindblom (1990) discussed for H&H theory. Future research should focus on investigating measures of listener tolerance for variability, along the lines of how Redmon (2020) investigated acoustic contrast of obstruents with known acoustic correlates. Work involving acoustic absement and acoustic lexical templates must focus less on specific acoustic measurements and more so on how contrast occurs acoustically.

What's more, it may be worthwhile to invert the relationship between acoustics and articulation, seeking to first describe acoustic events and then find articulatory correlates of the acoustic properties. Articulatory correlates have previously been discussed in relation to phonological phenomena like ambisyllabicity (Gick, 2003) and stress (Engstrand, 1988). However, in a view of speech production where a speaker seeks to meet particular acoustic goals informed by perceptual goals, it seems natural to look first at the acoustic goals and then determine what articulations meet those goals, thereby finding motor equivalences (consult Perrier & Fuchs, 2015) for specific acoustic goals. This is related to the claim that part of the process for infants learning to produce speech is imitating sounds and adapting their articulations until they produce a sound that matches the target they are aiming for (Kuhl et al., 2008). Motor equivalences for speech sounds—such as bunched, curled, and hybrid [ɹ] in American English (Delattre & Freeman, 1968; Westbury et al., 1998)—would then be learned options that achieve the same acoustic goal.

This approach is not meant to replace the concept of contrastive units like phonemes, which have clearly been useful tools and models. There is indeed empirical evidence in the form of factorizability (Nearey, 2001) and CVC recognition accuracies (J. Allen, 1994) for abstract units used in speech perception. The focus on competition dynamics and acoustic models of words would form the basis upon which lexical contrast as exemplified by phoneme-like units could occur. The initial hypothesis is that lexical contrast occurs when the acoustic input provides sufficient discriminability to differentiate between similar sounds. This sort of discriminability is likely quantifiable with statistical distributions of the sort that Pigoli et al. (2018) used. These distributions and the notion of discriminability must be constrained in such a way that they inter-operate with the solutions to the previous section's word-segmentation problem.

This alternate view of the lack of invariance and of speech communication is still nascent and vague, almost by obligation. Many more analyses, reanalyses, and experiments would need to be performed in the future to bear it out.

5.5 Conclusion

The concepts of acoustic distance and acoustic absement touch on many concepts in phonetics, from the subphonemic level all the way up to the word and phrase levels. Having a perceptually-validated method of performing appropriate comparisons at these levels is paramount if these concepts are to be seriously integrated into phonetic theory. The results from this dissertation provide a candidate for calculating acoustic distance and acoustic absement from recordings instead of performing textual comparisons on phoneme strings. There are a number of possible research areas that might grow out of the use and refinement of this algorithm, such as mathematical modeling of the dynamics of non-symbolic competition and activation and perceptually-informed algorithm development. The calculation of acoustic absement also uses speech production data to quantify aspects of speech perception. As such, new explorations into the relationship between speech production and speech perception may be possible as well. Additionally, there is yet another inroad between speech science and automatic speech recognition which may foment productive collaboration. Perhaps most importantly, though, researchers in phonetics can now measure sound similarity between words as its name suggests: with sound.

References

Abanda, A., Mori, U., & Lozano, J. A. (2019). A review on distance based time series classification. *Data Mining and Knowledge Discovery*, *33*(2), 378–412. https://doi.org/10.1007/s10618-018-0596-4

Adank, P., & McQueen, J. M. (2007). The effect of an unfamiliar regional accent on spoken-word comprehension. *16th International Congress of Phonetic Sciences (ICPhS 2007)*, 1925–1928.

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255–265. https://doi.org/10.1093/biomet/60.2.255

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Albright, A., & Hayes, B. (2006). Modelling Productivity with the Gradual Learning Algorithm: The Problem of Accidentally Exceptionless Generalizations. In G. Fanselow, C. Féry, M. Schlesewsky, & R. Vogel (Eds.), *Gradience in Grammar: Generative Perspectives* (pp. 185–204). Oxford University Press. Retrieved December 20, 2018, from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199274796.001.0001/acprof-9780199274796-chapter-10

Allen, B., & Becker, M. (2015). *Learning alternations from surface forms with sublexical phonology.*

Allen, J. (1994). How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, *2*(4), 567–577. https://doi.org/10.1109/89.326615

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., … Zhu, Z. (2016). Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin. *International Conference on Machine Learning*, 173–182. Retrieved March 10, 2021, from http://proceedings.mlr.press/v48/amodei16.html

Appelbaum, I. (1996). The lack of invariance problem and the goal of speech perception. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, *3*, 1541–1544 vol.3. https://doi.org/10.1109/ICSLP.1996.607912

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The Discriminative Lexicon: A Unified Computational Model for the Lexicon and Lexical Processing in Comprehension and Production Grounded Not in (De)Composition but in Linear Discriminative Learning. *Complexity*. https://doi.org/10.1155/2019/4895891

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–

481. https://doi.org/10.1037/a0023851

Baayen, R. H., Shaoul, C., Willits, J., & Ramscar, M. (2016). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience*, *31*(1), 106–128. https://doi.org/10.1080/23273798.2015.1065336

Bagge Carlson, F. (2020). DynamicAxisWarping.jl. https://github.com/baggepinnen/DynamicAxisWarping.jl

Bagge Carlson, F., & Chitre, M. (2020). New Metrics Between Rational Spectra and their Connection to Optimal Transport. *arXiv:2004.09152 [cs, eess, math, stat]*. Retrieved May 16, 2021, from http://arxiv.org/abs/2004.09152

Bagwell, C. (2015). SoX - Sound eXchange. http://sox.sourceforge.net/

Banerjee, R., Bhattacharya, J., & Majumdar, P. (2021). Exponential-growth prediction bias and compliance with safety measures related to COVID-19. *Social Science & Medicine*, *268*, 113473. https://doi.org/10.1016/j.socscimed.2020.113473

Barreda, S. (2015). phonTools: Functions for phonetics in R. https://cran.r-project.org/web/packages/phonTools/index.html

Barreda, S. (2021). Fast Track: Fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard*, *7*(1). https://doi.org/10.1515/lingvan-2020-0051

Bartelds, M., Richter, C., Liberman, M., & Wieling, M. (2020). A new acoustic-based pronunciation distance measure. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.00039

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beguš, G., & Zhou, A. (2021). Interpreting intermediate convolutional layers in unsupervised acoustic word classification. *arXiv:2110.02375 [cs, eess]*. Retrieved October 7, 2021, from http://arxiv.org/abs/2110.02375

Benedetto, J., Heil, C., & Walnut, D. (1992). Uncertainty principles for time-frequency operators. In I. Gohberg (Ed.), *Continuous and discrete Fourier transforms, extension problems and Wiener-Hopf equations* (pp. 1–25). Birkhäuser. https://doi.org/10.1007/978-3-0348-8596-6_1

Bennett, R., Tang, K., & Sian, J. A. (2018). Statistical and acoustic effects on the perception of stop consonants in kaqchikel (mayan). *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *9*(1).

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Boersma, P., & Weenink, D. (2020). Praat: Doing phonetics by computer. https://www.praat.org

Boring, E. G. (1939). The Psychophysics of Color Tolerance. *The American Journal of Psychology*, *52*(3), 384–394. https://doi.org/10.2307/1416749

Browman, C. P., & Goldstein, L. (1992). Articulatory Phonology: An Overview. *Phonetica*, *49*(3-4), 155–180. https://doi.org/10.1159/000261913

Browman, C. P., & Goldstein, L. (1995). Dynamics and articulatory phonology. In R. F. Port & T. van Gelder (Eds.), *Mind as Motion: Explorations in the Dynamics of Cognition* (pp. 175–193). MIT Press.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658. https://doi.org/10.1121/1.1815131

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function.

Mathematics of Control, Signals and Systems, *2*(4), 303–314. https://doi.org/10.1007/BF02551274

Davies, M. (2008). The Corpus of Contemporary American English (COCA). retrieved June 25, 2021, from https://www.english-corpora.org/faq.asp

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Delattre, P., & Freeman, D. C. (1968). A dialect study of American R's by x-ray motion picture. *Linguistics*, *6*(44), 29–68. https://doi.org/10.1515/ling.1968.6.44.29

Deng, L., Cui, X., Pruvenok, R., Huang, J., Momen, S., Chen, Y., & Alwan, A. (2006). A Database of Vocal Tract Resonance Trajectories for Research in Speech Processing. *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, *1*, I–I. https://doi.org/10.1109/ICASSP.2006.1660034

Deng, L., Lee, L., Attias, H., & Acero, A. (2004). A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, I–557. https://doi.org/10.1109/ICASSP.2004.1326046

Deng, L., & O'Shaughnessy, D. (2003). *Speech processing: A dynamic and optimization-oriented approach*. CRC Press.

Dharmaretnam, D., Foster, C., & Fyshe, A. (2021). Words as a window: Using word embeddings to explore the learned representations of Convolutional Neural Networks. *Neural Networks*, *137*, 63–74. https://doi.org/10.1016/j.neunet.2020.12.009

Diehl, R. L., & Kluender, K. R. (1989). On the Objects of Speech Perception. *Ecological Psychology*, *1*(2), 121–144. https://doi.org/10.1207/s15326969eco0102_2 _eprint: https://doi.org/10.1207/s15326969eco0102_2

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., & Keogh, E. (2008). Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, *1*(2), 1542–1552. https://doi.org/10.14778/1454159.1454226

Dissen, Y., Goldberger, J., & Keshet, J. (2019). Formant estimation and tracking: A deep learning approach. *The Journal of the Acoustical Society of America*, *145*(2), 642–653. https://doi.org/10.1121/1.5088048

Dissen, Y., & Keshet, J. (2016). Formant Estimation and Tracking Using Deep Learning. *Interspeech 2016*, 958–962. https://doi.org/10.21437/Interspeech.2016-490

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. (2019). Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions. *arXiv:1901.03729 [cs]*. Retrieved May 21, 2021, from http://arxiv.org/abs/1901.03729

Engstrand, O. (1988). Articulatory correlates of stress and speaking rate in Swedish VCV utterances. *The Journal of the Acoustical Society of America*, *83*(5), 1863–1875. https://doi.org/10.1121/1.396522

Espy-Wilson, C., Lammert, A. C., Seneviratne, N., & Quatieri, T. F. (2019). Assessing neuromotor coordination in depression using inverted vocal tract variables. *Interspeech 2019*, 1448–1452. https://doi.org/10.21437/Interspeech.2019-1815

Farouk, M. (2018). *Application of Wavelets in Speech Processing*. Springer.

Farrar, C. D. (1981). *A Prototype Theory of Speech Perception* (Doctoral dissertation). The Ohio State University. United States – Ohio. Retrieved September 2, 2021, from http://www.proquest.com/docview/303183271/abstract/FEAA2CEE10AE4E39PQ/1

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*(2-3), 285–307. https://doi.org/10.1080/01638539809545029

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd). Sage. http://tinyurl.com/carbook

Fridland, V., & Kendall, T. (2017). Speech in the Silver State. *The Publication of the American Dialect Society*, *102*(1), 139–164. https://doi.org/10.1215/00031283-4295222

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity Avoidance and the OCP. *Natural Language & Linguistic Theory*, *22*(1), 179–228. https://doi.org/10.1023/B:NALA.0000005557.78535.3c

Gabor, D. (1946). Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, *93*(26), 429–441.

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*(3), 474–496.

Gahl, S., & Strand, J. F. (2016). Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language*, *89*, 162–178. https://doi.org/10.1016/j.jml.2015.12.006

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren,

N. L., & Zue, V. (1993). *DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1* (Technical Report No. 93). NASA/STI Recon.

Gick, B. (2003). Articulatory correlates of ambisyllabicity in English glides and liquids. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic interpretation: Papers in laboratory phonology VI* (pp. 222–236).

Goldin, G. A. (2014). Mathematical Representations. In S. Lerman (Ed.), *Encyclopedia of Mathematics Education* (pp. 409–413). Springer Netherlands. https://doi.org/10.1007/978-94-007-4978-8_103

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. https://doi.org/10.1037/0033-295X.105.2.251

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, *31*(3), 305–320. https://doi.org/10.1016/S0095-4470(03)00030-5

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language; New York*, *28*(5), 501–518. Retrieved March 8, 2017, from http://search.proquest.com.ezproxy1.library.arizona.edu/docview/1297352569/citation/6C8C08592D514B25PQ/1

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gorroochurn, P. (2016). On Galton's Change From "Reversion" to "Regression". *The American Statistician*, *70*(3), 227–231. https://doi.org/10.1080/00031305.2015.1087876

Graves, A. (2012). *Supervised sequence labelling with recurrent neural networks*. Springer.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning*, 369–376. https://doi.org/10.1145/1143844.1143891

Graves, A., & Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st international conference on machine learning* (pp. 1764–1772). PMLR. https://proceedings.mlr.press/v32/graves14.html

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18*(5), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

Grossberg, S. (1982). How does a brain build a cognitive code? In S. Grossberg (Ed.), *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control* (pp. 1–52). Springer Netherlands. https://doi.org/10.1007/978-94-009-7758-7_1

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Elements of Statistical Learning: Data mining, inference, and prediction* (Second). Springer. Retrieved December 16, 2019, from https://web.stanford.edu/~hastie/ElemStatLearn/

Heeringa, W. J. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance* (Doctoral dissertation). University of Groningen. Retrieved October 18, 2021, from https://research.rug.nl/en/publications/measuring-dialect-pronunciation-differences-using-levenshtein-dis

Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift fur Physik*, *43*, 172–198. https://doi.org/10.1007/BF01397280

Hendriks, R., Heusdens, R., & Jensen, J. (2004). Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *4*, iv–iv. https://doi.org/10.1109/ICASSP.2004.1326795

Henry, F. M. (1948). Discrimination of the duration of a sound. *Journal of Experimental Psychology*, *38*(6), 734–743. Retrieved November 24, 2021, from https://oce-ovid-com.login.ezproxy.library.ualberta.ca/article/00004782-194812000-00010/HTML

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111. https://doi.org/10.1121/1.411872

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., & Sainath, T. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, *29*. Retrieved October 27, 2019, from https://www.microsoft.com/en-us/research/publication/deep-neural-networks-for-acoustic-modeling-in-speech-recognition/

Hirsh, I. J. (1959). Auditory Perception of Temporal Order. *The Journal of the Acoustical Society of America*, *31*(6), 759–767. https://doi.org/10.1121/1.1907782

Howell, D. C. (2008). *Fundamental statistics for the behavioral sciences* (6th ed.). Thomson Higher Education.

Hsieh, I.-H., & Saberi, K. (2016). Imperfect pitch: Gabor's uncertainty principle and the pitch of extremely brief sounds. *Psychonomic Bulletin & Review*, *23*(1), 163–171. https://doi.org/10.3758/s13423-015-0863-y

Inman, H. F. (1984). *Behavior and Properties of the Overlapping Coefficient as a Measure of Agreement Between Distributions (association, Dissimilarity)*

(Ph.D. Dissertation). The University of Alabama at Birmingham. United States – Alabama. Retrieved May 21, 2019, from http://search.proquest.com/docview/303295844/abstract/6A03AD442D6C405BPQ/1

Innes, M., Saba, E., Fischer, K., Gandhi, D., Rudilosso, M. C., Joy, N. M., Karmali, T., Pal, A., & Shah, V. (2018). Fashionable modelling with flux. *CoRR*, *abs/1811.01457*. https://arxiv.org/abs/1811.01457

Innes, M. (2018). Flux: Elegant machine learning with julia. *Journal of Open Source Software*. https://doi.org/10.21105/joss.00602

Ito, M., Tsuchida, J., & Yano, M. (2001). On the effectiveness of whole spectral shape for vowel perception. *The Journal of the Acoustical Society of America*, *110*(2), 1141–1149. https://doi.org/10.1121/1.1384908

Iverson, P., Bernstein, L. E., & Auer Jr, E. T. (1998). Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition. *Speech Communication*, *26*(1), 45–63. https://doi.org/10.1016/S0167-6393(98)00049-1

Jakobson, R., Fant, G., & Halle, M. (1952). *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates.*

Johnson, K. (1997). *The auditory/perceptual basis for speech segmentation* (Working Paper). Ohio State University Department of Linguistics. Retrieved March 31, 2021, from https://kb.osu.edu/handle/1811/81782

Johnson, K. (2012). *Acoustic and auditory phonetics* (Third). Wiley-Blackwell.

Johnston, J. D., Dunn, C. J., & Vernon, M. J. (2019). Tree traits influence response to fire severity in the western Oregon Cascades, USA. *Forest Ecology and Management*, *433*, 690–698. https://doi.org/10.1016/j.foreco.2018.11.047

Jones, S., & Brandt, S. (2019). Density and distinctiveness in early word learning:Evidence from neural network simulations. *Cognitive Science*. Retrieved

December 16, 2019, from https://eprints.lancs.ac.uk/id/eprint/139403/

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252–1263. https://doi.org/10.1121/1.1288413

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing* (Second). Pearson Prentice Hall.

Kapatsinski, V. (2005). *Sound similarity relations in the mental lexicon: Modeling the lexicon as a complex network* (Progress Report No. 27). Indiana University Speech Research Laboratory. Bloomington, IN.

Kawahara, H., & Matsui, H. (2003). Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, *1*, I–I. https://doi.org/10.1109/ICASSP.2003.1198766

Keller, M. T., & Trotter, W. T. (2017). *Applied combinatorics.* https://www.appliedcombinatorics.org

Kelley, M. C. (2020). Phonetics.jl. https://github.com/maetshju/Phonetics.jl

Kelley, M. C., & Tucker, B. V. (2020). A comparison of four vowel overlap measures. *The Journal of the Acoustical Society of America*, *147*(1), 137–145. https://doi.org/10.1121/10.0000494

Kelley, M. C., & Tucker, B. V. (2021a). Acoustic absement files. Retrieved October 6, 2021, from https://doi.org/10.7939/r3-mekk-5635

Kelley, M. C., & Tucker, B. V. (2021b). *Perception and timing of acoustic distance* [Manuscript in preparation].

Kelley, M. C., & Tucker, B. V. (2021c). *Using acoustic distance and acoustic absement to quantify lexical competition* [Manuscript in preparation].

Kendall, T., & Fridland, V. (2012). Variation in perception and production of mid front vowels in the U.S. Southern Vowel Shift. *Journal of Phonetics*, *40*(2), 289–306. https://doi.org/10.1016/j.wocn.2011.12.002

Kent, R. D., & Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of Communication Disorders*, *74*, 74–97. https://doi.org/10.1016/j.jcomdis.2018.05.004

Kewley-Port, D., & Watson, C. S. (1994). Formant-frequency discrimination for isolated English vowels. *The Journal of the Acoustical Society of America*, *95*(1), 485–496. https://doi.org/10.1121/1.410024

Kewley-Port, D., Watson, C. S., & Foyle, D. C. (1988). Auditory temporal acuity in relation to category boundaries; speech and nonspeech stimuli. *The Journal of the Acoustical Society of America*, *83*(3), 1133–1145. https://doi.org/10.1121/1.396058

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, san diego, CA, USA, may 7-9, 2015, conference track proceedings*. http://arxiv.org/abs/1412.6980

Kingston, J., & Diehl, R. L. (1995). Intermediate properties in the perception of distinctive feature values. In B. Connell & A. Arvanti (Eds.), *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV* (pp. 7–27). Cambridge University Press. https://doi.org/10.1017/CBO9780511554315.002

Kirchner, R., Moore, R. K., & Chen, T.-Y. (2010). Computing phonological generalization over real speech exemplars. *Journal of Phonetics*, *38*(4), 540–547. https://doi.org/10.1016/j.wocn.2010.07.005

Klatt, D. (1981). Lexical representations for speech production and perception. In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of* *speech* (pp. 11–31). North-Holland Publishing Company.

Kohler, K. J. (1995). Phonetics - A language science in its own right? *Proceedings of the XIIIth International Congress of Phonetic Sciences*, *1*, 10–17.

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, 288–295. Retrieved December 20, 2018, from http://dl.acm.org/citation.cfm?id=974305.974343

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25*, *25*, 1097–1105. Retrieved May 21, 2021, from https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Krug, A., Knaebel, R., & Stober, S. (2018). Neuron activation profiles for interpreting convolutional speech recognition models. *NeurIPS 2018 Interpretability and Robustness for Audio, Speech and Language Workshop (IRASL'18)*. Retrieved May 9, 2019, from https://openreview.net/forum?id=Bylpgfjen7

Kuhl, P. K. (1992). Psychoacoustics and speech perception: Internal standards, perceptual anchors, and prototypes. In L. A. Werner & E. W. Rubel (Eds.), *Developmental psychoacoustics* (pp. 293–332). Kuhl, Patricia K.: U Washington, Dept of Speech & Hearing Sciences, Seattle, WA, US. https://doi.org/10.1037/10119-012

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 979–1000. https://doi.org/10.1098/rstb.2007.2154

Ladefoged, P., & Broadbent, D. E. (1957). Information Conveyed by Vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104. https://doi.org/10.1121/1.1908694

Lapid, E., Ulrich, R., & Rammsayer, T. (2008). On estimating the difference limen in duration discrimination tasks: A comparison of the 2AFC and the reminder task. *Perception & Psychophysics*, *70*(2), 291–305. https://doi.org/10.3758/PP.70.2.291

Lavalle, S. M. (2006). *Planning algorithms*. Cambridge University Press.

Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477. https://doi.org/10.1121/1.1912375

Levy, M. R., & Tasoff, J. (2017). Exponential-growth bias and overconfidence. *Journal of Economic Psychology*, *58*, 1–14. https://doi.org/10.1016/j.joep.2016.11.001

Lewandowski, N. (2012). *Talent in nonnative phonetic convergence* (Doctoral Dissertation). Retrieved October 8, 2017, from http://elib.uni-stuttgart.de/handle/11682/2875

Lewandowski, N., & Jilka, M. (2019). Phonetic Convergence, Language Talent, Personality and Attention. *Frontiers in Communication*, *4*. https://doi.org/10.3389/fcomm.2019.00018

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403–439). Springer Netherlands. https://doi.org/10.1007/978-94-009-2037-8_16

Lindblom, B. (1978). Phonetic aspects of linguistic explanation. *Studia Linguistica*, *32*(1-2), 137–153. https://doi.org/10.1111/j.1467-9582.1978.tb00335.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9582.1978.tb00335.x

Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *The Journal of the Acoustical Society of America*, *99*(3), 1683–1692. https://doi.org/10.1121/1.414691

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word-journal of The International Linguistic Association*, *20*(3), 384–422.

Llorach, E. A. (1950). *Fonología española* (1st ed.). Gredos.

Lohmann, A. (2018). Time and thyme are not homophones: A closer look at gahl's work on the lemma-frequency effect, including a reanalysis. *Language*, *94*(2), e180–e190.

Luce, P. A. (1986). *Neighborhoods of Words in the Mental Lexicon* (Technical Report No. 6). Retrieved October 10, 2018, from https://eric.ed.gov/?id=ED353610

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics*, *62*(3), 615–625. https://doi.org/10.3758/BF03212113

Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and hearing*, *19*(1), 1–36. Retrieved April 29, 2018, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3467695/

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., & Rueckl, J. G. (2020). EARSHOT: A Minimal Neural Network Model of Incremental Human Speech Recognition. *Cognitive Science*, *44*(4), e12823. https://doi.org/10.1111/cogs.12823

_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12823

Mallat, S. (2009). *A Wavelet Tour of Signal Processing* (Third). Elsevier.

Mann, S., Hao, M. L., Tsai, M., Hafezi, M., Azad, A., & Keramatimoezabad, F. (2018). Effectiveness of Integral Kinesiology Feedback for Fitness-Based Games. *2018 IEEE Games, Entertainment, Media Conference (GEM)*, 1–9. https://doi.org/10.1109/GEM.2018.8516533

Mann, S., Janzen, R., & Post, M. (2006). Hydraulophone design considerations: Absement, displacement, and velocity-sensitive music keyboard in which each key is a water jet. *Proceedings of the 14th ACM International Conference on Multimedia*, 519–528. https://doi.org/10.1145/1180639.1180751

Marslen-Wilson, W., Moss, H. E., & van Halen, S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(6), 1376–1392. https://doi.org/10.1037/0096-1523.22.6.1376

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63. https://doi.org/10.1016/0010-0285(78)90018-X

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McCloy, D. R. (2013). *Prosody, intelligibility and familiarity in speech perception* (Thesis). Retrieved February 5, 2019, from https://digital.lib.washington.edu:443/researchworks/handle/1773/23472

Mermelstein, P. (1976). Distance measures for speech recognition—psychological and instrumental (C. H. Chen, Ed.). *Pattern recognition and artificial intelligence*, *116*, 374–388.

Mielke, J. (2012). A phonetically based metric of sound similarity. *Lingua*, *122*(2), 145–163. https://doi.org/10.1016/j.lingua.2011.04.006

Mohr, B., & Wang, W. S.-Y. (1968). Perceptual Distance and the Specification of Phonological Features. *Phonetica*, *18*, 31–45.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165–178.

Mukai, Y., Järvikivi, J., & Tucker, B. V. (2018). The effect of phonological-orthographic consistency on the processing of reduced and citation forms of Japanese words: Evidence from pupillometry. *Proceedings of the 2018 Annual Conference of the Canadian Linguistics Association*.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, *85*(5), 2088–2113. https://doi.org/10.1121/1.397861

Nearey, T. M. (2001). Phoneme-like units and speech perception. *Language and Cognitive Processes*, *16*(5-6), 673–681. https://doi.org/10.1080/01690960143000173

Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *The Journal of the Acoustical Society of America*, *80*(5), 1297–1308. https://doi.org/10.1121/1.394433

Nelson, N. R., & Wedel, A. (2017). The phonetic specificity of competition: Contrastive hyperarticulation of voice onset time in conversational English. *Journal of Phonetics*, *64*(Supplement C), 51–70. https://doi.org/10.1016/j.wocn.2017.01.008

Nenadić, F. (2020). *Computational Modelling of Spoken Word Recognition in the Auditory Lexical Decision Task* (Doctoral Dissertation). University of Alberta. Edmonton, Alberta, Canada. https://doi.org/10.7939/r3-whrd-a130

Nenadić, F., Coulter, P., Nearey, T. M., & Kiefte, M. (2020). Perception of vowels with missing formant peaks. *The Journal of the Acoustical Society of America*, *148*(4), 1911–1921. https://doi.org/10.1121/10.0002110

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357

Nycz, J., & Hall-Lew, L. (2015). Best practices in measuring vowel merger. *Proceedings of Meetings on Acoustics*, *20*(1), 060008. https://doi.org/10.1121/1.4894063

Oates, T., Firoiu, L., & Cohen, P. R. (2000). Using Dynamic Time Warping to Bootstrap HMM-Based Clustering of Time Series. In R. Sun & C. L. Giles (Eds.), *Sequence Learning: Paradigms, Algorithms, and Applications* (pp. 35–52). Springer. https://doi.org/10.1007/3-540-44565-X_3

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *The Journal of the Acoustical Society of America*, *99*(3), 1718–1725. https://doi.org/10.1121/1.414696

Palaz, D., Magimai-Doss, M., & Collobert, R. (2015). Analysis of CNN-based Speech Recognition System using Raw Speech as Input. *Proceedings of Interspeech*, 11–15. Retrieved July 10, 2017, from https://infoscience.epfl.ch/record/210039/files/Palaz_Idiap-RR-23-2015.pdf

Parhizkar, R., Barbotin, Y., & Vetterli, M. (2015). Sequences with minimal time–frequency uncertainty. *Applied and Computational Harmonic Analysis*, *38*(3), 452–468. https://doi.org/10.1016/j.acha.2014.07.001

Parr, T., & Wilson, J. D. (2020). Technical Report: Partial Dependence through Stratification. *arXiv:1907.06698 [cs, stat]*. Retrieved May 27, 2021, from http://arxiv.org/abs/1907.06698

Pastore, M., & Calcagnì, A. (2019). Measuring Distribution Similarities Between

Samples: A Distribution-Free Overlapping Index. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01089

Pastore, R. E., Ahroon, W. A., Baffuto, K. J., Friedman, C., Puleo, J. S., & Fink, E. A. (1977). Common-factor model of categorical perception. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(4), 686–696. https://doi.org/10.1037/0096-1523.3.4.686

Perre, L., & Ziegler, J. C. (2008). On-line activation of orthography in spoken word recognition. *Brain research*, *1188*, 132–138.

Perrier, P., & Fuchs, S. (2015). Motor equivalence in speech production. In M. A. Redford (Ed.), *The handbook of speech production* (pp. 225–247). Wiley Blackwell. https://doi.org/10.1002/9781118584156.ch11

Peterson, G. E., & Harary, F. (1961). Foundations of Phonemic Theory. *Proceedings of Symposia in Applied Mathematics*, *12*, 139–165.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184. https://doi.org/10.1121/1.1906875

Petitjean, F., Forestier, G., Webb, G. I., Nicholson, A. E., Chen, Y., & Keogh, E. (2014). Dynamic Time Warping Averaging of Time Series Allows Faster and More Accurate Classification. *2014 IEEE International Conference on Data Mining*, 470–479. https://doi.org/10.1109/ICDM.2014.27

Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, *44*(3), 678–693. https://doi.org/10.1016/j.patcog.2010.09.013

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137–157). John Benjamins.

Pigoli, D., Hadjipantelis, P. Z., Coleman, J. S., & Aston, J. a. D. (2018). The statistical analysis of acoustic phonetic data: Exploring differences between spoken Romance languages. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *67*(5). Retrieved March 13, 2021, from https://ora.ox.ac.uk/objects/uuid:1ddd672f-1b77-4fc4-8ee2-7c9536b5ea0d

Pike, K. (1943). *Phonetics: A critical analysis of phonetic theory and a technic for the practical description of sounds*. The University of Michigan Press.

Porretta, V., & Tucker, B. V. (2013). Perception of non-native consonant length in naïve English listeners. *Proceedings of Meetings on Acoustics*, *19*(1), 060274. https://doi.org/10.1121/1.4800677

Port, R. F., & Leary, A. P. (2005). Against formal phonology. *Language*, *81*(4), 927–964. https://doi.org/10.1353/lan.2005.0195

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical Recipes in C: The Art of Scientific Computing, Second Edition* (2 edition). Cambridge University Press.

R Core Team. (2018). *R: A language and environment for statistical computing.* Manual. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

R Core Team. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rabiner, L., & Schmidt, C. (1980). Application of dynamic time warping to connected digit recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 377–388. https://doi.org/10.1109/TASSP.1980.1163422

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., & Keogh, E. (2012). Searching and mining trillions of time series subsequences under dynamic time warping. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 262–270. https://doi.org/10.1145/2339530.2339576

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). Springer. Retrieved May 18, 2021, from https://login.ezproxy.library.ualberta.ca/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat03710a&AN=alb.3296008&site=eds-live&scope=site

Ratanamahatana, C. A., & Keogh, E. (2004). Everything you know about dynamic time warping is wrong. *Third Workshop on Mining Temporal and Sequential Data*, *32*.

Redmon, C. H. (2020). *Lexical acoustics: Linking phonetic systems to the higher-order units they encode* (Ph.D.). University of Kansas. United States – Kansas.

Reiss, C. (2017). Substance free phonology. In S. J. Hannahs & A. Bosch (Eds.), *The Routledge handbook of phonological theory* (pp. 425–452). https://doi.org/10.4324/9781315675428-15

Sakoe, H. (1979). Two-level DP-matching–A dynamic programming-based pattern matching algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *27*(6), 588–595. https://doi.org/10.1109/TASSP.1979.1163310

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), 43–49. https://doi.org/10.1109/TASSP.1978.1163055

Sakoe, H., & Chiba, S. (1970). A similarity evaluation of speech patterns by dynamic programming. *Nat. Meeting of Institute of Electronic Communications Engineers of Japan*, 136.

Saltzman, E. L., & Munhall, K. G. (1989). A Dynamical Approach to Gestural Patterning in Speech Production. *Ecological Psychology*, *1*(4), 333–382. https://doi.org/10.1207/s15326969eco0104_2

Sanders, N. C., & Chin, S. B. (2009). Phonological Distance Measures. *Journal of quantitative linguistics*, *16*(1), 96–114. https://doi.org/10.1080/09296170802514138

Saporta, S. (1955). Frequency of Consonant Clusters. *Language*, *31*(1), 25–30. https://doi.org/10.2307/410889

Scharenborg, O. (2010). Modeling the use of durational information in human spoken-word recognition. *The Journal of the Acoustical Society of America*, *127*(6), 3758–3770. https://doi.org/10.1121/1.3377050

Scharenborg, O., Norris, D., ten Bosch, L., & McQueen, J. M. (2005). How Should a Speech Recognizer Work? *Cognitive Science*, *29*(6), 867–918. https://doi.org/10.1207/s15516709cog0000_37

Schonger, M., & Sele, D. (2020). How to better communicate the exponential growth of infectious diseases. *PLOS ONE*, *15*(12), e0242839. https://doi.org/10.1371/journal.pone.0242839

Seyfarth, S., Garellek, M., Gillingham, G., Ackerman, F., & Malouf, R. (2018). Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience*, *33*(1), 32–49.

Shankweiler, D., & Fowler, C. A. (2015). Seeking a reading machine for the blind and discovering the speech code. *History of Psychology*, *18*(1), 78–99. https://doi.org/10.1037/a0038299

Sheu, Y.-h. (2020). Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research. *Frontiers in Psychiatry*, *11*. https://doi.org/10.3389/fpsyt.2020.551299

Stango, V., & Zinman, J. (2009). Exponential Growth Bias and Household Finance. *The Journal of Finance*, *64*(6), 2807–2849. https://doi.org/10.1111/j.1540-6261.2009.01518.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2009.01518.x

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, *64*(5), 1358–1368. https://doi.org/10.1121/1.382102

Stevens, K. N. (1998). *Acoustic phonetics*. MIT Press.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185–190. https://doi.org/10.1121/1.1915893

Taft, M., Castles, A., Davis, C., Lazendic, G., & Nguyen-Hoan, M. (2008). Automatic activation of orthography in spoken word recognition: Pseudohomograph priming. *Journal of Memory and Language*, *58*(2), 366–379.

ten Bosch, L., Boves, L., & Ernestus, M. (2013). Towards an end-to-end computational model of speech comprehension: Simulating a lexical decision task. *INTERSPEECH-2013*, 2822–2826.

ten Bosch, L., Boves, L., & Ernestus, M. (2015). Diana, an end-to-end computational model of human word comprehension. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*, 1–4.

ten Bosch, L., Boves, L., Tucker, B. V., & Ernestus, M. (2015). DIANA: Towards computational modeling reaction times in lexical decision in North American English. *INTERSPEECH-2015*, 1576–1580.

ten Bosch, L., Ernestus, M., & Boves, L. (2018). Analyzing Reaction Time Sequences from Human Participants in Auditory Experiments. *Interspeech 2018*, 971–

975. https://doi.org/10.21437/Interspeech.2018-1728

Terasawa, H., Slaney, M., & Berger, J. (2005). Perceptual distance in timbre space. *Proceedings of ICAD 05*, 61–68. Retrieved October 27, 2019, from https://smartech.gatech.edu/handle/1853/50176

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. *2015 IEEE Information Theory Workshop (ITW)*, 1–5. https://doi.org/10.1109/ITW.2015.7133169

Tomaschek, F. (2013). *Behavioral and neural correlates of vowel length in German and of its interaction with the tense/lax contrast* (Dissertation). Universität Tübingen.

Tomaschek, F., Truckenbrodt, H., & Hertrich, I. (2011). Processing german vowel quantity: Categorical perception or perceptual magnet effect? *ICPhS XVII*, 2002–2005.

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*, *51*(3), 1187–1204. https://doi.org/10.3758/s13428-018-1056-1

Uchida, S., & Sakoe, H. (2005). A Survey of Elastic Matching Techniques for Handwritten Character Recognition. *IEICE TRANSACTIONS on Information and Systems*, *E88-D*(8), 1781–1790. Retrieved September 3, 2021, from https://search.ieice.org/bin/summary.php?id=e88-d_8_1781&category=D&year=2005&lang=E&abst=

van de Geijn, R., & Myers, M. (2020). *Advanced Linear Algebra: Foundations to Frontiers*.

van Gelder, T. (1995). What Might Cognition Be, If Not Computation? *The Journal of Philosophy*, *92*(7), 345–381. https://doi.org/10.2307/2941061 van Gelder, T. (1997). The Dynamical Hypothesis in Cognitive Science. *Behavioral and Brain Sciences*, *21*, 615–665.

van Leeuwen, D. (2019). MFCC.jl. https://github.com/JuliaDSP/MFCC.jl

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2017). itsadug: Interpreting time series and autocorrelated data using gamms. https://cran.r-project.org/web/packages/itsadug/index.html

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language & Hearing Research*, *51*(2), 408–422. Retrieved April 29, 2018, from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=rzh&AN=105776080&site=eds-live&scope=site

Vitevitch, M. S. (2021). What can network science tell us about phonology and language processing? *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12532

_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12532

Vitevitch, M. S., & Luce, P. A. (2016). Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics*, *2*(1), 75–94. https://doi.org/10.1146/annurev-linguistics-030514-124832

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, Neighborhood Activation, and Lexical Access for Spoken Words. *Brain and*

Language, *68*(1), 306–311. https://doi.org/10.1006/brln.1999.2116

Vitevitch, M. S., & Rodríguez, E. (2005). Neighborhood density effects in spoken word recognition in Spanish. *Journal of Multilingual Communication Disorders*, *3*(1), 64–73. https://doi.org/10.1080/14769670400027332

Vitevitch, M. S., & Storkel, H. L. (2013). Examining the Acquisition of Phonological Word Forms with Computational Experiments. *Language and Speech*, *56*(4), 493–527. https://doi.org/10.1177/0023830912460513

Vitz, P. C., & Winkler, B. S. (1973). Predicting the judged "similarity of sound" of english words. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 373–388. https://doi.org/10.1016/S0022-5371(73)80016-7

Warner, N., Fountain, A., & Tucker, B. V. (2009). Cues to perception of reduced flaps. *The Journal of the Acoustical Society of America*, *125*(5), 3317–3327. https://doi.org/10.1121/1.3097773

Warner, N., Jongman, A., Sereno, J., & Kemps, R. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from dutch. *Journal of phonetics*, *32*(2), 251–276.

Warner, N., & Tucker, B. V. (2017). An effect of flaps on the fourth formant in English. *Journal of the International Phonetic Association*, *47*(1), 1–15. https://doi.org/10.1017/S0025100316000219

Westbury, J. R., Hashi, M., & J. Lindstrom, M. (1998). Differences among speakers in lingual articulation for American English /ɹ/. *Speech Communication*, *26*(3), 203–226. https://doi.org/10.1016/S0167-6393(98)00058-2

William Marslen-Wilson, & Pienie Zwisterlood. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human perception and performance*, *15*(3), 576–585.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood, S. N. (2020). mgcv. https://cran.r-project.org/web/packages/mgcv/index.html

Wu, R., & Keogh, E. J. (2020). FastDTW is approximate and Generally Slower than the Algorithm it Approximates. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/TKDE.2020.3033752

Yoneyama, K. (2002). *Phonological neighborhoods and phonetic similarity in Japanese word recognition* (Doctoral dissertation). The Ohio State University.

You, H., & Magnuson, J. S. (2018). TISK 1.0: An easy-to-use Python implementation of the time-invariant string kernel model of spoken word recognition. *Behavior Research Methods*, *50*(3), 871–889. https://doi.org/10.3758/s13428-017-1012-5

Zeghidour, N., Usunier, N., Synnaeve, G., Collobert, R., & Dupoux, E. (2018). End-to-End Speech Recognition from the Raw Waveform. *Interspeech 2018*, 781–785. https://doi.org/10.21437/Interspeech.2018-2414

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 818–833). Springer International Publishing. https://doi.org/10.1007/978-3-319-10590-1_53

Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Laurent, C., Bengio, Y., & Courville, A. (2016). Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks. *Interspeech 2016*, 410–414. https://doi.org/10.21437/Interspeech.2016-1446

Ziegler, J. C., & Ferrand, L. (1998). Orthography shapes the perception of speech:

The consistency effect in auditory word recognition. *Psychonomic Bulletin & Review*, *5*(4), 683–689.

Zue, V., & Cole, R. (1979). Experiments on spectrogram reading. *ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, *4*, 116–119. https://doi.org/10.1109/ICASSP.1979.1170735

Zwicker, E. (1961). Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen). *The Journal of the Acoustical Society of America*, *33*(2), 248–248. https://doi.org/10.1121/1.1908630

Appendix A

Concurvity and the Frobenius norm

A.1 Understanding concurvity

To better understand the indices of concurvity that `mgcv` provides in the `concurvity` function, consider that one part of what occurs when a matrix A is multiplied with a vector b as in Ab is that b can change in length. Determining the size of a matrix, as it were, is often done by determining just how much A can stretch any particular vector b. Now, consider that a GAMM is given as a model matrix M containing the data that the model was fit to and a vector b of coefficients. When M is multiplied with b as in Mb, the GAMM's predicted values are obtained. What follows is merely providing some extra detail to the documentation and source code contained in the `mgcv` package (Wood, 2020). The `concurvity` function determines how much of the stretching that a smooth term does on b can be explained by other smooth and parametric terms in the model. This is performed by decomposing M using the QR decomposition to cast the effects of the smooth term in question in terms of the orthnormalized effects of all the other terms in the model as well as what is left to be explained by the smooth term.

Because the smooth is now decomposed in this way, the proportion of how much of the smooth term's stretching can be explained by the other terms to how much stretching the smooth term does in total on b can be calculated. When the proportion is high, it indicates that the effect of the particular smooth in question can be explained by other terms in the model. Because there are infinite particular instantiations of b, the `concurvity` function provides three indices of this value.

The first index, termed "worst", is the greatest value that the proportion can be for all possible particular instantiations of b. The second index, termed "observed", is the value of this proportion as calculated using the coefficients from the fitted model for b. And, the third index, termed "estimate", is found by comparing the size of the matrix containing the entire smooth effect to the size of the submatrix that contains the components of the smooth explained by other terms in the model, where the size of the matrix is measured with the Frobenius norm. Because the Frobenius norm is equal to a scaled version of the average or expected value of the stretching of a vector, this estimate index can also roughly be seen as what the proportion would be given an average-case instantiation of b. We do not believe that this is commonly known in our field, and it is difficult to look up, so we provide a proof of this statement regarding the Frobenius norm (specifically that for an m-by-n matrix A, $\|A\|_F^2 = n\, E_{\|x\|_2=1} \|Ax\|_2^2$) subsequently.

A.2 Frobenius norm as average case stretching

It is also well-known in linear algebra that the Frobenius norm $||A||_F$ is equal to the ℓ_2 norm of its singular values. That is,

$$\|A\|_F = \sqrt{\sum_i \sigma_i^2}\,. \tag{A.1}$$

We wish to prove the following theorem.

Theorem 1. *Given a matrix A with n columns, $\|A\|_F^2 = n \mathop{\mathbb{E}}_{\|x\|_2=1} \|Ax\|_2^2$.*

Proof. By the singular value decomposition theorem, we have $A = U\Sigma V^H$. By substituting for A, we have

$$\begin{aligned}
\|Ax\|_2^2 &= \|U\Sigma V^H x\|_2^2 \\
&= (U\Sigma V^H x)^H U\Sigma V^H x && <\|x\|_2^2 = x^H x> \\
&= x^H V\Sigma^H U^H U\Sigma V^H x && <(AB)^H = B^H A^H> \\
&= x^H V\Sigma^2 V^H x && <U \text{ is unitary, algebra}> \\
&= b^H \Sigma^2 b && <\text{Let } b = V^H x> \\
&= \sum_i \beta_i^2 \sigma_i^2 && <\text{algebra}>
\end{aligned}$$

We can now substitute this expression for $\|Ax\|_2^2$ in the initial expression to get

$$
\begin{aligned}
n \mathop{\mathbb{E}}_{\|x\|_2=1} \|Ax\|_2^2 &= n \mathop{\mathbb{E}}_{\|x\|_2=1} \left[\sum_i \beta_i^2 \sigma_i^2 \right] \\
&= n \sum_i \mathop{\mathbb{E}}_{\|x\|_2=1} \left[\beta_i^2 \right] \sigma_i^2 && \text{<}\mathbb{E} \text{ is linear>}
\end{aligned}
$$

We know that $\mathbb{E}_{\|x\|_2=1} \sum_i \beta_i^2 = 1$ because $\|x\|_2 = 1$ and V^H is unitary. We also know that b must be on the unit sphere. This implies that each $\beta_i \in b$ is identically distributed, and so their expectations must be equivalent, meaning that $\mathbb{E}_{\|x\|_2=1} \beta_i^2 = \frac{1}{n}$. We thus have

$$
n \sum_i \mathop{\mathbb{E}}_{\|x\|_2=1} \left[\beta_i^2 \right] \sigma_i^2 = n \frac{1}{n} \sum_i \sigma_i^2 = \sum_i \sigma_i^2 \,,
$$

which is clearly equivalent to the definition of $\|A\|_F$ in Equation A.1 after taking the square root. □

We offer a note to readers who attempt to numerically verify this proof using linear algebra software that such software often does not return the 0 values associated with rank-deficient matrices, and this must be accounted for when determining conformability of matrix and vector sizes and counting the total number of singular values if using rank-deficient matrices.

Appendix B

Formant values from distance rating task

Table B.1: Table of the formant values used to synthesize each vowel stimulus in the distance rating task. Each row is numbered according to the order in which the stimuli were generated. The F1 and F2 values for the first stimulus are presented in columns $F1_1$ and $F2_1$, respectively. The F1 and F2 values for the second stimulus are presented in columns $F1_2$ and $F2_2$, respectively.

Vowel	Stimulus Number	$F1_1$	$F2_1$	$F1_2$	$F2_2$
i	1	510.28	2340.71	600.87	2590.07
i	2	509.18	2562.81	520.97	2646.71
i	3	391.95	2458.43	479.32	2781.47
i	4	564.43	2696.75	509.58	2485.34
i	5	490.17	2436.15	561.30	2574.82
i	6	585.23	2404.26	524.36	2478.27
i	7	451.67	2476.35	590.72	2712.41

Vowel	Stimulus Number	$F1_1$	$F2_1$	$F1_2$	$F2_2$
i	8	534.83	2486.72	487.03	2411.40
i	9	670.27	2576.03	593.57	2607.43
i	10	444.03	2504.54	563.24	2365.88
i	11	676.50	2362.89	513.04	2421.21
i	12	475.01	2407.27	604.93	2490.56
i	13	611.09	2305.77	495.77	2496.56
i	14	492.45	2618.84	642.07	2616.94
i	15	480.10	2343.59	501.73	2584.00
i	16	585.12	2520.33	518.14	2397.44
i	17	614.03	2562.38	548.77	2621.79
i	18	483.78	2762.13	605.90	2599.76
i	19	567.04	2401.98	494.39	2649.29
i	20	589.30	2673.16	558.72	2273.70
i	21	461.81	2600.36	580.17	2602.96
i	22	541.05	2418.51	605.67	2612.32
i	23	509.35	2428.76	505.24	2607.50
i	24	567.77	2320.42	525.21	2723.55
i	25	549.81	2536.49	526.79	2510.41
i	26	587.81	2504.54	623.54	2387.07
i	27	542.55	2254.14	579.99	2371.83
i	28	585.35	2556.66	528.18	2374.05
i	29	557.97	2305.87	543.40	2524.75
i	30	501.65	2662.83	477.63	2465.93

Vowel	Stimulus Number	$F1_1$	$F2_1$	$F1_2$	$F2_2$
i	31	576.33	2592.88	596.64	2421.51
i	32	526.99	2414.41	516.10	2587.12
i	33	526.51	2547.82	566.69	2472.52
u	1	559.88	1177.76	621.08	1344.67
u	2	574.19	1053.53	535.22	1106.74
u	3	597.90	1126.23	408.96	1211.97
u	4	629.84	1088.74	667.30	981.55
u	5	627.63	1051.89	686.34	1240.28
u	6	550.14	1162.44	481.32	1033.96
u	7	665.65	1361.24	643.40	1325.98
u	8	624.70	1141.76	468.97	906.50
u	9	584.49	1095.42	617.51	1269.21
u	10	569.35	1179.94	567.13	1292.86
u	11	582.72	1319.87	627.19	1108.98
u	12	523.28	1090.00	573.55	1241.17
u	13	503.84	1211.21	593.59	1159.53
u	14	530.15	1064.51	558.78	1086.75
u	15	459.84	1235.96	525.08	1185.93
u	16	510.91	1054.03	650.47	1338.28
u	17	586.21	1231.99	563.10	1249.08
u	18	651.59	1394.69	556.92	1254.80
u	19	721.86	1329.05	546.14	1121.28
u	20	502.21	1165.89	517.79	1241.97

Vowel	Stimulus Number	$F1_1$	$F2_1$	$F1_2$	$F2_2$
u	21	611.70	1309.10	628.04	1216.01
u	22	512.13	1143.72	512.93	1214.00
u	23	617.64	1254.77	569.06	1356.40
u	24	592.48	1085.49	626.73	1263.14
u	25	467.71	1189.88	648.83	1054.51
u	26	517.79	1120.96	787.79	1206.49
u	27	457.50	1244.55	634.95	1221.02
u	28	483.54	845.56	532.35	1125.74
u	29	610.10	1085.69	750.45	1029.71
u	30	559.41	872.41	658.99	1193.18
u	31	627.35	1194.95	539.60	1105.97
u	32	537.68	1225.29	582.13	1215.57
u	33	633.56	1167.00	570.66	1249.50
ə	1	778.04	1429.57	815.07	1309.53
ə	2	933.16	1368.51	778.55	1452.86
ə	3	838.78	1456.97	775.73	1410.38
ə	4	845.84	1376.02	767.81	1202.38
ə	5	837.69	1318.23	773.15	1347.70
ə	6	853.58	1325.44	774.30	1344.66
ə	7	792.89	1357.82	908.70	1424.09
ə	8	739.79	1464.33	892.30	1500.72
ə	9	862.03	1330.83	892.56	1348.92
ə	10	900.38	1259.24	724.14	1655.74

Vowel	Stimulus Number	$F1_1$	$F2_1$	$F1_2$	$F2_2$
ə	11	872.48	1361.88	879.18	1271.11
ə	12	872.66	1452.25	757.31	1315.89
ə	13	870.38	1371.25	968.21	1342.98
ə	14	811.86	1332.15	779.52	1457.91
ə	15	731.75	1370.80	778.41	1114.93
ə	16	782.63	1412.63	867.42	1376.79
ə	17	755.92	1222.94	830.98	1464.10
ə	18	912.98	1555.66	857.60	1348.75
ə	19	844.28	1379.24	894.19	1411.04
ə	20	793.95	1425.47	889.41	1460.98
ə	21	892.50	1273.67	778.52	1427.90
ə	22	835.49	1457.55	822.42	1232.15
ə	23	838.14	1364.11	846.40	1437.09
ə	24	875.35	1368.97	830.91	1411.47
ə	25	713.76	1359.63	800.36	1456.90
ə	26	816.04	1278.89	790.11	1329.38
ə	27	854.45	1346.35	783.84	1352.50
ə	28	775.38	1406.19	828.07	1404.49
ə	29	725.20	1421.14	828.86	1321.80
ə	30	921.49	1468.46	844.37	1404.76
ə	31	806.55	1428.19	810.61	1221.87
ə	32	857.27	1473.75	788.54	1251.05
ə	33	816.34	1436.13	1002.89	1373.86

Vowel	Stimulus Number	$F1_1$	$F2_1$	$F1_2$	$F2_2$
ɔ	1	817.48	1180.40	962.05	1307.97
ɔ	2	916.21	1184.51	895.59	1345.42
ɔ	3	877.59	1320.87	824.10	1375.18
ɔ	4	864.71	1187.01	772.25	1030.28
ɔ	5	883.97	1336.21	872.48	1183.25
ɔ	6	802.58	1316.06	817.65	1282.49
ɔ	7	983.84	1251.53	813.93	1286.52
ɔ	8	1007.64	1353.82	838.40	1274.36
ɔ	9	891.03	1106.30	887.05	1200.90
ɔ	10	969.07	1294.75	907.94	1203.98
ɔ	11	847.46	1181.31	881.27	1314.16
ɔ	12	836.02	1220.46	831.77	1216.01
ɔ	13	868.59	1209.88	871.72	1072.14
ɔ	14	962.05	1352.61	830.19	1132.39
ɔ	15	860.08	1108.01	820.97	1161.75
ɔ	16	915.51	1208.04	754.32	987.92
ɔ	17	950.40	1207.22	881.48	1291.10
ɔ	18	930.93	1305.06	839.65	1309.09
ɔ	19	865.28	1178.05	897.31	1152.84
ɔ	20	856.43	1305.37	885.41	1306.73
ɔ	21	818.94	1334.20	861.67	1265.92
ɔ	22	896.32	1175.01	934.81	1311.21
ɔ	23	833.61	1353.25	933.28	1175.17

Vowel	Stimulus Number	$F1_1$	$F2_1$	$F1_2$	$F2_2$
ɔ	24	893.05	1276.18	885.94	1251.28
ɔ	25	792.26	1256.88	811.92	1170.87
ɔ	26	844.55	1161.43	786.11	1322.78
ɔ	27	797.29	1231.72	950.55	1247.70
ɔ	28	837.19	1059.21	788.46	1088.21
ɔ	29	881.63	1236.63	907.24	1110.09
ɔ	30	842.52	1303.23	895.89	1337.76
ɔ	31	849.89	1170.27	898.17	1183.91
ɔ	32	819.53	1121.95	898.54	1265.67
ɔ	33	969.85	1378.61	843.25	1279.10
ʊ	1	782.02	1370.06	704.39	1125.69
ʊ	2	798.70	1334.28	728.41	1363.91
ʊ	3	541.88	1382.92	651.37	1250.49
ʊ	4	589.90	1322.57	667.91	1391.61
ʊ	5	749.18	1517.39	633.44	1381.10
ʊ	6	664.00	1310.47	578.44	1482.08
ʊ	7	614.71	1300.29	604.45	1270.63
ʊ	8	775.14	1291.99	568.81	1301.72
ʊ	9	592.10	1274.63	671.36	1429.83
ʊ	10	676.83	1204.38	566.28	1315.29
ʊ	11	587.87	1321.81	641.63	1418.35
ʊ	12	700.70	1268.41	601.65	1296.49
ʊ	13	589.28	1147.13	724.09	1286.97

Vowel	Stimulus Number	$F1_1$	$F2_1$	$F1_2$	$F2_2$
ʊ	14	599.50	1451.79	692.63	1354.48
ʊ	15	653.51	1325.46	673.39	1198.14
ʊ	16	632.96	1383.91	682.88	1213.61
ʊ	17	587.99	1312.97	685.07	1400.92
ʊ	18	644.96	1328.12	702.22	1487.95
ʊ	19	741.97	1294.28	801.21	1345.69
ʊ	20	680.88	1393.00	616.01	1229.91
ʊ	21	594.80	1186.46	714.82	1287.66
ʊ	22	725.78	1290.40	765.44	1187.64
ʊ	23	634.34	1238.83	707.24	1271.25
ʊ	24	689.46	1363.47	636.31	1241.14
ʊ	25	745.17	1496.36	654.86	1373.40
ʊ	26	575.69	1209.14	678.84	1243.04
ʊ	27	698.40	1356.16	786.24	1370.92
ʊ	28	715.29	1380.36	619.61	1272.61
ʊ	29	610.58	1320.78	661.48	1293.96
ʊ	30	615.37	1328.41	615.02	1324.77
ʊ	31	684.25	1290.06	590.71	1351.78
ʊ	32	715.08	1477.17	608.48	1321.48
ʊ	33	602.82	1441.48	627.43	1311.20
ɛ	1	714.29	2023.31	737.81	1978.08
ɛ	2	858.57	1828.94	735.73	2067.48
ɛ	3	843.42	1983.02	716.14	1931.77

Vowel	Stimulus Number	$F1_1$	$F2_1$	$F1_2$	$F2_2$
ɛ	4	785.02	1831.79	717.56	1925.70
ɛ	5	824.45	2247.13	794.91	1981.52
ɛ	6	812.50	1881.05	786.77	2005.95
ɛ	7	823.34	2087.69	908.93	1999.10
ɛ	8	750.06	1903.24	840.28	2037.27
ɛ	9	709.37	2061.12	805.79	1959.86
ɛ	10	791.09	2211.13	714.85	1808.69
ɛ	11	780.72	1880.42	719.18	2040.95
ɛ	12	896.22	1962.31	779.66	2022.99
ɛ	13	737.15	1884.19	758.01	2081.14
ɛ	14	702.39	1997.12	712.91	2014.79
ɛ	15	818.97	1955.48	762.16	1953.44
ɛ	16	778.84	1952.39	776.04	2017.38
ɛ	17	657.33	1782.88	917.18	1931.68
ɛ	18	839.66	2306.16	828.26	2080.94
ɛ	19	745.63	2012.21	725.88	2004.86
ɛ	20	807.04	2023.87	907.36	2025.63
ɛ	21	815.03	1831.29	768.30	2035.36
ɛ	22	931.69	1986.53	741.00	2219.84
ɛ	23	817.23	2025.01	761.54	2200.22
ɛ	24	732.46	1918.32	874.58	2133.09
ɛ	25	687.71	2034.11	717.92	1929.10
ɛ	26	826.86	2210.28	870.40	2168.77

Vowel	Stimulus Number	$F1_1$	$F2_1$	$F1_2$	$F2_2$
ɛ	27	829.54	2053.08	802.63	2175.09
ɛ	28	715.93	1905.50	816.03	1924.62
ɛ	29	748.39	1873.98	705.65	1894.30
ɛ	30	765.75	2148.50	839.39	2278.61
ɛ	31	704.63	1954.25	725.45	1966.33
ɛ	32	649.86	1883.33	861.46	2014.75
ɛ	33	910.08	2219.67	776.37	2069.07
æ	1	829.86	2118.97	967.09	2268.71
æ	2	899.08	2014.48	643.92	2330.25
æ	3	873.14	2147.93	782.99	2298.62
æ	4	874.08	2118.37	798.87	2125.38
æ	5	703.43	2279.33	759.17	1679.76
æ	6	812.89	2237.50	841.22	2286.19
æ	7	740.44	2169.58	806.10	2084.80
æ	8	805.17	2049.80	733.06	2112.05
æ	9	789.46	2111.14	842.14	2489.22
æ	10	891.60	2135.74	871.75	1861.37
æ	11	801.16	2122.49	868.64	2094.28
æ	12	783.68	2182.70	793.31	2291.21
æ	13	801.87	2350.94	736.42	2028.46
æ	14	811.97	2154.86	772.20	2101.17
æ	15	734.57	1882.83	874.81	2025.83
æ	16	759.61	2233.17	794.28	1880.67

Vowel	Stimulus Number	$F1_1$	$F2_1$	$F1_2$	$F2_2$
æ	17	781.19	2407.98	698.84	1978.81
æ	18	848.41	1980.10	693.07	2229.35
æ	19	743.00	2403.31	788.89	2084.58
æ	20	663.57	2187.94	777.86	2026.76
æ	21	751.70	2007.44	709.19	2281.97
æ	22	807.50	1991.05	646.00	2133.09
æ	23	805.28	2127.83	966.38	2145.13
æ	24	763.42	2143.66	731.38	1946.81
æ	25	814.86	2039.60	837.34	2072.54
æ	26	735.04	2144.12	900.41	2361.23
æ	27	777.29	2055.22	747.68	2391.50
æ	28	723.21	2090.80	859.85	2438.04
æ	29	823.83	2434.92	820.76	2245.89
æ	30	689.29	2040.18	922.11	2058.19
æ	31	859.52	1982.89	732.70	2022.86
æ	32	715.24	2185.69	760.49	1812.85
æ	33	852.63	2063.70	771.65	2051.16
ɑ	1	966.88	1455.95	1010.29	1669.34
ɑ	2	923.54	1535.90	892.86	1435.69
ɑ	3	1040.62	1469.76	973.44	1566.22
ɑ	4	962.92	1421.60	1041.12	1190.14
ɑ	5	957.32	1577.16	1014.12	1523.41
ɑ	6	1045.44	1302.91	883.58	1601.05

Vowel	Stimulus Number	$F1_1$	$F2_1$	$F1_2$	$F2_2$
ɑ	7	976.79	1529.40	1021.78	1525.54
ɑ	8	845.75	1561.82	906.80	1366.19
ɑ	9	1063.25	1779.76	923.18	1541.84
ɑ	10	954.95	1500.73	1002.43	1559.12
ɑ	11	963.13	1329.99	913.04	1593.08
ɑ	12	1102.15	1576.16	873.65	1607.95
ɑ	13	925.14	1295.60	1020.95	1590.74
ɑ	14	985.83	1629.23	873.84	1568.88
ɑ	15	1055.51	1625.52	789.06	1579.75
ɑ	16	908.35	1328.29	951.90	1688.57
ɑ	17	1047.70	1708.84	946.51	1463.90
ɑ	18	972.91	1297.03	1129.54	1592.36
ɑ	19	866.80	1509.01	855.29	1465.67
ɑ	20	930.22	1565.22	992.87	1352.30
ɑ	21	897.19	1469.08	949.73	1441.23
ɑ	22	1049.36	1620.52	984.32	1497.57
ɑ	23	924.18	1663.71	963.31	1705.90
ɑ	24	1194.79	1480.53	811.32	1567.46
ɑ	25	899.84	1558.39	919.47	1402.59
ɑ	26	911.74	1385.37	1049.10	1554.01
ɑ	27	1037.70	1432.08	1118.76	1591.32
ɑ	28	888.78	1418.96	839.47	1384.44
ɑ	29	1029.71	1510.03	870.46	1553.64

Vowel	Stimulus Number	$F1_1$	$F2_1$	$F1_2$	$F2_2$
ɑ	30	970.10	1393.63	880.41	1544.87
ɑ	31	837.14	1539.03	945.04	1326.82
ɑ	32	937.17	1397.52	880.41	1630.43
ɑ	33	939.56	1447.48	964.64	1525.07
ɪ	1	695.86	2241.57	710.95	2097.35
ɪ	2	580.79	2337.95	718.52	2417.91
ɪ	3	602.29	2231.35	548.17	2100.31
ɪ	4	623.39	1990.93	454.58	2277.69
ɪ	5	681.01	2364.81	724.99	2038.11
ɪ	6	558.50	2213.59	678.33	2109.57
ɪ	7	538.80	2388.72	629.95	2332.06
ɪ	8	665.29	2235.77	687.92	2148.20
ɪ	9	703.92	2139.10	662.58	2126.04
ɪ	10	557.40	2149.02	593.96	2297.43
ɪ	11	651.29	2249.32	605.15	2168.08
ɪ	12	583.82	2126.48	652.50	2428.41
ɪ	13	620.78	2286.91	600.48	2273.34
ɪ	14	555.74	2201.01	774.69	2051.68
ɪ	15	640.58	2216.76	707.08	2081.53
ɪ	16	633.98	2252.88	705.47	2042.09
ɪ	17	555.89	2150.60	676.21	2192.95
ɪ	18	671.52	2289.94	617.66	2306.16
ɪ	19	623.56	2263.40	644.41	1985.91

Vowel	Stimulus Number	$F1_1$	$F2_1$	$F1_2$	$F2_2$
ɪ	20	591.43	2333.85	673.03	2262.14
ɪ	21	640.32	2197.24	695.18	2203.02
ɪ	22	669.85	2212.82	644.40	2349.07
ɪ	23	740.05	2183.86	523.91	2124.45
ɪ	24	613.33	2325.38	620.77	2047.32
ɪ	25	664.11	2406.85	601.60	2434.46
ɪ	26	620.20	2205.11	670.29	2207.36
ɪ	27	640.15	2387.36	704.25	2231.15
ɪ	28	586.59	2121.77	704.72	2188.02
ɪ	29	601.39	2170.73	641.45	2082.69
ɪ	30	645.73	1874.01	629.60	2187.12
ɪ	31	562.52	2254.11	695.87	1990.12
ɪ	32	568.01	2191.27	540.94	2131.57
ɪ	33	630.70	2291.52	634.53	2289.76