Using acoustic distance to quantify lexical competition

Matthew C. Kelley[1], Benjamin V. Tucker[1]

[1]Department of Linguistics, University of Alberta, Edmonton, AB T6G 2E7, Canada

December 20, 2018

Abstract

The present study quantifies the effects of lexical competition during spoken word recognition using acoustic distance, 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 distance to all other 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 neigbhorhood 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 motiviations for why a researcher may want to use acoustic disinctiveness over phonological neighborhood density in future experiments. [This document was prepared as part of a generals paper course in the fall 2018 term.]

1 Introduction

In the process of spoken word recognition, a listener must discriminate the word contained in an audio signal from 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 on the basis of how well the acoustic information in the signal matches the listener's expectations for each word. These potential matches compete until one receives enough activation that it is recognized. The present study introduces a new measure of competition based on the acoustic distance between words and uses it to model responses to an auditory lexical decision task.

A group of words that sound similar and are expected to compete have been called phonological neighborhoods (Luce, 1986; Luce and Pisoni, 1998). In Luce (1986) and Luce and Pisoni (1998), words are defined to be neighbors on the basis of being one substitution, addition, or deletion away. That is, for a word A, if a word B can be transformed into word A by adding, subtracting, or substituting exactly one phoneme, word B is considered a neighbor to word A. So, *cat*, for example, is neighbors with *bat* but not *bad*.

Competition is then quantified by counting the number of words in the lexicon that are neighbors with a given word. Luce (1986) and Luce and Pisoni (1998) term this competition measure "phonological neighborhood density." It is taken as an indicator of how many words will be competing for activation when a particular word is being heard. It 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 (Luce and Pisoni, 1998; Goldinger et al., 1989; Vitevitch and Luce, 1999; Tucker et al., 2018), where the words with high values are more difficult to process or take longer to respond to. However, facilitatory effects were found in Spanish (Vitevitch and Rodríguez, 2005), where words with a high phonological neighborhood density value were easier to process or faster to recognize. See Vitevitch and Luce (2016) for a review of other tasks that this measure has been used for.

Despite its popularity, when Luce (1986) introduced the notion of phonological neighborhood density, he remarked that a measure more sophisticated than the one-edit definition of neighbors should eventually be used. He notes that the one-edit definition applies equal weight to phoneme substitutions wherever they occur in the word, meaning that *kit* would be considered as similar to *sit* as it is to *kiss*. What's more, equal weight is also assigned to any possible phoneme change, so *pow* would be connsidered to be as close to *bow* as it is to *now*. There have been attempts to replace or modify phonological neighborhood density, to varying efficacy.

1.1 Previously proposed alterantives to phonological neighborhood density

Luce (1986) details more sophisticated methods of quantifying competition, ending on the frequency-weighted neighborhood probability rule. It is an indication of the probability of a listener correctly identifying a word. In addition to how many phonological neighbors a word has, it accounts for the the lexical frequency of the word's neighbors, as well as the confusability of the word's constituent phonemes. Note that this method still uses the one-edit definition for a neighbor that was used in phonological neighborhood density.

A modification to the one-edit definition of neighbors is proposed by Kapatsinski (2005). A word is taken to be a neighbor to another word if at least two thirds of the segments in the words are the same. This is determined by dividing the word in question's number of segments by 3 and rounding down if needed. The result of this division is taken as the maximum number of edits that can be performed to a word before it is no longer a neighbor to the word in question. With this modified definition, only 7% of the words Kapatsinski examined in English had no neighbors, as compared to 58% when using the one-edit definition of neighbors. Notably, this modification keeps the results of historical experiments that used monosyllabic CVC stimuli, since the determined edit distance would still be 1. However, this modification does not address the original concerns expressed about using edit distance as a way of determining whether words are neighbors or not.

Iverson et al. (1998) presents a method of applying the phi-square coefficient to phoneme similarity. The phi-square coefficient is calculated based on experimental phoneme identification data. For two phonemes, it is a measure of how distinct the distributions the phonemes come from are. If it has a value of 0, the two distributions are taken to be identical. And, if it has a value of 1, the two distributions are taken to be distinct. This coefficient allows for perceptual similarity of phonemes to be taken into account when finding some measure of competition. However, this coefficient is often calculated using confusion data based on monosyllabic stimuli. Iverson et al. used C-/ɑ/ for consonants and /h/-V-/d/ for vowels, which is not representative of the larger variety of patterns observed in everyday language use.

Gahl and Strand (2016) also use Iverson et al.'s (1998) measure. Gahl and Strand were comparing phonological neighborhood density and and the phi-square coefficient to see if one was a better predictor of competition. They did slightly expand the type of stimuli the phi-square coefficient has been used on by adding CVC structures. However, it seems the phi-square coefficient only seems to have been used with CVC style stimuli. These kinds of stimuli likely present fewer confounds to account for when calculating the coefficient. Such confounds could be coarticulatory or contextual cues also being present when the phoneme is pronounced. Nonetheless, the phonetic signals that listeners encounter in everyday life involve more than just monosyllabic words. So, the phi-square coefficient as it has been calculated before is not as ecological to the task of spoken word recognition as it could be.

Turning to production, where it's been suggested that lexical competition impacts a speaker's speech, Nelson and Wedel (2017) and Wedel et al. (2018) examined the hyperarticulation of stop consonants and vowels in conversational English. They found that the presence of a minimal pair for a word was a significant predictor of whether a speaker would hyperarticulate the portion of the word that related to the minimal pair. In Wedel et al. (2018), they also found that phonological neighborhood density was not predictive of the acoustic phenomena they were studying. They reason that phonological neighborhood density's lack of predictive power may be due to speakers' attention to phonetic cues rather than form-based repreesentations of words. Nevertheless, 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 that Luce (1986) raised when introducing phonological neighborhood density as an indication of competition.

1.2 Previous measures of similarity

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 distance between segment pairs for each language. Measure of distance between two segments often compare feature vectors and count the number of values that don't match (Saporta, 1955; Mohr and Wang, 1968; Allen and Becker, 2015; Hall et al., 2017), though not all of them use fully-specified features vectors (Frisch et al., 2004; Albright and Hayes, 2006). Still others use different sets of features, like Peterson and Harary (1961) who used hierarchical parameters based on physiology and assigned greater or lesser weights depending on the category of the features. Sanders and Chin (2009), Heeringa (2004), and Kondrak (2000) also use feature sets different from the traditional distinctive feature sets. It bears noting that Kondrak's (2000) metric calculates similarity and not distance like the others metrics and measures discussed here. Some of these comparison measures have been used to create phonetic alignments between phone strings (Allen and Becker, 2015; Albright and Hayes, 2006; Kondrak, 2000), while others were used to compare co-occurrence of consonants (Frisch et al., 2004; Saporta, 1955).

Inherently, one of the challenges of using such featural comparisons is that they are not grounded in perceptual or acoustic evidence. They may very well be useful when creating phonetic alignments (Allen and Becker, 2015; Albright and Hayes, 2006; Kondrak, 2000), but it cannot be assumed *a priori* that these measures will be relevant in perceptual studies. Additionally, while these measures will allow for more gradience in determining how similar two individual segments are, and aligning algorithms like Levenshtein distance can be used to compare similar segments to each other (as in Allen and Becker 2015 and Albright and Hayes 2006), these measures do not all address phonological neighborhood density's shortcoming regarding the position of differences in a word. Those comparison measures that do in some way address the position of the change do so due to the alignment algorithm forcing comparisons between similar segments. Thus, though the outcome of comparing words by using these measures may be affected by the position of segmental differences, it is not due to differences in speech production or perception observed in the data. Rather, it is due to incidental effects of the algorithms used for alignment.

Other researchers comparing linguistic units have instead focused on using acoustic data. In addition to featural comparison between segments, Heeringa (2004) also examined comparing spectrograms to each other and formant tracks (including F1 and F2) to each other. These representations were compared using Euclidean distance. At the word level, the formant representation included F1, F2, and F3. Word-level comparison include a speech rate normalization to ensure a consistent duration for every segment in the word's transcription. The words are then compared by using a modified Levenshtein distance which uses the Euclidean distance between spectrogram slices or formant values as the sbustitution cost. A notable shortcoming of this method for use in perceptual work is its reliance on speech rate normalization, since speaking rate is everpresent in the speech that listeners hear.

Mielke (2012) also introduced a method of calculating phonetic similarity between phone or phoneme categories. It involves converting the acoustic signal from the time representation to a frequency-time representation using Mel-frequency cepstral coefficients. Dynamic time warping is then applied to the two signals in frequency-time representation to calculate the minimal distance between the two signals. The dynamic time warping process works by comparing each time slice in one signal to all the time slices in the other signal that it can be compared to, and it then performs the comparisons that result in the smallest distance between the two signals. The overall distance between the two signals being compared is then taken as the accumulation of the distance between each time slice based on the prevoiusly determiend alignment.

Lewandowski (2012) presented a method of calculating acoustic similarity based on the amplitude envelopes of specific frequency bands of the signals in question. Calculating the similarity between two signals with this method involves finding the amplitude envelopes of 4 frequency bands in the two signals and comparing them using cross-correlation on a dynamic time warped version of them. Cross-correlation is a mathematical function that determines the extent to which two signals match each other, which can be taken as a form of similarity between the signals. Lewandowski used this method in evaluating phonetic convergence between speakers. Overall, it is similar to the Mel-frequency cepstral coefficient method, but only four frequency bands are used, so it provides a coarser representation of the signal. However, it is more interpretable because the frequency bands aren't transformed into cepstral coefficients, which are harder to interpret.

Being that both of these methods operate on the entire acoustic signal, they could both account for the shortcomings of the one-edit definition of neighbors. The type of change should be reflected in the acoustics; if the change is drastic, the acoustic signal will be different, and thus the distance measure will be impacted accordingly. The position or timing of the change can be addressed as well, at least in the sense that acoustic differences based on where a segment is in a word (e.g., word-initial, word-final, before a voiced segment) will also be incorporated into the ditance measure. There is not a straightforward definition of what a neighbor is with these measures, though.

Rather, each individual item in the lexicon would have a distance to every other item in the lexicon.

1.3 The present study

The remainder of the paper presents a measure of lexical competition bsaed on acoustic comparisons between words, and presents the analyses that were run on auditory lexical decision data.

The first analysis is a proof-of-concept where the stimuli that are used in the experiment are compared with each other to determine an overall acoustic distinctiveness value for each word. The second analysis builds on the first but compares different ways calculating the acoustic distinctiveness of a word, including using recordings from speakers that aren't used in the auditory lexical decision and using average sequences from different recordings. These results are compared with a model that uses neighborhood density instead of acoustic distinctiveness. The third analysis investigates the extent to which acoustic distinctiveness and phonological neighborhood density overlap in the models. These anlyses are followed by a general discussion of the results and why a researcher might choose to use acoustic distinctiveness over phonological neighborhood density.

2 Analyses and results

The data that are used in the analysis come from Tucker et al.'s (2018) freely available Massive Auditory Lexical Decision (MALD) data set. MALD is an auditory lexical decision megastudy, and it allows researchers to perform analyses and vritual experiments with a large sample size and vast number of real words. Auditory lexical decision is a behavioral task where participants are presented with a series of audio stimuli and asked to respond on a button box whether each stimulus they are hearing is a real word or not a real word in the target language. In this case, the target language was English. The time it takes to respond to each word is recorded, and the accuracy of the participants' responses is also recorded. These response latencies can be analyzed using standard behavioral chronometry techniques like regression analyses as in Tucker et al.. Further discussion of chronometric methods is discussed in Baayen and Milin (2010).

For the first version of the MALD data set, over 26,000 real words were recorded by a young male speaker of western Canadian English, and 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 reponses to both real words and pseudowords). Stimuli sets have also been recorded for two other speakers: a young female and and 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 measures of acoustic distance detailed later on in the present study. As such, only words that are in common between these three speakers will be used, so that no particular word is left being uncomparable in the different representations developed herein. There are 26,005 words in common between the speakers.

See Tucker et al. (2018) for more information on the recording process for the young male speaker, the auditory lexical decision task, and the variables included in the data set.

2.1 Analysis 1

The first analysis performed was to use the auditory lexical decision stimuli themselves as templates to compare each word against. In this sense, the recordings or some manipulations thereof were taken as a cognitive representation of the word in the lexicon. This analysis using the same stimuli as were used in the experiment as the lexical representation was run as a proof-of-concept so as to make sure that the method worked in the first place and to see how it compares to phonological neighborhood density. Subsequent analyses will address the question of the ecology of the representation, as well as its potential for generalizability.

2.1.1 Calculating acoustic distinctiveness

Only the subset of words with recordings from all three speakers mentioned above were used. 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 frequency representation, similar in some ways to a spectrogram. More specifically, this process involves windowing the signal, calculating Mel filterbanks for each frame from the windowing, and determining the cepstrum coefficients for each filterbank. In the present analysis, a typical format used in speech recognition was used, where the window length was 25 ms, the step size for the windows was 10 ms, and there were 12 coefficients used along with the log energy of the frames for a total of 13 values representing each frame of audio. Unlike in speech recognition, delta and delta-delta coefficients were not calculated. This choice was made on the grounds that the goal is to calculate the distance between 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 at the point in time when you want their distance. Additionally, similarity and distance measures such as Euclidean distance, cosine similarity, and Levenshtein distance do not or cannot make use of derivatives, so the derivatives are not used in the calculation of acoustic distance to be consistent with other measures of similarity or distance.

Once the words were converted to an MFCC-by-time representation, each individual word was compared to all the other words using the fast dynamic time warping algorithm with a radius set to 10 (Salvador and Chan, 2007). Dynamic time warping is the process of finding the ideal warping between two time signals. That is, it takes a signal and a template and matches the time steps in the signal to their closest matches in the template. The restriction is that a time step in the signal cannot be matched to a time step in the template that comes before the previously matched time step in the template. So, if a 10-item sequence is being mapped to a 6-item template, and the 5th item in the 10-item sequence was matched to the 2nd item in the template, the 6th item in the 10-item sequence cannot be matched to the 1st item in the template. At the end of this process, the distance between all of the matches is summed to output a sequence-level cost value of how far apart the sequence is from the template.

The dynamic time warping algorithm is similar to the edit distance or Levenshtein distance calculation used for phonological neighborhood density. Like Levenshtein distance, dynamic time warping calculates addition or deletion costs at each time step and accumulates the costs as it processes further. The difference is that additions and deletions are explicitly penalized with a discrete value in Levenshtein distance calculations. Whereas, in dynamic time warping, the cost for an addition or deletion is determined based on the similarity between frames. For this reason, dynamic time warping is more general and allows for more nuanced comparisons between frames, when working with continuous values. In the present analysis, the MFCC frames take on real values, so dynamic time warping is a fitting choice for the comparisons, while also maintaining some similarity to phonological neighborhood density.

The dynamic time warping process for the number of items in the present analysis involves a large number of calculations and would take an unreasonable amount of time to process using the full dynamic time warping process, so the fast dynamic time warping algorithm with a radius of 10 was used, which produces a good approximation of the dynamic time warping output (Salvador and Chan, 2007).

After comparing each word to all the other words, the mean of its distance to all the other words was calculated. This mean value was taken as an indicator of the word's acoustic distinctiveness, or how distinct it is overall from the other words in the lexicon. So that the calculations were carried out in a reasonable amount of time, they were performed using the `DynamicTimeWarp.jl` package (Fowler, 2016) in the `Julia` programming language (Bezanson et al., 2017).

It was found that one recording, for the word "accounts" had an excessive period of silence at the end, so it was left out of the analysis. Additionally, a handful of words were recorded but not used in the experiment, so they were used in the process of calculating the acoustic distinctiveness for other words, but those words' acoustic distinctiveness values themselves were not able to be used in the modeling process.

2.1.2 Statistical analysis

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 intuition behind this relationship is that acoustic distinctiveness is a measure of how unique a word is from other words, whereas phonological neighborhood density is a measure of how similar a word is to other words. This trend is represented in a linear correlation value of -0.47 between these two variables.

This acoustic distinctiveness values were used as a predictor of response latency in generalized additive mixed models (GAMMs) using the `mgcv` (Wood, 2011) and `itsadug` (van Rij et al., 2017) packages in the `R` programming language (R Core Team, 2018). Response latency was measured from stimulus offset so as to factor stimulus duration out of the respone latency values

Table 1: Table of coefficients for the GAMM

	edf	Ref.df	F	p-value
Trial number	8.12	8.79	127.54	< 0.001
Phonological uniqueness point	5.42	6.38	303.99	< 0.001
Log COCA frequency+1	6.52	7.60	238.01	< 0.001
Log acoustic distinctiveness	4.84	5.99	1148.64	< 0.001

themselves. Only correct responses that were made after stimulus offset were kept. Only correct responses to real words were used in the analysis. 51 words were outliers in terms of acoustic distinctiveness (beyond the 0.998 quantile) and were trimmed from the data as well. Manual inspection of a handful of these words suggested they had untrimmed periods of silence in the recording, meaning that their inflated values of distinctiveness were likely due to extraction errors, providing a strong motiviation to remove them from the data. 95,902 responses remained for the modeling process.

The model fitting process consisted of a forward-fitting 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 analagously but gradually removing complexity. This backward fitting process resulted in a smooth term for age and a parametric term for sex being removed from the model for not contributing to the overall fitness of the model. Age was also checked as a parametric term, but it was not found to contribute significantly enough to the model fitness to warrant the extra complexity. The table of coefficients for the fixed smooth terms can be seen in Table 1. The random effect structure consisted solely of random intercepts by subject. 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, so there would be not much explanatory power added by having the by-item random intercepts, and the potential for overfitting increases. (Some words had only one response once the data were subset, so a by-item random intercept would end up being the exact response time value itself, which is overfitting.)

The smooth for uniqueness point could be taken as linear, and it is are generally decreasing. That is, for words where the uniqueness point is later, the response time is faster. For trial, there is a sharp decrease at the beginning of the experiment, and then a more gradual decrease as the experiment goes on. In general, participants get faster throughout the experiment. In part, this may be due to a learning effect. this this means that the further into the experiment the participant is, the faster they respond. The smooth for frequency is nonlinear, and the trend changes for highly frequent words. For low and middle frequency words, the higher the frequency, the faster the response. However, for higher frequency words, the higher the frequency, the slower the response, which may be due to increased competition at the higher frequency values.

A plot of the smooth effect of acoustic distinctiveness can be seen in Figure 1. The relationship is virtually linear and monotonically decreasing. That is, words that are acoustically similar to other words are responded to slower. Analogously, words that are acoustically distinct from other words are responded to faster. In the frame of competition, words with many potential competitors (words that are acoustically similar to many other words) take longer to be responded to, and words with few potential competitors (words that are more acoustically distinct) take less time to be responded to. This is the same trend as neighborhood density shows, at least for English.

The concurvity was also calculated for this model. The results may be seen in Table 2. Concurvity is a generalization of collinearity for nonlinear trends (Wood, 2011). Since GAMMs model nonlinear trends, it is appropriate to use concurvity here. It uses the same scale as correlation, where a value of 0 means no concurvity and a value of 1 means indiscernability from other smooths. The values taken were the estimate values, which are supposed to be between the pessimistic and optimistic values the concurvity calculation outputs (Wood, 2011). These values are hard to interpret, and the only consensus seems to be concern for values near to 1. As such, in this model, there do not appear to be issues of concurvity among the smooth terms.

There are some implications about 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 derived from physical measurements. The MFCC templates used for calculating acoustic distinctiveness are grounded in the acoustic production of the speaker, and each coefficient in each frame of the template indicates frequency information. If competition were to arise at an abstract 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

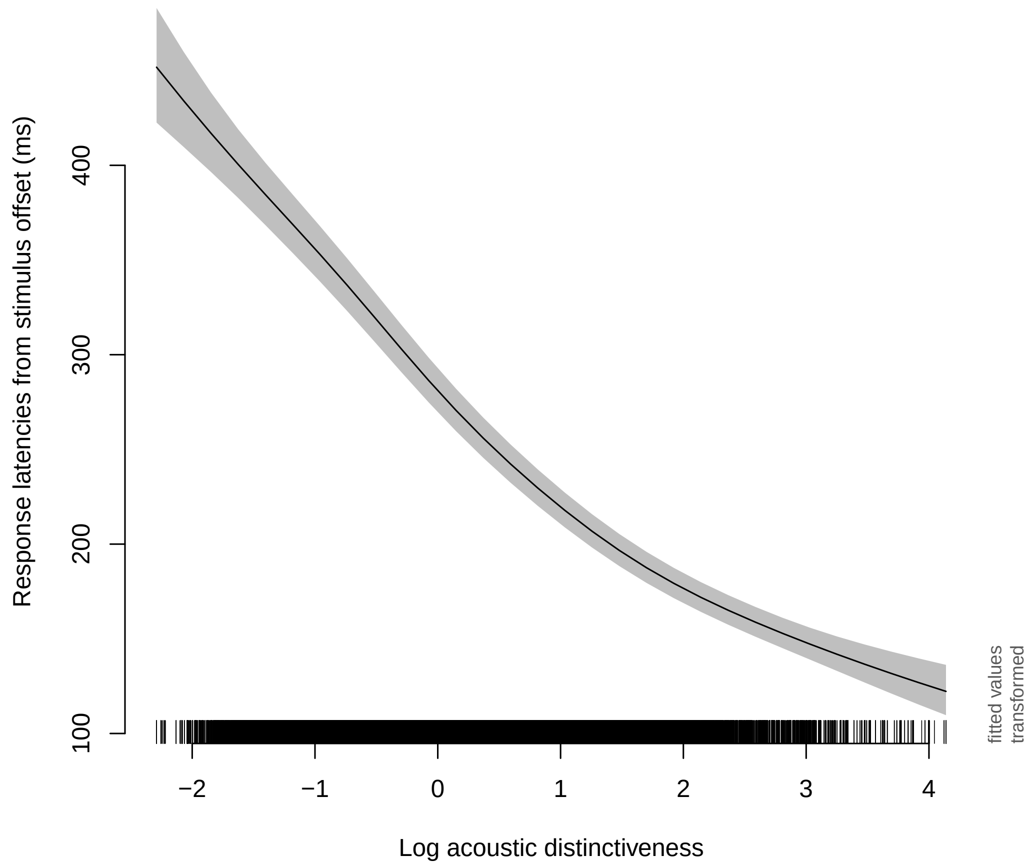

Figure 1: Smooth effect for the log acoustic distinctiveness value. The y-axis is the respone latency after backtransformation from log scale. The x-axis is the scaled and centered log acoustic distinctiveness.

Table 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.

	Estimate concurvity
Trial number	0.02
Uniqueness point	0.21
Log frequency+1	0.08
Log acoustic distinctiveness	0.28
Subject	0.00

cognitive information that is producing the competition effect. However, since acoustic distinctiveness produced a competition-style effect, its effect in this model challenges the the idea that lexical competition only plays out among candidates that are represented by phoneme strings.

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. However, there is a potential shortcoming of using the stimuli themselves as the template against which the stimuli are compared against 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 summed 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 their viability in modeling spoken word reocgnition, with attention also paid to how they compare to phonological neighborhood density.

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 Analyss 1. However, additional speakers' recordings were

Table 3: Example comparisons for finding time series medoid.

	Series 1	Series 2	Series 3
Series 1	0	5	100
Series 2	5	0	10
Series 3	100	10	0

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, two other sets of values were calculated by using the average of the recordings of the words. One set consisted of averages between the female and older speakers' recordings, and the second set consisted of averages between all three speakers. 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 auditory shape, using an average of multiple speakers' recordings should produce a template that is closer to a listner's cognitive representation, providing a better index than a single speaker would.

2.2.1 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. As with the distinctiveness calculations, this process begain by converting each of the recordings to MFCC-by-time sequences. Next, the medoid of the sequence was found. The medoid is a central tendency—like the mean and median—for a set of data. The element in the data set which is closest to all the other elements in the set is taken as the medoid. For a set of time series, it is time series that is closest to all the other time series. For example, given the dummy sequences and distances in Table 3, Series 2 would be taken as the medoid because the sum of its distance to Series 1 and Series 3 is smaller than the sum of Series 1's distances and the sum of Series 3's distances. The distances are found using the dynamic time warping algorithm's ouptut.

Once the medoid is found, it is taken to be the time series that will be modified to find the average sequence. Subsequently, the medoid is mapped onto the other series with dynamic time warping. In effect, each frame of the current average sequence is mapped onto relevant frames in the other time series. After each frame in the current average has been mapped frames in all the series being averaged, each frame in the current average sequence is replaced by the average (or barycenter) of all the frames that were associated with it. The process of mapping to relevant frames and taking the average is repeated iteratively for a user-specified number of times, and the resulting sequence is taken as the average. It has been shown that this process will converge to final values with enough iterations (Petitjean et al., 2011, 2014). For this analysis, 10 iterations were used for finding each average.

2.2.2 Statistical analysis

To compare the effect each of the different methods of calculating the acoustic distinctiveness had on the model, the model from Analysis 1 without the acoustic distinctiveness value was taken as a baseline. 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 fREML is a measure of model fitness that penalizes model complexity. Using this measure helps ensure that the terms used in the model actually contribute meaningfully to the fitness of the model. Lower values are better. The decreases were calculated using the values from the GAMM models themselves. They are presented visually in Figure 2.

There was a decrease in fREML for each method used for deriving the templates to compare the audio simuli against, providing support for both hypotheses. The second hypothesis was not fully supported since using the young male speaker's recordings as the templates produced the greatest increase to model fitness. This is not completely unexpected, though, 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 improved model fitness the least. However, based on the fREML value, it was still a significant increase in fitness from the baseline model. Both templates that included the speaker of the stimuli for the lexical decision task incresed the model fitness the most. This result is not surprising, since the speaker's productions themselves should be the best templates for the words, since they should be the closest to his speaking patterns.

What is more striking is that using the older male speaker's productions as templates does not improve model fitness to the same degree as the other acoustic distinctiveness values. It suggests that the recordings from the older male speaker are different enough acoustically from the stimuli used in the experiment so as to not be good templates for the stimuli. By extension, the larger increases to model fitness from the other methods of template calculation 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 speaker's recordings as the templates to compare the stimuli against produces the greatest increase to model fitness.

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 concurvity value for neighborhood density was 0.34, and the uniqueness point was 0.40. This suggests that there is a moderate degree of overlap in terms of what uniqueness point and neighborhood density are predicting in the model. It is difficult to interpret concurvity values absolutely until they approach 1. The relatively higher concurvity in the model with neighborhood density suggests that neighborhood density is partially explained by other predictors in the model, but it is unlikely that it this level is high enough to raise concern about the model.

In the face of these observations, it is clear that acoustic distinctiveness increases model fitness more so than does neighborhood density. Overall, this means that acoustic distinctiveness is a better predictor of response latencies 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.

What's more, the results suggest that this measure can be generalized to be used in other studies. 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 use input their stimuli to a program that will compare the stimuli to the items in the database and give back their acoustic distinctiveness results.

It is still unclear, though, if acoustic distinctiveness values represent the same kind of information as neighborhood density does. To answer this

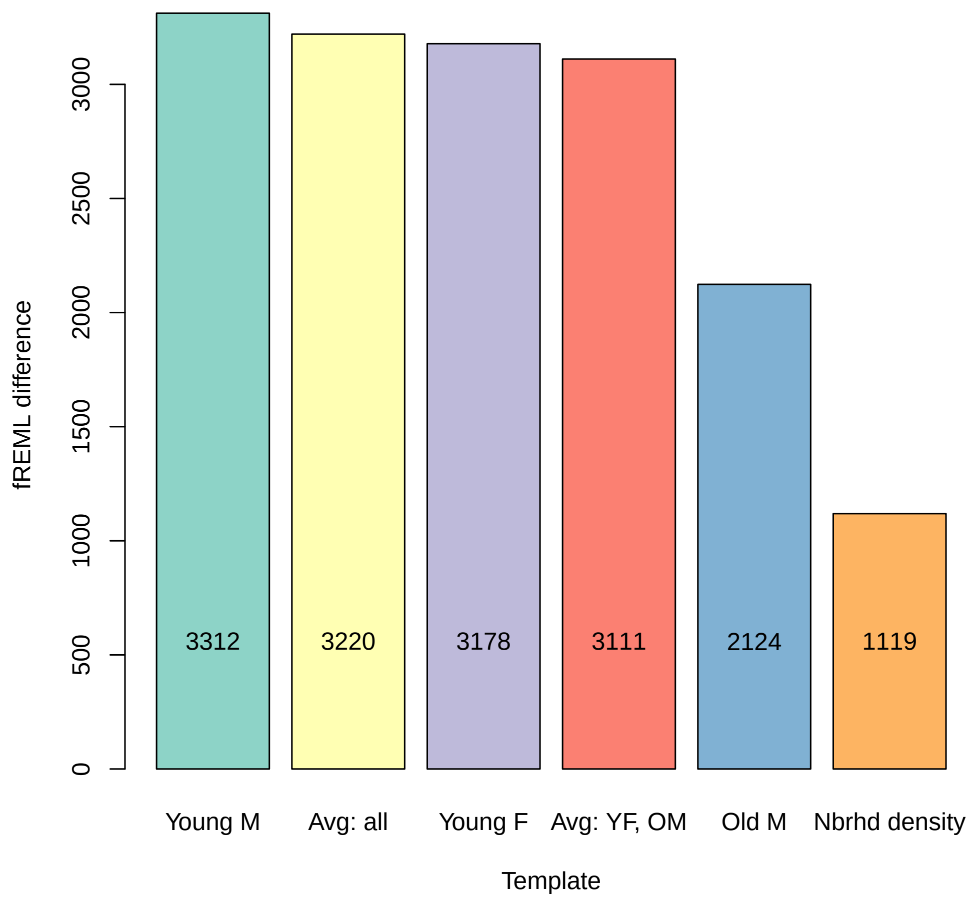

Figure 2: fREML differences between acoustic distinctiveness calculations and neighborhood density. All the changes were decreases, indicating better model fit. Larger values indicate greater increases to model fitness.

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.

2.3 Analysis 3

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

2.3.1 Statistical analysis

Phonological neighborhood density was added to each of the models from from Analysis 2, and the changes in the fREML values were observed. The results of this analysis are presented in Figure 3. Overall, neighborhood density contributed significantly to improving the fitness of all the models, which is taken as evidence against the motivating hypothesis for this analysis.

Note that the level of increase was greatest for the model using the older male speaker's recordings as the template for acoustic distinctiveness. There is a parallel trend 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. This implies that using the older male speaker's productions as the templates for the 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 is seen in the concurvity values. Table 4 shows the estimate concurvity values for the model using the older male speaker recordings as templates and the model using the young male speaker recordings as templates. The better acoustic models produced from the young male result in higher concurvity values for neighborhood density and uniqueness point. Whereas, the less well matching acoustic models from the old male result in lower concurvity values for neighborhood density.

In sum, the more the acoustic representation contained in the templates matches the stimuli, the more that acoustic distinctiveness explains parts neighborhood density's domain. Further against the hypothesis motivating

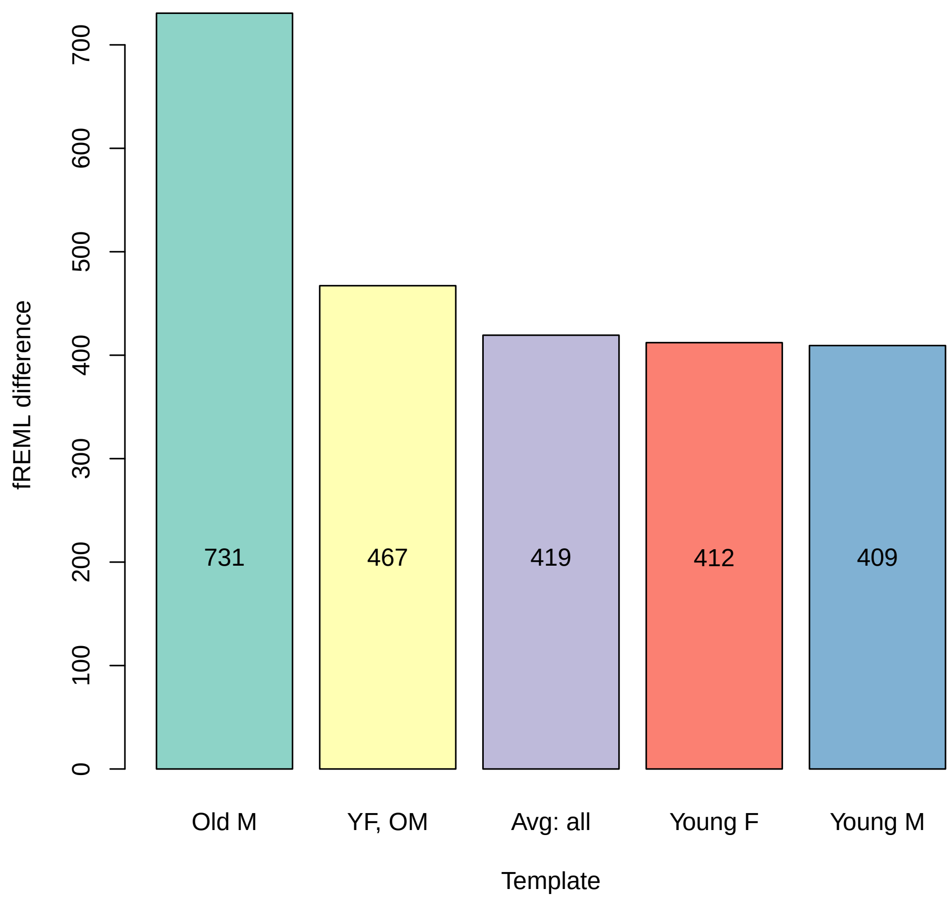

Figure 3: fREML differences between acoustic distinctiveness calculations and neighborhood density. All the changes were decreases, indicating better model fit. Larger values indicate greater increases to model fitness.

Table 4: Estimate concurvity values for the models using the young speaker and the old speaker to create the templates.

	Young male	Old male
Trial number	0.02	0.01
Uniqueness pt	0.44	0.42
Log freq+1	0.17	0.17
Log ac distinctiveness	0.25	0.16
Nbrhd density	0.40	0.37
Subject	0.01	0.01

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. This could be due in part to how neighborhood density's reliance on phonological, phoneme-based representations is multiple degrees divorced from the observed acoustic signal, or that phonological neighborhood density's reliance on phonemes may cause it to be confounded by the effects of orthography. It may also be due to phonemic representations capturing 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 relevant to speech processing at all.

3 General discussion

The overall results presented in the current study are that acoustic distinctiveness is a significant predictor of 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 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. And, in reality, they aren't. Acoustic distinctiveness is measuring an overall tendency of how easily a given word can be distinguished from other words on the basis of acoustics. Phonological neighborhood density is providing an index of approximately how many words there are that sound like a given word. Both of these measures can be interpreted as some indication of lexical competition, though.

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. Nevertheless, because the one-edit-away measure of similarity was seemingly chosen simply as a tool to model lexical competition and not strictly due to theoretical constraints as to how words are represented in the mind, it stands to reason that what is important in any index of lexical competition is that it models trends seen in the data. This is as opposed to necessitating that such an index rest on the same theoretical underpinnings as phonological neighborhood density. As such, it does not appear that what is understood about lexical competition based on phonological neighborhoods or otherwise similar sounding words is tied directly to the characteristics of phonological neighborhood density itself.

It emerges then that 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 instructive 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 are 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 grounds. These grounds may be representational, in that they concern the nature of lexical representations; applicability, in that one of the measures can explain something another can't; statistical, in that one of the measures provides a better fit to the data; or feasible, in that the measure can be calculated easily and efficiently by researchers without being experts in high-performance computing.

Concerning representational grounds for using either acoustic distinctivenes or phonological neighborhood density, 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 have 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. Such an acoustically based representation is similar in essence to those proposed in Port and Leary (2005), Baayen et al. (2016), Ramscar and Port (2016), and Port (2010).

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 necessary for discrete symbols—phonemes—must be recognized, or if some less abstract, acoustic features suffice for representing items in the lexicon. The averaged MFCC sequences represent a level of abstraction between the raw signal and phonemes. For any given word, provided a sufficient number of observations of a word are available, it is likely that the average sequence would converge toward one sequence to represent that word, such that the addition of new observations similar to the representation do little to alter the average sequence. In other words, the sequence is stable. At that point, each element in the sequence could be treated as a symbol with a numerical specification. Note as well that such an acoustic representation of words allows for neighborhood anlyses in the style of phonological neighborhood density if sections of the sequence are treated as symbols.

The reverse, however, is not true; a phonemic representation of words in the lexicon can not be expanded easily into an acoustic representation of the words. And this point leads into the question of the applicability grounds, since the processes of creating the acoustic specifications associated with acoustic distinctiveness are transparent and can be mapped to explaining a variety of lingusitic phenomena.

One such example is when a listener adapts to an unfamiliar speaker. 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. 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. The conclusion in Ohala (1996) highlights some difficulties and potential remedies to finding invariant cues for phonemes, but to date, the constellations of cues that unvaryingly lead to the perception of phonemes are unknown. Additionally, though deep neural networks have seen some success in phoneme recognition. For example, Graves and Schmidhuber (2005) have achieved competitive performance in classifying 25 ms frames as belonging to English phoneme classes, Zhang et al. (2016) have found success in determining the phonemes in an acoustic signal without using time-aligned transcriptions to trian, and Kelley and Tucker (2018) have used phoneme classification to build a forced alignment system. However, it remains unclear what these networks are learning and how to use them as models of cognitive processes.

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, Warner et al. (2004), Gahl (2008) (renalyzed and confirmed by Lohmann (2018)), and Seyfarth et al. (2018) have found production differences in homophones. Additionally, Warner et al. (2004) have 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 grounds for using one of phonlogical 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 of the words, acoustic distinctiveness increased model fit more so than did neighborhood density. There was as well a non-insignificant amount of concurvity in the models once both acoustic distinctiveness and phonological neighborhood density were included in the model. The parts of phonological neighborhood density that were not subsumed by acoustic distincitveness 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 are due to the effects orthography, which has been found to have profound and varied effects on speech perception in a variety of studies, such as in Ziegler and Ferrand (1998), Perre and Ziegler (2008), Taft et al. (2008), and Mukai et al. (2018).

In terms of feasibility, phonological neighborhood density has some factors in its favor. It is conceptually easier to program, especially compared to the average sequencing procedure. It is also used with textual data, which is easier to manipulateand gather, and it takes 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, which will give researchers an accessible programatic interface for calculating it on their stimuli. Additionally, a database could be released for researchers to look up the acoustic distinctiveness for words, or to upload their stimuli to and have them compared against an acoustic database. Those steps still need to be taken for acoustic distinctiveness to be as feasible and convenient a variable as phonological neighborhood density is.

Thus, on representational grounds, where acoustic representations of lexical items can provide more transparent explanations of phenomena than phonemic representations; scope grounds, where acoustic distinctiveness seems applicable to a wider variety of experiments performed in phonetic and linguistic research; and statistical grounds, where acoustic distinctiveness contributes more to model fitness than phonological neighborhood density and does not seem to be confounded with the effects of orthography, it is possible that phonological neighborhood density can be supplanted with acoustic distinctiveness.

Future work should focus on improving the acoustic representation used to model the lexical representation of words. There should also be work done on providing an acoustic specification of phonemes. These specifications can be used to ensure that acoustic representations without phonemes can explain the general tendencies that phonemes do. In doing so, lexical representations can be specified from observed, physical data, rather than hypothesized units like phonemes.

It will also be necessary to use acoustic distinctiveness 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.

4 Conclusion

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 forms, there is now cause to consider reasoning about competition in terms of acoustics. Lexical representations may encode acoustic information, rather than the mind merely using acoustics to get to the lexical representations. Additionally, the timing of the onset of competition effects may be earlier than once thought, and future models of spoken word recognition will need to be intentional in how they depict competition.

If nothing else, the advent of large databases of speech and more powerful computers has ushered in the possibilitty 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 they should.

References

Albright, A. and Hayes, B. (2006). Modelling Productivity with the Gradual Learning Algorithm: The Problem of Accidentally Exceptionless Generalizations. In Fanselow, G., Féry, C., Schlesewsky, M., and Vogel, R., editors, *Gradience in Grammar: Generative Perspectives*, pages 185–204. Oxford University Press, Oxford.

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

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

Baayen, R. H., Shaoul, C., Willits, J., and Ramscar, M. (2016). Comprehension without segmentation: a proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience*, 31(1):106–128.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1):65–98.

Fowler, J. (2016). DynamicTimeWarp.jl. original-date: 2014-12-05T00:59:23Z.

Frisch, S. A., Pierrehumbert, J. B., and Broe, M. B. (2004). Similarity Avoidance and the OCP. *Natural Language & Linguistic Theory*, 22(1):179–228.

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. and Strand, J. F. (2016). Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language*, 89:162–178.

Goldinger, S. D., Luce, P. A., and 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.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610.

Hall, K. C., Allen, B., Fry, M., Johnson, K., Lo, R., Mackie, S., and McAuliffe, M. (2017). Phonological CorpusTools.

Heeringa, W. J. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD thesis, University of Groningen.

Iverson, P., Bernstein, L. E., and Auer Jr, E. T. (1998). Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition. *Speech Communication*, 26(1):45–63.

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

Kapatsinski, V. (2005). Sound similarity relations in the mental lexicon: Modeling the lexicon as a complex network. Progress Report 27, Indiana University Speech Research Laboratory, Bloomington, IN.

Kelley, M. C. and Tucker, B. V. (2018). A comparison of input types to a deep neural network-based forced aligner. In *Proc. Interspeech 2018*, pages 1205–1209.

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 288–295. Association for Computational Linguistics.

Lewandowski, N. (2012). *Talent in nonnative phonetic convergence*. Doctoral dissertation.

Llorach, E. A. (1950). *Fonología española*. 1 edition.

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. Research on Speech Perception. Technical Report No. 6*.

Luce, P. A. and Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and hearing*, 19(1):1–36.

Mielke, J. (2012). A phonetically based metric of sound similarity. *Lingua*, 122(2):145–163.

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

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

Nelson, N. R. and 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.

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

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

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

Petitjean, F., Forestier, G., Webb, G. I., Nicholson, A. E., Chen, Y., and Keogh, E. (2014). Dynamic Time Warping Averaging of Time Series Allows Faster and More Accurate Classification. In *2014 IEEE International Conference on Data Mining*, pages 470–479.

Petitjean, F., Ketterlin, A., and Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693.

Port, R. F. (2010). Rich memory and distributed phonology. *Language Sciences*, 32(1):43–55.

Port, R. F. and Leary, A. P. (2005). Against Formal Phonology. *Language*, 81(4):927–964.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramscar, M. and Port, R. F. (2016). How spoken languages work in the absence of an inventory of discrete units. *Language Sciences*, 53:58–74.

Salvador, S. and Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580.

Sanders, N. C. and Chin, S. B. (2009). Phonological Distance Measures. *Journal of quantitative linguistics*, 16(1):96–114.

Saporta, S. (1955). Frequency of Consonant Clusters. *Language*, 31(1):25–30.

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

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

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., and Sims, M. (2018). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*.

van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2017). itsadug: Interpreting time series and autocorrelated data using gamms. R package version 2.3.

Vitevitch, M. S. and Luce, P. A. (1999). Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition. *Journal of Memory and Language*, 40(3):374–408.

Vitevitch, M. S. and Luce, P. A. (2016). Phonological Neighborhood Effects in Spoken Word Perception and Production. *Annual Review of Linguistics*, 2(1):75–94.

Vitevitch, M. S. and Rodríguez, E. (2005). Neighborhood density effects in spoken word recognition in Spanish. *Journal of Multilingual Communication Disorders*, 3(1):64–73.

Warner, N., Jongman, A., Sereno, J., and 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.

Wedel, A., Nelson, N., and Sharp, R. (2018). The phonetic specificity of contrastive hyperarticulation in natural speech. *Journal of Memory and Language*, 100:61–88.

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.

Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Laurent, C., Bengio, Y., and Courville, A. (2016). Towards end-to-end speech recognition with deep convolutional neural networks. In *Interspeech 2016*, pages 410–414.

Ziegler, J. C. and Ferrand, L. (1998). Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin & Review*, 5(4):683–689.