Using acoustic distance and acoustic absement to quantify lexical competition

Matthew C. Kelley and Benjamin V. Tucker

Citation: The Journal of the Acoustical Society of America **151**, 1367 (2022); doi: 10.1121/10.0009584
View online: https://doi.org/10.1121/10.0009584
View Table of Contents: https://asa.scitation.org/toc/jas/151/2
Published by the Acoustical Society of America

JASA ARTICLE

Using acoustic distance and acoustic absement to quantify lexical competition[a)]

Matthew C. Kelley[b)] and Benjamin V. Tucker[c)]
Department of Linguistics, University of Alberta, Edmonton, Alberta T6G 2E7, Canada

ABSTRACT:
Using phonological neighborhood density has been a common method to quantify lexical competition. It is useful and convenient but 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 to be 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 that of phonological neighborhood density. Additionally, acoustic distinctiveness consistently increases model fitness more than phonological neighborhood density regardless of which kind of acoustic representation is used. However, acoustic distinctiveness does not seem to explain all of the same things as phonological neighborhood density. The different areas that these two predictors explain are discussed in addition to the potential theoretical implications of the usefulness of acoustic distinctiveness in the models. The present paper concludes with some reasons why a researcher may want to use acoustic distinctiveness over phonological neighborhood density in future experiments. © *2022 Acoustical Society of America*.
https://doi.org/10.1121/10.0009584
(Received 6 October 2021; revised 27 December 2021; accepted 28 January 2022; published online 25 February 2022)
[Editor: James F. Lynch] Pages: 1367–1379

I. INTRODUCTION

In spoken word recognition, a listener must discriminate or recognize the word contained in an audio signal from among other potential candidates based on cues from auditory and other modalities. 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 and 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/. The number of edits between phoneme strings is assessed using Levenshtein distance, which is the smallest number of symbol additions, deletions, or substitutions required to convert one string into another string. In this sense, sound similarity between words is assessed using text in the form of phoneme strings. Competition is then quantified by counting the number of words that are neighbors with a given word in the lexicon. 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 and Pisoni, 1998). However, facilitatory effects were found for Spanish (Vitevitch and 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 the phonetic differences that would occur. For example, /kɪt/ would be considered to be as similar to /sɪt/ as it is to /kɪs/. What's more, equal weight is also assigned to any possible segmental change, and, as such, /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 and

[a)]This paper is part of a special issue on Reconsidering Classic Ideas in Speech Communication.
[b)]Also at: University of Washington, Seattle, WA 98195, USA. Electronic mail: matthew.c.kelley@ualberta.ca, ORCID: 0000-0002-7218-5599.
[c)]ORCID: 0000-0001-8965-7890.

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 using 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, ultimately proposing 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 the 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, though, does not resolve the concerns about the timing or type of change to the phonetic signal when assessing sound similarity.

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 and Hayes, 2006; Allen and Becker, 2015; Frisch *et al.*, 2004; Mohr and Wang, 1968). Other feature sets have also been used (Heeringa, 2004; Kondrak, 2000; Peterson and Harary, 1961; Sanders and 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 the phi-square coefficient, which is equivalent to the squared Pearson correlation between binary variables (Howell, 2008). This method was later adopted in Gahl and Strand (2016). However, phoneme confusion data are difficult to extend to the word level, and confusion in simple syllables may not relate well to confusion in longer words because of 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 among different productions of the same vowel.

Johnson (1997) and Yoneyama (2002) created acoustically derived exemplar models of the words. A vector-quantization technique was used 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 the 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 the temporal order. This method was later adopted by 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 the pronunciation distance between words and also used a temporal normalization technique similar to averaging. Dynamic time warping has also been used in McCloy (2013) to align pitch and intensity contours and in Kirchner *et al.* (2010) to create averages of speech exemplars.

A. 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 the phonological neighborhood density did not reflect perceptual similarity, and Yoneyama (2002) reported better performance using more acoustic comparisons. However, there has yet to be a large-scale comparison between phonological neighborhood density and more acoustically grounded methods.

The remainder of this paper describes a measure of lexical competition based on acoustic comparisons between words and analyzes auditory lexical decision data. This paper extends

https://doi.org/10.1121/10.0009584

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.

Considering 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 as 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.

We wish to briefly address some conceptual and terminological concerns about the output from the dynamic time warping algorithm. Some previous work using dynamic time warping has referred to its accumulated cost value output as a "distance metric" (e.g., Bennett *et al.*, 2018). However, in the strict sense of a mathematical distance metric, this label is inaccurate because the output of dynamic time warping does not meet all of the criteria necessary to be a distance metric. Bartelds *et al.* (2020) and Mielke (2012) avoided this terminological problem by finding the 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 caused by 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 value is computed between two chunks of audio. Then, the accumulated distance that is output 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 in which 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 are not 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 times. 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.

II. ANALYSES AND RESULTS

The data 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 27 000 real words recorded by a young male speaker of western Canadian English. Each word was responded to in an auditory lexical decision task 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 in the present study. As such, only words that are common between these three speakers will be used such 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 on the recording process for the young male speaker, the auditory lexical decision task, and the variables included in the data set are available in Tucker *et al.* (2019). The young female and older male speakers were recorded in a similar environment and with similar methods and equipment as for the young male speaker.

A. Analysis 1

The first analysis used the stimuli from the auditory lexical decision task itself as templates to compare against each word. In this way, a word was acoustically represented as the frequency information in its associated recording.

1. Calculating acoustic distinctiveness

Each word was first converted to an MFCC representation similar to that in Mielke (2012). At a high level, this process converts the waveform of the audio into a transform of the frequency representation, which is similar in some

ways to a spectrogram. More specifically, this process involves multiplying the 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 in which the window length was 25 ms and the step size for the windows was 10 ms. Thirteen coefficients were calculated, and the zeroth coefficient was replaced with the log energy of the frames.

The delta and delta-delta coefficients were not calculated, unlike the standard practice in speech recognition and in 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 the time slices in the signals, and it does not make sense to use derivatives in such calculations. To motivate this choice, consider the question of how many kilometers there are between Edmonton and Calgary. A response of "How fast will you be going?" would not address the question because the distance does not depend on the rate of travel.

The choice not to use delta and delta-delta coefficients should not be interpreted as discounting the importance of spectral change on speech perception and spoken word recognition. Indeed, it has already been demonstrated that listeners are sensitive to spectral change in speech (Nearey and Assmann, 1986; Souza *et al*., 2015). Additionally, note that spectral change still comes to bear on the absement value from dynamic time warping because the MFCC representation itself changes over time. Yet, folding rate of change variables into the acoustic representation disturbs the natural kinematic metaphor between distance and absement and makes it more difficult to reason about the variables used in our modeling. We are, as such, being intentionally strict with our terminology, here, to be able to isolate and test the effect of local distance accumulated into global absement on spoken word recognition.

Once the words were converted to an MFCC-by-time representation using the MFCC.JL package (version 0.3.1, van Leeuwen, 2019) in the JULIA programming language (version 1.4.2, Bezanson *et al*., 2017), each individual word was acoustically compared to all of the 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 of the words, the mean of its absement to all of the 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 of 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, 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 acoustic absement. The acoustic distinctiveness value would then be the average connection weight of the word. These calculations were performed using the PHONETICS.JL (version 0.1, Kelley, 2020) and DYNAMICAXISWARPING.JL (version 0.2.5, Bagge Carlson, 2020) packages.

There were some words left over from the recording process for the MALD data set. They were not responded to in the lexical decision task because there were not enough of them to make an additional experimental session. These words were used in calculating acoustic distinctiveness for other words, but the acoustic distinctiveness values of those words themselves were not used in the modeling process.

2. Statistical analysis

The acoustic distinctiveness values correlated highly with the duration of the stimuli ($r = 0.89$, $p < 0.001$). This is to be expected, however. The interval over which the acoustic distances are summed to calculate 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, in this case, absement increases monotonically over time. The high correlation does not mean that these variables are the same, however. Consider that $f(x) = x^2$ and $g(x) = x$ are also very highly correlated 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 km. 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

https://doi.org/10.1121/10.0009584

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.

The acoustic distinctiveness values were used as a predictor of the response latency in generalized additive mixed models (GAMMs) using the MGCV (version 1.8.3, Wood, 2011) and ITSADUG (version 2.3, van Rij *et al.*, 2017) packages in the R programming language (version 3.6.3, *R* Core Team, 2020). GAMMs were chosen to model nonlinear relationships between the variables. We feel that modeling possible nonlinear relationships is especially important when introducing a new variable. Response time was measured from stimulus offset to help factor stimulus duration out of the response time values themselves. These response times were then logged. Only the 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 complexity was gradually removed instead of added. 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 frequency + 1 from the Corpus of Contemporary American English (COCA, Davies, 2008), acoustic distinctiveness, phonological uniqueness point, and log moving average response latency. The phonological uniqueness point of a word is the point at which the 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; Marslen-Wilson and Zwisterlood, 1989). The log moving average response latency is a decaying average of a participant's previous responses. It was calculated using the algorithm from ten Bosch *et al.* (2018) with the α variable set to 0.1 globally. Phonological uniqueness point and log moving average response latency were included in the model as control variables.

The random effect structure consisted solely of a by-subject random intercept. 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 large amount of RAM to run. Additionally, most items had four or fewer responses after subsetting, thus, 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, and, as such, 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 the model fitting), and the model was refit. The table of coefficients for the fixed smooth terms in this model can be viewed in Table I.

TABLE I. The coefficients for the GAMM after model criticism.

Predictor	edf	Ref.df	*F*	*p*-value
Trial number	3.32	4.12	23.22	<0.001
Log COCA frequency + 1	5.76	6.72	333.01	<0.001
Acoustic distinctiveness	5.39	6.59	1154.60	<0.001
Phonological uniqueness point	5.62	6.53	441.76	<0.001
Log moving average response time (RT)	8.41	8.91	540.72	<0.001

The smooths for the control variables were as expected. And, a plot of the smooth effect of acoustic distinctiveness can be observed in Fig. 1. Smooth effect plots for the other effects 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 to many other words are responded to more slowly. Analogously, words that are acoustically distinct from many other words are responded to more quickly. 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 was reported for phonological neighborhood density, at least for English (Luce and Pisoni, 1998). In terms of speech perception, these results

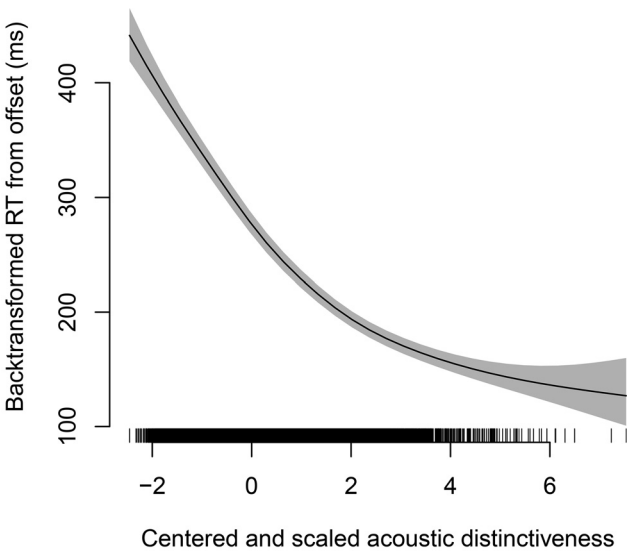

FIG. 1. The smooth effect for the acoustic distinctiveness values when all of the other predictors are held constant. The *y* axis is the response latency after backtransforming from the log scale. The *x* axis is the centered and scaled acoustic distinctiveness. Each point in the function represents how much additional time it would take to respond to a word with that particular acoustic distinctiveness value.

suggest that it takes longer for the competition process to play out in the mind when hearing a word that sounds like many other words.

Concurvity was also calculated for this model. The results are reported in Table II. Concurvity is a generalization of collinearity for nonlinear trends (Wood, 2011). Because 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 zero means no concurvity and a value of one means indiscernibility from other smooths, although the intermediate values cannot necessarily be mapped onto standard correlation thresholds.

Of the three indices that MGCV provides, we chose 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 the supplementary material.[1]

As far as we are aware, there is not yet a systematic way to interpret the concurvity values as indicators of 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 the 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 that of phonological neighborhood density, is explained by all of 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.

There is one predictor for which the measure crosses our threshold—that of the log moving average reaction time. It is a control predictor that does not really relate to the research questions. Thus, it 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 intercept for subject, where the value of the observed index was 0.49. The concurvity between the log moving average reaction time and the random intercept subject is to be expected, though, because the log moving average reaction time is calculated on a by-subject basis.

TABLE II. The estimated concurvity for the smooths in the GAMM model. A value of zero indicates no concurvity and a value of one 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 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 productions 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, because acoustic distinctiveness produced a competition-style effect in the model, it challenges the idea that word-level competition plays out among the candidates represented as symbol strings (e.g., phonemes or diphones) and not acoustics, such as was suggested by the networks in TRACE (McClelland and Elman, 1986) and TISK (You and 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 the 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, whereas 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 the lexical predictors in the model may be more easily interpreted when using acoustic distinctiveness than when using duration due to the 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 templates 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

https://doi.org/10.1121/10.0009584

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 that is currently being listened to. 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.

B. Analysis 2

To answer the question of how using different and multiple speakers to create 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, the recordings of additional speakers were used. These were the previously mentioned young female and older male speakers. Both of these speakers' recordings were used as the template sets for determining the acoustic distinctiveness of the stimuli used in the lexical decision task. Additionally, the values were calculated using each possible combination of the 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 (1) that 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) because a listener has multiple experiences with a given word's acoustic characteristics, using an average of the 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, and (7) the average of all three speakers.

1. Calculating average sequences

The averaging process used was dynamic barycenter averaging (Petitjean *et al.*, 2014, 2011), and this process was designed for time series data, generally. We started with the MFCC-by-time representations described previously. Next, the medoid of the sequence was found. The medoid is a central tendency—similar to the mean and median—for a set of data. It is the element in the data set which is closest to all of the other elements in the set given a cost function. In this case, the absement between the sequences (dynamic time warping cost) was used as the cost function to minimize. Here, the medoid is found by computing all of the pairwise absement values and choosing the recording with 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 the relevant frames in the other time series. Each frame in the current average sequence is then replaced with the average (or barycenter) of all of the frames mapped to it during 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 performed using the AVGSEQ function in the PHONETICS.JL package.

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

2. Statistical analysis

To compare the effects that each of the different methods of calculating acoustic distinctiveness had on the model, the same model from analysis 1 but without the acoustic distinctiveness variable was taken as a baseline model. The acoustic distinctiveness values from the 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 the fREML for each method used to calculate the acoustic distinctiveness as well as for phonological neighborhood density. The magnitudes of these decreases are presented in Fig. 2. The fREML decreases support both hypotheses outlined for this analysis. The second hypothesis was not fully supported, though, because using the young male speaker's recordings as the templates produced the greatest increase to model fitness. This is not completely unexpected as 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 the fitness from the baseline model. Generally, all of the 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 productions of the older male speaker as the templates to compare the experimental stimuli against does not improve the model fitness to the same degree as the other acoustic distinctiveness values. It suggests that the speech of the older male speaker is not a good model for the younger male speaker 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 stimuli. Support for this idea is also found in the fact that using the recordings of the younger male speaker as the templates produces the greatest increase to model

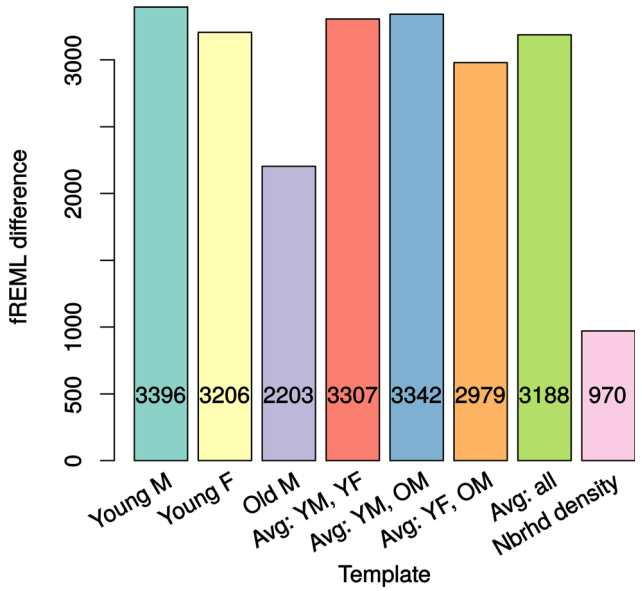

FIG. 2. (Color online) The fREML differences between the acoustic distinctiveness calculations and neighborhood density. All of the changes were decreases, indicating a better model fit. The larger values indicate greater increases to the model fitness. "YM" refers to the young male speaker, "YF" refers to the young female speaker, and "OM" refers to the older male speaker.

fitness. These results also suggest that age differences produce greater acoustic differences in the 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 of the model that used neighborhood density instead of acoustic distinctiveness. In that model, 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 log frequency at a value of 0.23. Overall, these concurvity results suggest 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 due to the uniqueness point and log frequency.

In the face of these observations, it is clear that acoustic distinctiveness increases the 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," which 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 performed, which examined the degree to which neighborhood density further increased model fitness for models that already had distinctiveness values as predictors.

C. Analysis 3

To answer the question of whether the 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.

1. 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 Fig. 3. Overall, neighborhood density contributed significantly to improving the fitness of all of 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 the fREML decrease (that is, the level of the 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 in which using the recordings of the older male speaker as the templates increased the model fitness the least amount compared to the other acoustic distinctiveness values. Together, these results imply, again, that using the productions of the older male speaker as the templates for the productions of the younger male speaker is a worse fit, potentially due to greater acoustic differences between the two speakers.

A similar trend to those from analysis 2 is observed in the concurvity values for the models in the present analysis. The best case for phonological neighborhood density was when it was added to the model using the recordings of the older male as templates. In this case, phonological neighborhood density had an observed concurvity index of 0.53 with

https://doi.org/10.1121/10.0009584

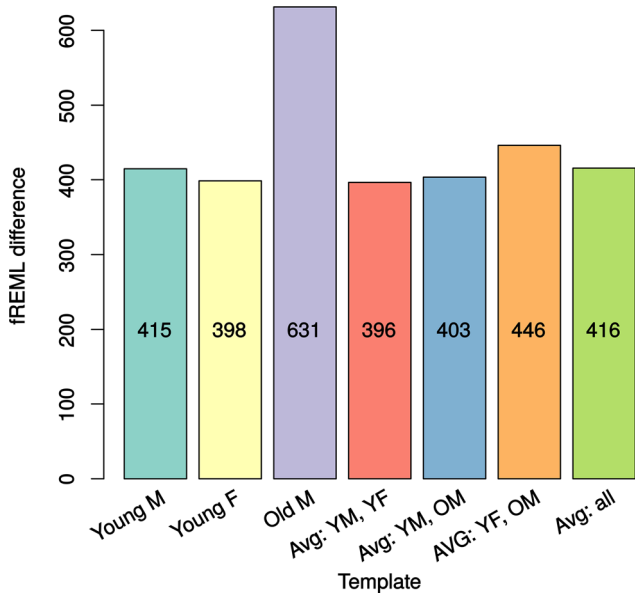

FIG. 3. (Color online) The fREML differences between the acoustic distinctiveness calculations and the neighborhood density. All of the changes were decreases, indicating a better model fit. The larger values indicate greater increases to model fitness.

uniqueness point, log frequency, and acoustic distinctiveness values being the greatest contributors in the pairwise comparisons with values of 0.40, 0.23, and 0.12, respectively. The worst case for phonological neighborhood density was the model using the recordings of the young female as templates, in which 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 the 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 that the acoustic representation contained in the templates matched the stimuli, the more that acoustic distinctiveness explained parts of the effect of the neighborhood density. Further, against the hypothesis motivating this analysis, it may not be possible for acoustic distinctiveness to completely subsume the effect of neighborhood density because 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, which are multiple degrees removed from the observed acoustic signal, whereas acoustic distinctiveness does not. (2) The reliance of phonological neighborhood density 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 the effect of the neighborhood density in the presence of acoustic distinctiveness is still relevant to sound similarity.

III. 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 conceptual and statistical 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. Although 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 of the 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 the initial investigations using phonological neighborhood density, the focus was on examining the role of the structure of words on the 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 because of the 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 observed in the data. Similarly, 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 the 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 the case, it would suggest that they could be operationalizing the same characteristics of words as each other and would be interchangeable for nontheoretical 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 the 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 the 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, as sequences of frequency information. A representation based on acoustics is similar in spirit to approaches to phonetic and psycholinguistic analyses that do not coerce the continuous acoustic or articulatory signal to the discrete symbols (Baayen *et al.*, 2016; Goldinger and Azuma, 2003; Kohler, 1995; Pike, 1943; Port and 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 that 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 to 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 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 and McQueen, 2007; Clarke and 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 that phonemes are the units of lexical representation, possibly as the listener adjusting the weights they have in the connections between the 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) highlighted some of the difficulties and potential remedies to finding invariant cues for phonemes such as looking for cues to diphones or 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 because 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 the differences in production. It would also be applicable to studies examining the effects of speech production on perception, where phonological neighborhood density could not.

Turning now to the statistical reasons for using either phonological neighborhood density or acoustic distinctiveness over the other, the case for acoustic distinctiveness is stronger. The analyses presented in the current study show that acoustic distinctiveness is more predictive than neighborhood density in a variety of different methods of deriving the acoustic representation. Whether using the stimuli that were being presented to the participants, recordings of the same words by different speakers, or averages of the recordings, acoustic distinctiveness increased the model fitness 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. Because 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 and Ziegler, 2008; Taft *et al.*, 2008; Ziegler and Ferrand, 1998). Although, 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.

There are also task-related reasons why one might choose to use acoustic distinctiveness over phonological neighborhood density or, possibly, vice versa. Such a choice

should, in principle, be motivated by the differences in what the two variables represent. Namely, it seems that acoustic distinctiveness is a better fit to the stimuli itself, whereas phonological neighborhood density may be a better fit to more abstract representations of the stimulus. When a task deals more in how acoustic differences contribute to speech perception and spoken word recognition—such as speech in noise tasks—acoustic distinctiveness will more closely relate to the task. As acoustic distinctiveness and absement are calculated using production data, they may also have a clearer link to the production tasks in which certain speech patterns arise based on lexical competition (Gahl, 2015; Wright, 2004). Moreover, effect trends of phonological neighborhood density can change depending on whether it is being used in perception or production tasks or the language it is being used to analyze (Vitevitch and Rodríguez, 2005). Although empirical evidence still needs to be gathered, it is possible that acoustic distinctiveness will provide a more uniform effect across tasks and tested languages. It will, at least, either counterbalance or bolster the effects observed with phonological neighborhood density.

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; as such, 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, such as 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 (2021) for other researchers to be able to use acoustic absement in their own work. We have also added acoustic distinctiveness as a variable to the MALD data set (Tucker *et al.*, 2019).

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, the acoustic representations of lexical items can provide more transparent explanations of phenomena than phonemic representations. In terms of applicability, acoustic distinctiveness seems usable for 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 that the time has arrived to reconsider quantifying lexical competition with the one-edit rule and phonological neighborhood density with some considerations given to the experimental task. The 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 on 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. Alternative representations should also 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. Other representation formats may also allow for more local, fine-grained acoustic differences, such as formant transitions, to be better accounted for. It may also be fruitful to explore the methods used in Kirchner *et al.* (2010). Finally, future research may need to further investigate more explicit measures of spectral change, like total variation as was used in Kelley and Aalto (2019). By so doing, listener sensitivity to the rate of spectral change can be accounted for when necessary.

IV. 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 do not completely account for the same information.

Although competition has often been reasoned about using abstract symbolic forms, acoustic distinctiveness opens the door to thinking about competition in terms of acoustics. Lexical representations may encode acoustic information itself rather than acoustics being a mere tool to get to the abstract symbols used for representation. Similar suggestions about acoustics being part of lexical representations have been made by Johnson (1997), and work like that in Mullennix *et al.* (1989) has highlighted the importance of acoustic information in lexical modeling. On a related note, 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 now 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 winter 2019 University of Alberta Department of Linguistics Generals Paper conference, and 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. Thanks are also owed to Michael Vitevitch, Michael Kiefte, Andrea MacLeod, and Stephanie Archer for their thoughts and comments on previous versions of this study. We also thank two anonymous reviewers for their insightful comments and questions on this work. This project was funded, in part, by Social Sciences and Humanities Research Council Grant No. 435-2014-0678.

[1]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0009584 for additional smooth effect plots from the first model and the explanations of the concurvity indices.

Adank, P., and McQueen, J. M. (**2007**). "The effect of an unfamiliar regional accent on spoken-word comprehension," in *the 16th International Congress of Phonetic Sciences (ICPhS 2007)*, Pirrot, Saarbrücken, Germany, pp. 1925–1928.

Albright, A., and Hayes, B. (**2006**). "Modelling productivity with the gradual learning algorithm: The problem of accidentally exceptionless generalizations," in *Gradience in Grammar: Generative Perspectives*, edited by G. Fanselow, C. Féry, M. Schlesewsky, and R. Vogel (Oxford University Press, Oxford), pp. 185–204, available at http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199274796.001.0001/acprof-9780199274796-chapter-10 (Last viewed 12/20/2018).

Allen, B., and Becker, M. (**2015**). "Learning alternations from surface forms with sublexical phonology," University of British Columbia and Stony Brook University (unpublished).

Baayen, R. H., and Milin, P. (**2010**). "Analyzing reaction times," Int. J. Psychol. Res. **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," Lang., Cognit. Neurosci. **31**(1), 106–128.

Bagge Carlson, F. (**2020**). "DynamicAxisWarping.jl," available at https://github.com/baggepinnen/DynamicAxisWarping.jl (Last viewed 2/23/2021).

Bartelds, M., Richter, C., Liberman, M., and Wieling, M. (**2020**). "A new acoustic-based pronunciation distance measure," Front. Artif. Intell. **3**, 39.

Bennett, R., Tang, K., and Sian, J. A. (**2018**). "Statistical and acoustic effects on the perception of stop consonants in Kaqchikel (Mayan)," Lab. Phonol.: J. Assoc. Lab. Phonol. **9**(1), 9.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. (**2017**). "Julia: A fresh approach to numerical computing," SIAM Rev. **59**(1), 65–98.

Clarke, C. M., and Garrett, M. F. (**2004**). "Rapid adaptation to foreign-accented English," J. Acoust. Soc. Am. **116**(6), 3647–3658.

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

Frisch, S. A., Pierrehumbert, J. B., and Broe, M. B. (**2004**). "Similarity avoidance and the OCP," Nat. Lang. Linguist. 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. (**2015**). "Lexical competition in vowel articulation revisited: Vowel dispersion in the Easy/Hard database," J. Phonetics **49**, 96–116.

Gahl, S., and Strand, J. F. (**2016**). "Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception," J. Mem. Lang. **89**, 162–178.

Goldinger, S. D., and Azuma, T. (**2003**). "Puzzle-solving science: The quixotic quest for units in speech perception," J. Phonetics **31**(3), 305–320.

Goldinger, S. D., Luce, P. A., and Pisoni, D. B. (**1989**). "Priming lexical neighbors of spoken words: Effects of competition and inhibition," J. Mem. Lang. **28**(5), 501–518.

Heeringa, W. J. (**2004**). "Measuring dialect pronunciation differences using Levenshtein distance," Ph.D. thesis, University of Groningen, available at https://research.rug.nl/en/publications/measuring-dialect-pronunciation-differences-using-levenshtein-dis (Last viewed 10/18/2021).

Howell, D. C. (**2008**). *Fundamental Statistics for the Behavioral Sciences*, 6th ed. (Thomson Higher Education, Belmont, CA).

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

Jakobson, R., Fant, G., and Halle, M. (**1952**). *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates*, Technical Report No. 13 (Acoustics Laboratory, Cambridge, MA).

Johnson, K. (**1997**). "The auditory/perceptual basis for speech segmentation," Working Paper, available at https://kb.osu.edu/handle/1811/81782 (Last viewed 9/3/2021).

Johnston, J. D., Dunn, C. J., and Vernon, M. J. (**2019**). "Tree traits influence response to fire severity in the western Oregon Cascades, USA," Ecol. Manage. **433**, 690–698.

Kapatsinski, V. (**2005**). "Sound similarity relations in the mental lexicon: Modeling the lexicon as a complex network," Progress Report 27 (Research on Spoken Language Processing, Bloomington, IN), pp. 133–152.

Kelley, M. C. (**2020**). "Phonetics.jl," available at https://github.com/maetshju/Phonetics.jl (Last viewed 2/10/2022).

Kelley, M. C., and Aalto, D. (**2019**). "Measuring the dispersion of density in head and neck cancer patients' vowel spaces: The vowel dispersion index," Can. Acoust. **47**(3), 114–115.

Kelley, M. C., and Tucker, B. V. (**2021**). "Acoustic absement files," available at https://doi.org/10.7939/r3-mekk-5635 (Last viewed 2/10/2022).

Kirchner, R., Moore, R. K., and Chen, T.-Y. (**2010**). "Computing phonological generalization over real speech exemplars," J. Phonetics **38**(4), 540–547.

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

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, Association for Computational Linguistics*, pp. 288–295, available at http://dl.acm.org/citation.cfm?id=974305.974343 (Last viewed 12/20/2018).

Lewandowski, N., and Jilka, M. (**2019**). "Phonetic convergence, language talent, personality and attention," Front. Commun. **4**, 18.

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

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 6, available at https://eric.ed.gov/?id=ED353610 (Last viewed 9/12/2019).

Luce, P. A., and Pisoni, D. B. (**1998**). "Recognizing spoken words: The neighborhood activation model," Ear Hear. **19**(1), 1–36.

Mann, S., Hao, M. L., Tsai, M., Hafezi, M., Azad, A., and Keramatimoezabad, F. (**2018**). "Effectiveness of integral kinesiology feedback for fitness-based games," in *2018 IEEE Games, Entertainment, Media Conference (GEM)*, pp. 1–9.

Mann, S., Janzen, R., and Post, M. (**2006**). "Hydraulophone design considerations: Absement, displacement, and velocity-sensitive music keyboard in which each key is a water jet," in *Proceedings of the 14th ACM International Conference on Multimedia, MM '06*, Association for Computing Machinery, New York, pp. 519–528, available at https://doi.org/10.1145/1180639.1180751 (Last viewed 10/03/2021).

Marslen-Wilson, W., and Zwisterlood, P. (**1989**). "Accessing spoken words: The importance of word onsets," J. Exp. Psychol.: Human Percept. Perform. **15**(3), 576–585.

McClelland, J. L., and Elman, J. L. (**1986**). "The TRACE model of speech perception," Cognit. Psychol. **18**(1), 1–86.
McCloy, D. R. (**2013**). "Prosody, intelligibility and familiarity in speech perception," thesis, University of Washington, available at https://digital.lib.washington.edu:443/researchworks/handle/1773/23472 (Last viewed 2/10/2022).
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*.
Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**(1), 365–378.
Nearey, T. M., and Assmann, P. F. (**1986**). "Modeling the role of inherent spectral change in vowel identification," J. Acoust. Soc. Am. **80**(5), 1297–1308.
Nelson, N. R., and Wedel, A. (**2017**). "The phonetic specificity of competition: Contrastive hyperarticulation of voice onset time in conversational English," J. Phonetics **64**(Supplement C), 51–70.
Ohala, J. J. (**1996**). "Speech perception is hearing sounds, not tongues," J. Acoust. Soc. Am. **99**(3), 1718–1725.
Perre, L., and Ziegler, J. C. (**2008**). "On-line activation of orthography in spoken word recognition," Brain Res. **1188**, 132–138.
Peterson, G. E., and Harary, F. (**1961**). "Foundations of phonemic theory," in *Proceedings of Symposia in Applied Mathematics*, Vol. 12, pp. 139–165.
Petitjean, F., Forestier, G., Webb, G. I., Nicholson, A. E., Chen, Y., and Keogh, E. (**2014**). "Dynamic time warping averaging of tme series allows faster and more accurate classification," in *2014 IEEE International Conference on Data Mining*, pp. 470–479.
Petitjean, F., Ketterlin, A., and Gançarski, P. (**2011**). "A global averaging method for dynamic time warping, with applications to clustering," Pattern Recognit. **44**(3), 678–693.
Pigoli, D., Hadjipantelis, P. Z., Coleman, J. S., and Aston, J. A. D. (**2018**). "The statistical analysis of acoustic phonetic data: Exploring differences between spoken Romance languages," J. R. Stat. Soc.: Ser. C (Appl. Stat.) **67**(5), 1103–1145.
Pike, K. (**1943**). *Phonetics: A Critical Analysis of Phonetic Theory and a Technic for the Practical Description of Sounds* (The University of Michigan Press, Ann Arbor).
Port, R. F., and Leary, A. P. (**2005**). "Against formal phonology," Language **81**(4), 927–964.
R Core Team. (**2020**). *R: A Language and Environment for Statistical Computing* (*R* Foundation for Statistical Computing, Vienna, Austria), available at https://www.R-project.org/ (Last viewed 2/10/2022).
Sanders, N. C., and Chin, S. B. (**2009**). "Phonological distance measures," J. Quant. Linguist. **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," Lang., Cognit. Neurosci. **33**(1), 32–49.
Souza, P. E., Wright, R. A., Blackburn, M. C., Tatman, R., and Gallun, F. J. (**2015**). "Individual sensitivity to spectral and temporal cues in listeners with hearing impairment," J. Speech, Lang., Hear. Res. **58**(2), 520–534.
Taft, M., Castles, A., Davis, C., Lazendic, G., and Nguyen-Hoan, M. (**2008**). "Automatic activation of orthography in spoken word recognition: Pseudohomograph priming," J. Mem. Lang. **58**(2), 366–379.
ten Bosch, L., Ernestus, M., and Boves, L. (**2018**). "Analyzing reaction time sequences from human participants in auditory experiments," in *Interspeech 2018*, ISCA, pp. 971–975, available at https://doi.org/10.21437/Interspeech.2018-1728 (Last viewed 2/10/2022).
Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., and Sims, M. (**2019**). "The Massive Auditory Lexical Decision (MALD) database," Behav. Res. Methods **51**(3), 1187–1204.
van Leeuwen, D. (**2019**). "MFCC.jl," available at https://github.com/JuliaDSP/MFCC.jl (Last viewed 10/3/2021).
van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (**2017**). "itsadug: Interpreting time series and autocorrelated data using GAMMs," available at https://cran.r-project.org/web/packages/itsadug/index.html (Last viewed 2/10/2022).
Vitevitch, M. S. (**2008**). "What can graph theory tell us about word learning and lexical retrieval?," J. Speech, Lang. Hear. Res. **51**(2), 408–422.
Vitevitch, M. S. (**2022**). "What can network science tell us about phonology and language processing?," Top. Cognit. Sci. **14**(1), 127–142.
Vitevitch, M. S., and Luce, P. A. (**2016**). "Phonological neighborhood effects in spoken word perception and production," Annu. Rev. Linguist. **2**(1), 75–94.
Vitevitch, M. S., and Rodríguez, E. (**2005**). "Neighborhood density effects in spoken word recognition in Spanish," J. Multilingual Commun. 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," J. Phonetics **32**(2), 251–276.
Wood, S. N. (**2011**). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **73**(1), 3–36.
Wood, S. N. (**2020**). "mgcv," available at https://cran.r-project.org/web/packages/mgcv/index.html (Last viewed 3/25/2021).
Wright, R. (**2004**). "Factors of lexical competition in vowel articulation," in *Papers in Laboratory Phonology VI*, edited by J. Local, R. Ogden, and R. Temple (Cambridge University Press, New York), pp. 75–87.
Yoneyama, K. (**2002**). "Phonological neighborhoods and phonetic similarity in Japanese word recognition," Ph.D. thesis, The Ohio State University.
You, H., and Magnuson, J. S. (**2018**). "TISK 1.0: An easy-to-use Python implementation of the time-invariant string kernel model of spoken word recognition," Behav. Res. Methods **50**(3), 871–889.
Ziegler, J. C., and Ferrand, L. (**1998**). "Orthography shapes the perception of speech: The consistency effect in auditory word recognition," Psychonom. Bull. Rev. **5**(4), 683–689.