Revisiting Word Neighborhoods for Speech Recognition

Preethi Jyothi*
Beckman Institute
University of Illinois, Urbana, IL
pjyothi@illinois.edu

Karen Livescu
Toyota Technological Institute at Chicago
Chicago, IL
klivescu@ttic.edu

Abstract

Word neighborhoods have been suggested but not thoroughly explored as an explanatory variable for errors in automatic speech recognition (ASR). We revisit the definition of word neighborhoods, propose new measures using a fine-grained articulatory representation of word pronunciations, and consider new neighbor weighting functions. We analyze the significance of our measures as predictors of errors in an isolated-word ASR system and a continuous-word ASR system. We find that our measures are significantly better predictors of ASR errors than previously used neighborhood density measures.

1 Introduction

An important pursuit for both human and machine speech recognition research is to understand the factors that affect word recognition accuracy. In the substantial body of work on human word recognition, it has been shown that it is harder to recognize words that have many "similar" neighboring words than words with few neighbors (Luce and Pisoni, 1998), and that frequent words are recognized faster and more accurately than are infrequent words (Marslen-Wilson, 1987; Luce and Pisoni, 1998; Vitevitch and Luce, 1999). In the ASR research community, prior work has also investigated various factors that benefit or disrupt recognition. Examples of such factors include word frequency, speaking rate, and prosodic factors (Fosler-Lussier and Morgan, 1999; Shinozaki and Furui, 2001; Hirschberg et al., 2004; Goldwater et al., 2010). There has also been prior work that uses word confusability measures to predict speech recognition errors (Fosler-Lussier et al., 2005; Jyothi and Fosler-Lussier, 2009).

*Supported by a Beckman Postdoctoral Fellowship.

Word neighborhood measures have been studied more heavily for human word recognition than as predictors of ASR errors. Although not studied specifically in prior work (Fosler-Lussier et al., 2005; Jyothi and Fosler-Lussier, 2009), word confusability measures used in predicting ASR errors could be utilized to build word neighborhoods. Goldwater et al. (2010) examine the behavior of certain standard neighborhood density measures as predictors of ASR errors. To our knowledge, this is the only study that explicitly considers word neighborhoods as a potential factor in ASR.

In this work, we investigate word neighborhood measures as predictors of ASR errors. We propose new neighborhood measures that we find to be more well-suited to ASR than standard neighborhood density measures. We also propose a new mechanism to incorporate frequency weighting within the measures. Finally, we analyze the measures as predictors of errors in an isolated-word recognition system and a continuous-word recognition system for conversational speech.

2 Related Work: Neighborhood Density Measures

In much of the prior work in the psycholinguistics literature, the notion of word similarity is quantified by a simple one-phone-away rule: A word w' is a neighbor of word w if w and w' differ by a single phone, via a substitution, deletion, or insertion. We refer to this density measure as "**ND**".

$$\text{ND} = \sum_{w'} \overline{\Delta}_{\text{ND}}(w, w')$$

where $\overline{\Delta}_{\text{ND}}(w, w') = 1$ if w and w' differ by a phone and 0 otherwise.

The frequencies of the neighbors are often accounted for in the neighborhood density measure by computing the sum of the raw (or log) frequencies of a word's neighbors (Luce and Pisoni, 1998; Vitevitch and Luce, 1999); the word frequencies

are derived from a large corpus. We refer to this frequency-weighted measure as "**wND**".

$$\text{wND} = \sum_{w'} \overline{\Delta}_{\text{ND}}(w, w') \cdot \pi(w')$$

where $\pi(w')$ is the frequency of the word w'.[1] Both **ND** and **wND** are popular measures for word neighborhoods that we consider to be our baselines; Goldwater et al. (2010) also make use of these two density measures.[2]

Neither of these measures account for the frequency of the word itself. In continuous ASR, which uses a language model, frequent words are more likely to be recognized correctly (Fosler-Lussier and Morgan, 1999). To account for this, instead of using absolute frequencies of the neighboring words, we use their relative frequencies to define a third baseline density measure,"**rwND**" (relative-**wND**):

$$\text{rwND} = \sum_{w'} \overline{\Delta}_{\text{ND}}(w, w') \cdot \frac{\pi(w')}{\pi(w)}$$

Relative frequencies have appeared in prior work (Luce, 1986; Luce and Pisoni, 1998; Scarborough, 2012). In fact, the measure used by Scarborough (2012) is the reciprocal of rwND.

3 Proposed Neighborhood Measures

Our new neighborhood measures are defined in terms of a distance function between a pair of words, Δ, and a weighting function, β. The proposed measures are not densities in the same sense as **ND, wND, rwND**, but are scores that we may expect to correlate with recognition errors. We define the neighborhood score for a word w as:

$$\text{score}(w) = \sum_{w' \neq w} \beta(w, w') \cdot \Delta(w, w') \quad (1)$$

Intuitively, β is an averaging function that weighs the importance of each neighboring word. For example, Yarkoni et al. (2008) use a neighborhood measure that gives equal importance to the top 20 closest neighbors and rejects the others. The rest of the section presents multiple choices for Δ and β which will define our various neighborhood measures via Equation 1.

3.1 Distance Functions

All of our distance functions are based on an edit distance between a pair of words, i.e., the minimum cost incurred in converting one word to the other using substitutions, insertions and deletions of the sub-word units in the word. In addition to binary edit costs, we consider edit costs that depend on sub-phonetic properties of the phones rather than a uniform cost across all phones. Second, instead of a single pronunciation for a word, we consider a distribution over multiple pronunciations. These distance functions can be easily computed via finite-state transducer (FST) operations, as explained below (see also Figure 1).

Edit Distance (Δ_{ED}): This is the simplest edit distance function that incurs an equal cost of 1 for any substitution, insertion, or deletion. To compute the distance between a pair of words, each word w is represented as a finite state acceptor, F_w, that accepts the pronunciations (phone sequences) of the word. We also introduce a memoryless transducer, T, that maps an input phone to any output phone, with arc weights equal to the corresponding substitution costs (mapping to or from epsilon indicates a deletion or an insertion). The weight of the shortest path in the composed FST, $F_w \circ T \circ F_{w'}$, gives the edit distance between w and w'. When either w or w' has more than one pronunciation, Δ_{ED} is the minimum edit distance among all pairs of pronunciations. This edit distance function has been previously proposed as a measure of phonological similarity between words (Hahn and Bailey, 2005). Similar distance functions have also been used for neighborhood density measures in visual word recognition studies (Yarkoni et al., 2008).

Simple Articulatory Feature-based Edit Distance (Δ_{AF}): The distance function Δ_{ED} penalizes an incorrect substitution equally regardless of the phone identity; for example, the phone [p] can be substituted with [b] or [aa] with equal cost according to Δ_{ED}, although we know it is more likely for [p] to be produced as [b] than as [aa]. To account for this, we adopt a finer-grained representation of the phone as a vector of discrete articulatory "features". Our features are derived from

[1] Here we use raw rather than log frequencies. The baseline density measures in this section perform better with raw rather than log frequencies on our evaluation data. Our proposed measures perform significantly better than the baseline measures using both raw and log frequencies.

[2] Goldwater et al. (2010) also consider the number of homophones (words that share a pronunciation with the target word) and frequency-weighted homophones as additional neighborhood measures. In our data there is insufficient homophony for these measures to be significant, so we do not report on experiments using them.

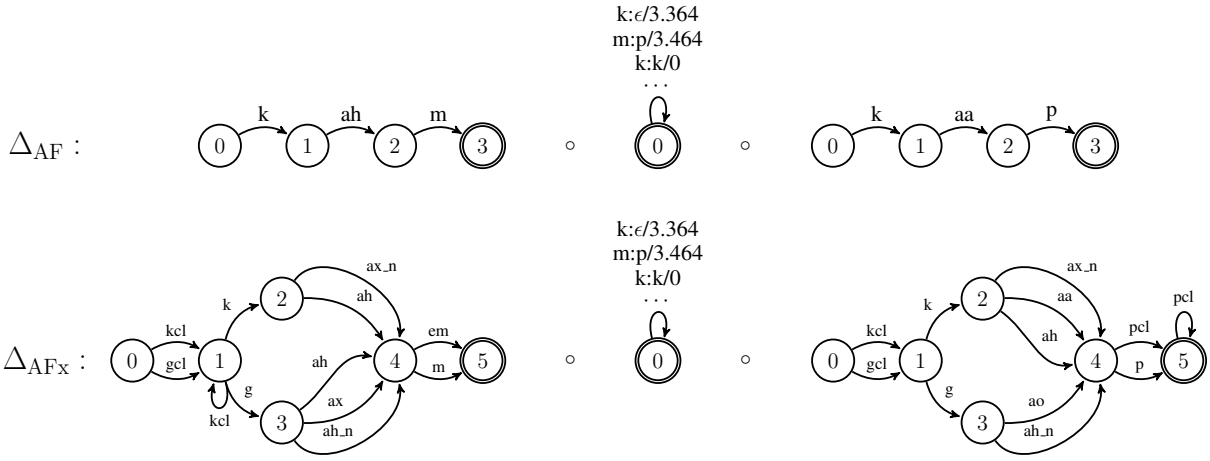


Figure 1: Distance functions implemented using finite-state machines.

the vocal tract variables of articulatory phonology (Browman and Goldstein, 1992), including the constriction degrees and locations of the lips, tongue tip, tongue body, velum and glottis.We borrow a particular feature set from (Livescu, 2005).[3] The substitution cost between two phones is defined as the L1 distance between the articulatory vectors corresponding to the phones. We set the insertion and deletion costs to the mean substitution cost between the articulatory vectors for all phone pairs. These new costs will appear as the arc weights on the edit transducer T. This is shown in Figure 1; apart from the difference in the arc weights on T, Δ_{AF} is the same as Δ_{ED}.

Extended Articulatory Feature-based Edit Distance (Δ_{AFx}): The words in our dictionary are associated with one or more canonical pronunciations written as sequences of phones. The distance functions Δ_{ED} and Δ_{AF} make use of this small set of canonical pronunciations and do not capture the various other ways in which a word can be pronounced. An alternative, explored in some prior work on pronunciation modeling (Deng and Sun, 1994; Richardson et al., 2003; Livescu and Glass, 2004; Mitra et al., 2011; Jyothi et al., 2011), is to model the pronunciation of a word as multiple, possibly asynchronous streams of fine-grained articulatory features, again inspired by articulatory phonology. Such a model can be implemented as a dynamic Bayesian network (DBN) with multiple variables representing the articulatory features in each time frame; please refer to (Livescu and Glass, 2004; Livescu, 2005; Jyothi et al., 2011) for more details. In this approach, deviations from a dictionary pronunciation are the result of either asynchrony between the articulatory streams (accounting for effects such as nasalization, rounding, and epenthetic stops) or the substitution of one articulatory feature value for another (accounting for many reduction phenomena).

Jyothi et al. (2012) describe an approach to encode such a DBN model of pronunciation as an FST that outputs an articulatory feature tuple for each frame of speech. We modify this FST by mapping each articulatory feature tuple to a valid phone as per the phone-to-articulatory-feature mapping used for Δ_{AF} (discarding arcs whose labels do not correspond to a valid phone). The resulting FSTs are used to define Δ_{AFx} by composing with the edit transducer T as in the definition of Δ_{AF}. For computational efficiency, we prune these FSTs to retain only paths that are within three times the weight of the shortest path. The pruned FSTs have hundreds of arcs and $\sim$50 states on average. A schematic diagram is used to illustrate the computation of Δ_{AFx} in Figure 1.

3.2 Weighting Functions

Our weighting functions can be appropriately defined to discount the contributions of words that are infrequent or are very far away. We note here that unlike the density measures in Section 2, the lower the distance-based score for a word (from Equation 1), the more confusable it would be with its neighbors. One approach, as pursued in Nosofsky (1986) and Bailey and Hahn (2001), is to use $\mathrm{score}(w) = \sum_{w'} g(\Delta(w, w'))$ where g is an expo-

[3]The mapping of phones to their articulatory feature values is defined in Appendix B of Livescu (2005). This mapping includes a probability distribution over feature values for certain phones; in these cases, we choose the articulatory feature value with the highest probability.

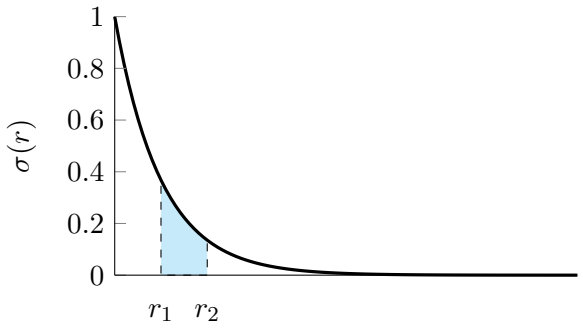


Figure 2: Let w_1 and w_2 be the two closest words to w. The area of the shaded region shows $\beta(w, w_2)$ where $r_i = R_w(w_i) = i$. In the weighted case given in Equation 4, $r_1 = R^\eta_w(w_1)$, $r_2 = R^\eta_w(w_2)$ and $r_2 - r_1 = \eta_w(w_2)$.

nentially decreasing function. This, however, has the disadvantage of being very sensitive to the distance measure used: Slight changes in the distance can alter the score significantly, even if the overall ordering of the distances is preserved. We propose an alternative approach that keeps the score as a linear function of the distances as long as the ordering is fixed. For this, we introduce $\beta(w, w')$ in Equation 1 and let it be a (possibly exponentially) decreasing function of the *rank* of w'.

Formally, we define the rank of w' with respect to w, $R_w(w')$, as follows: Fix an ordering of all $N-1$ words in the vocabulary other than w as $(w_1, w_2, \ldots, w_{N-1})$ such that $\Delta(w, w_i) \leq \Delta(w, w_{i+1})$ for all $i \in \{1, \ldots, N-2\}$. Then $R_w(w') = j$ if $w' = w_j$ in the above ordering. We then define β in terms of a "decay" function σ:

$$\beta(w, w') = \int_{R_w(w')-1}^{R_w(w')} \sigma(r)\mathrm{d}r \qquad (2)$$

If σ is monotonically decreasing, Equation 2 ensures that neighbors with a higher rank (i.e., further away) contribute less weight than neighbors with a lower rank. For example, a measure that gives equal weight to the k closest neighbors (Yarkoni et al., 2008) corresponds to

$$\sigma(r) = \begin{cases} 1 & \text{if } r \leq k \\ 0 & \text{otherwise} \end{cases}$$

Instead of a step function that gives equal weight to all k neighbors, we define σ as an exponentially decreasing function of rank: $\sigma(r) = e^{-r}$. Then, from Equation 2, we obtain $\beta(w, w') = (e-1)e^{-R_w(w')}$. Figure 2 shows the exponentially decreasing $\sigma(r)$ and a sample $\beta(w, w')$.

We know from prior work that it is also important to distinguish among the neighbors depending on how frequently they appear in the language. To account for this, we define a frequency-weighted rank function, $R^\eta_w(w')$:

$$R^\eta_w(w') = \sum_{i=1}^{R_w(w')} \eta_w(w_i) \qquad (3)$$

where η_w is a suitably defined frequency function (see below). We now redefine β as:

$$\beta(w, w') = \int_{R^\eta_w(w')-\eta_w(w')}^{R^\eta_w(w')} \sigma(r)\mathrm{d}r \qquad (4)$$

Note that when $\eta_w(w') = 1$ for all w', Equation 4 reduces to Equation 2. $\beta(w, w')$ is robust in that it is invariant to the ordering used to define rank, R^η_w, i.e. words with the same distance from w can be arbitrarily ordered. Also, multiple words at the same distance contribute to β equally to a single word at the same distance with a frequency that is the sum of their frequencies.

We use three choices for $\eta_w(w')$:

1. The first choice is simply $\eta_w(w') = 1$ for all w'.

2. Let $\pi(w')$ be the unigram probability of w'. We then define $\eta_w(w') = P \cdot \pi(w')$ where P is a scaling parameter. One natural choice for P is the perplexity of the unigram probability distribution, π, i.e., $2^{-\sum_w \pi(w)\log(\pi(w))}$. With this choice of P, when π is a uniform distribution over all words in the vocabulary, we have $\eta_w(w') = 1$ for all w', and $R^\eta_w(w') = R_w(w')$.

3. As defined above, $\eta_w(w')$ does not depend on w. Our third choice for the frequency function considers the frequency of w' relative to w: $\eta_w(w') = \pi(w')/\pi(w)$

To summarize, Equation 1 gives the neighborhood score for w in terms of β and Δ. We use three choices for Δ as specified in Section 3. β is defined by Equation 4 where R^η_w is defined by Equation 3 in terms of the frequency function η_w. We use the three choices described above for η_w. The resulting nine score functions are summarized in Table 1. For completeness, we also include the neighborhood density baseline measures and represent them using our notation with a distance function defined as $\overline{\Delta}_{\mathrm{ND}}(w, w') =$

Measure	$\sigma(r)$	$\Delta(w, w')$	$\eta_w(w')$
ND	1	$\overline{\Delta}_{\text{ND}}$	1
wND			$\pi(w')$
rwND			$\frac{\pi(w')}{\pi(w)}$
ED	e^{-r}	Δ_{ED}	1
wED			$\pi(w') \cdot P$
rwED			$\frac{\pi(w')}{\pi(w)}$
AF		Δ_{AF}	1
wAF			$\pi(w') \cdot P$
rwAF			$\frac{\pi(w')}{\pi(w)}$
AFx		Δ_{AFx}	1
wAFx			$\pi(w') \cdot P$
rwAFx			$\frac{\pi(w')}{\pi(w)}$

Table 1: Summary of neighborhood measures.

$\mathbf{1}(\Delta_{\text{ED}}(w, w') = 1)$ (i.e. $\overline{\Delta}_{\text{ND}}(w, w') = 1$ if $\Delta_{\text{ED}}(w, w') = 1$ and 0 otherwise) and $\sigma = 1$. With $\sigma = 1$ and $\beta(w, w') = \eta_w(w')$, the three choices of η_w give us ND, wND and rwND, as shown in Table 1. The notation $\overline{\Delta}_{\text{ND}}(w, w')$ is to highlight the inverse relationship of the density measures with our distance-based measures.

4 Experiments

We provide an individual analysis of each neighborhood measure as it relates to recognition error rate. We also present a matrix of pairwise comparisons among all of the neighborhood measures with respect to their ability to predict recognition errors. We study the relationship between neighborhood measures and ASR errors in two settings:

- **Isolated-word ASR:** Psycholinguistic studies typically use isolated words as stimuli to study the influence of neighborhood measures on recognition (e.g., see Goldwater et al. (2010) and references therein). Motivated by this, we build an ASR system that recognizes words in isolation and analyze the relationship between its errors and each neighborhood measure. Further details of this analysis are described in Section 4.1.
- **Continuous-word ASR:** ASR systems typically deal with continuous speech. However, the usefulness of neighborhood measures for continuous-word ASR has received little attention, with the notable exception of Goldwater et al. (2010). We further this line of investigation in our second set of experiments by analyzing the relationship between errors made by a continuous-word ASR system and our new measures. These are described in more detail in Section 4.2.

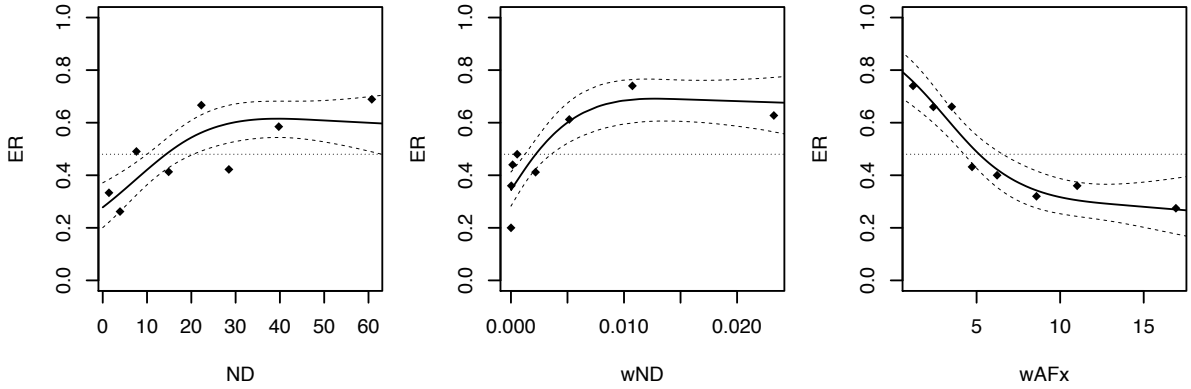
4.1 Isolated-Word ASR

Experimental Setup: We extract isolated words from a subset of the Switchboard-I conversational speech corpus (Godfrey et al., 1992) called the Switchboard Transcription Project, STP (Greenberg et al., 1996; STP, 1996), which is phonetically labeled at a fine-grained level. Isolated words were excised from continuous utterances in sets 20–22 in the STP corpus. We use a total of 401 word tokens (247 unique words) derived from the 3500 most frequent words in Switchboard-I, excluding non-speech events and partial words. These words make up the development and evaluation sets used in prior related work on pronunciation modeling (Livescu and Glass, 2004; Jyothi et al., 2011; Jyothi et al., 2012). We use the dictionary that accompanies the Switchboard-I corpus consisting of 30,241 words; $\sim$98% of these words are associated with a single pronunciation.

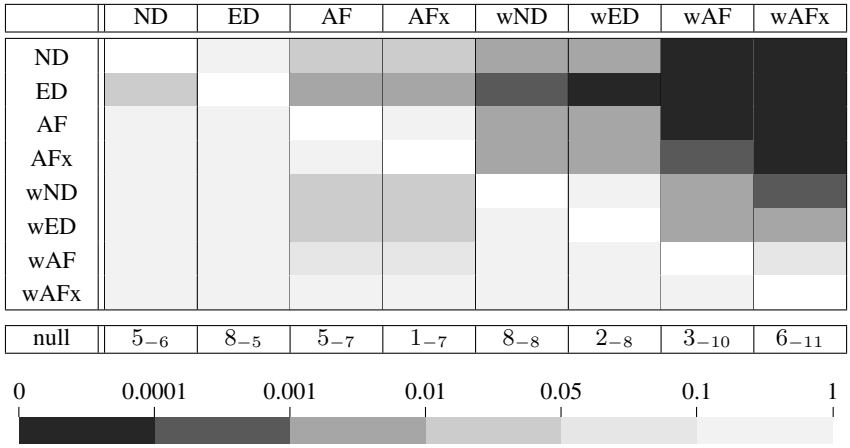
The recognition system for this isolated word dataset was built using the Kaldi toolkit (Povey et al., 2011; Kal, 2011). We use an acoustic model that is trained on all of Switchboard-I, excluding the sentences from which our 401-word set was drawn. The ASR system uses standard mel frequency cepstral coefficients with their first and second derivatives (deltas and double-deltas) as acoustic features, with standard normalization and adaptation techniques including cepstral mean and variance normalization and maximum likelihood linear regression. Linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT) feature-space transformations were applied to reduce the feature-space dimensionality (Povey et al., 2011). The acoustic models are standard Gaussian mixture model-Hidden Markov models (GMM-HMMs) for tied-state triphones. The recognition vocabulary includes 3328 words, consisting of the 3500 most frequent words from Switchboard excluding partial and non-speech words.[4] Since this is an isolated-word task, the ASR system does not use any language model.

Results and Discussion: In order to individually analyze each of the neighborhood measures,

[4]Large-vocabulary automatic recognition of isolated words is a hard task due to the absence of constraints from a language model. Using the entire Switchboard vocabulary would greatly deteriorate the recognition performance on an already hard task. Thus, we restrict the vocabulary to 1/10th of the original size in order to obtain reasonable performance from the isolated ASR system.



(a) Neighborhood measures ND, wND and wAFx as predictors of isolated-word error rate (ER).

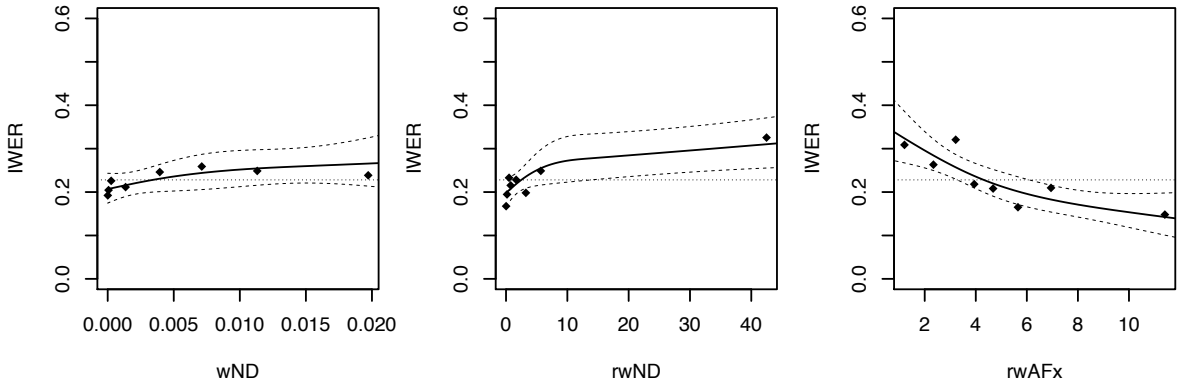


(b) Pairwise comparison of word neighborhood measures as predictors of errors from the isolated-word ASR system using p-values. Many low p-values (darker cells) along a column implies the corresponding measure is a significant predictor of ER.

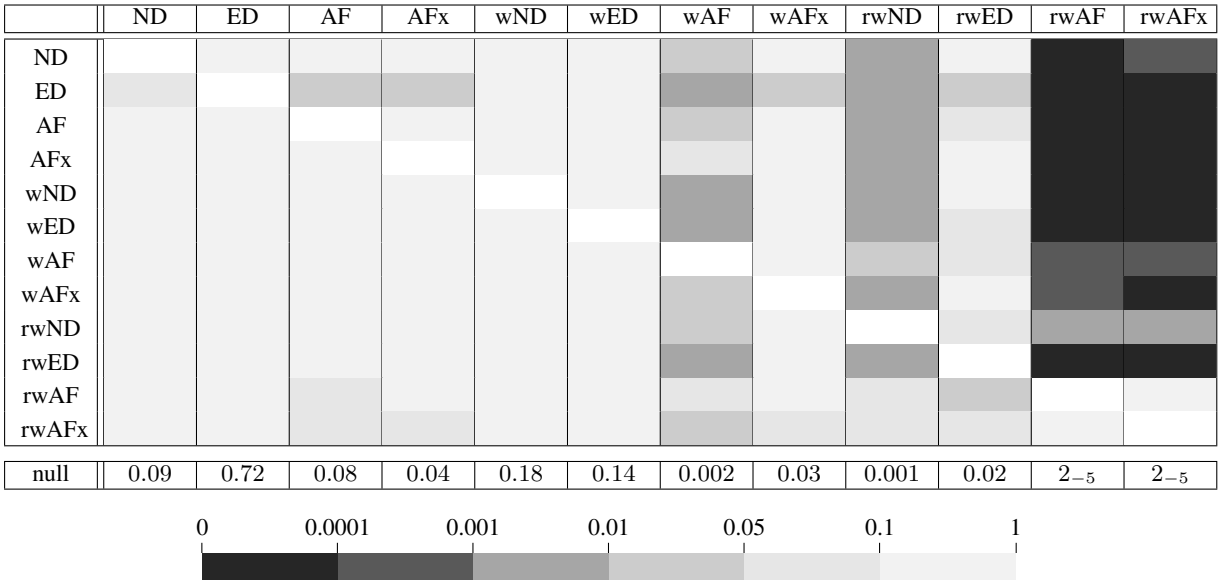
Figure 3: Analysis of neighborhood measures with isolated word ASR.

following Goldwater et al. (2010), we use a logistic regression model implemented using the *glm* function in R (R Development Core Team, 2005). The logistic regression model fits the log-odds of a binary response variable with a linear combination of one or more predictor variables. For our isolated-word task, the response variable takes a value of either 1 or 0 corresponding to the presence or absence of an error, respectively; we will refer to it as "ER". We build a separate logistic regression model for each neighborhood measure acting as the only predictor of ER. We use restricted cubic splines, using the *rcs* (Harrell Jr., 2012) function in R, to model non-linear predictive relationships. In order to determine whether a neighborhood measure is a significant predictor of ER, we use a likelihood-ratio test (using the *anova* function in R) that compares the fit of the model including only that neighborhood measure as a predictor against the fit of a baseline model including only an intercept and no other predictors. All of the neighborhood measures were found to be significant predictors, with our measures wAF and wAFx being most significant. The p-values from this test are shown in a separate row under the header "null" in Figure 3(b); here, 5_{-6} stands for 5×10^{-6} and so forth. We note that the neighborhood measures are significantly correlated with ER as individual predictors, but classifiers built with each individual measure as the only feature are not good predictors of ASR errors. This is unsurprising as we expect many other predictors other than neighborhood measures, as outlined in Goldwater et al. (2010), to influence ASR errors. This paper focuses only on analyzing each neighborhood measure as an individual predictor; joint models will be explored as part of future work.

Figure 3(a) shows the relationship between errors from the isolated ASR system and three neighborhood measures: the best-performing measure (wAFx) and the two standard density measures (ND, wND). The feature values are aggregated into roughly equal-sized bins and the average error rate for each bin is plotted. The



(a) Neighborhood measures wND, rwND and rwAFx as predictors of IWER.



(b) Pairwise comparison of all word neighborhood measures as predictors of IWER from the continuous-word ASR system.

Figure 4: Analysis of neighborhood measures with continuous-word ASR system.

solid line shows the probability of an error from the corresponding logistic regression model and the dashed lines show a 95% confidence interval. The dotted line is the average error rate from the entire data set of 401 words, 0.483. The plots clearly show the inverse relationship between our distance-based measure (wAFx) and the density measures (ND and wND). The slope of the fitted probabilities from the logistic regression model for a measure is indicative of the usefulness of the measure in predicting ER. All of the measures are significant predictors having non-zero slope with a slightly larger slope for wAFx than ND and wND. ND and wND being significant predictors of errors for isolated words is consistent with prior studies from human speech recognition. The proposed measures, wAF and wAFx, stand out as the best predictors of errors. We next analyze the differences between the measures more closely.

Figure 3(b) shows a pairwise comparison of the word neighborhood measures. Each cell $\{i, j\}$ shows a p-value range from a likelihood-ratio test that compares the fit of a logistic regression model using only measure i as a predictor with the fit of a model using both measures i and j as independent predictors. Lower p-values (darker cells) indicate that adding the measure in column j significantly improves the ability of the model to predict ER, as opposed to only using the measure along row i.[5] We use such nested models to compare the model fits using likelihood-ratio significance tests. It is clear from Figure 3(b) that our measures wAF and wAFx are the most significant predictors.

[5]The relative frequency-weighted measures (rwND, rwED, rwAF, rwAFx) were omitted since (wND, wED, wAF, wAFx) are significantly better predictors. This could be because the isolated-word system has no language model and is thus unaffected by the target word frequency.

4.2 Continuous-word ASR

Experimental Setup: For the continuous-word task, our evaluation data consists of full sentences from Switchboard-I that were used to extract the isolated words in Section 4.1. For our analysis, we include all the words in the evaluation sentences that are 3 or more phonemes long and occur 100 times or more in the training set. This gives us a total of 1223 word tokens (459 word types).

The continuous-word ASR system uses an acoustic model trained on all of Switchboard-I excluding the above-mentioned evaluation sentences. The acoustic models are GMM-HMMs for tied-state triphones using MFCC + delta + double-delta features with LDA and MLLT feature-space transformations and speaker adaptation. They are also trained discriminatively using boosted maximum mutual information training from the Kaldi toolkit. We use the entire Switchboard vocabulary of 30,241 words and a 3-gram language model trained on all of the training sentences. The word error rate on the evaluation sentences is 28.3%.[6]

Results and Discussion: Unlike the isolated-word task, the continuous-word ASR system gives word error rates over full utterances. Since we need to measure the errors associated with the individual words, we use the individual word error rate (IWER) metric proposed by Goldwater et al. (2010). The IWER for word w_i is $\alpha \cdot \text{in}_i + \text{del}_i + \text{sub}_i$ where in_i is the number of insertions adjacent to w_i; del_i or sub_i is 1 if w_i is either deleted or substituted, respectively. α is chosen such that $\alpha \cdot \sum_i \text{in}_i = \text{I}$ where I is the total number of insertions for the entire dataset.

As in the isolated-word task, we fit logistic regression models to analyze the neighborhood measures as predictors of IWER. Figure 4(a) shows fitted probabilities from a logistic regression model for IWER built individually using each of the measures wND, rwND and rwAFx as predictors. The number of frequency-weighted neighbors, wND (as well as the number of neighbors, ND), was not found to be a significant predictor of IWER. This is consistent with the findings in Goldwater et al. (2010) that show weak correlations between the number of frequency-weighted neighbors and the probability of misrecognizing a word. However, we find that using the number of frequency-weighted neighbors relative to the frequency of the word (rwND) improves the correlation with the probability of error (seen in Figure 4(a) as an increase in slope). Using our proposed distance measures with relative frequency weighting improves the correlation even further.

Figure 4(b) shows a pairwise comparison of all measures in Table 1; the interpretation is similar to Figure 3(b). We observe that the relative frequency-weighted measures (rwND, rwED, rwAF, rwAFx) are consistently better than their unweighted (ND, ED, AF, AFx) and frequency-weighted (wND, wED, wAF, wAFx) counterparts, with rwAF and rwAFx being most significant. This suggests that the relative frequency-weighted measures are taking precedence in the continuous-word task as significant predictors of IWER (unlike in the isolated-word task) due to the presence of a strong language model.

5 Conclusion

In this work, we propose new word neighborhood measures using distances between words that employ a fine-grained articulatory feature-based representation of the word. We present a new rank-based averaging method to aggregate the word distances into a single neighborhood score. We also suggest multiple ways of incorporating frequency weighting into this score. We analyze the significance of our word neighborhood measures as predictors of errors from an isolated-word ASR system and a continuous-word ASR system. In both cases, our measures perform significantly better than standard neighborhood density measures.

This work reopens the question of whether word neighborhood measures are a useful variable for ASR. There are many possible directions for future work. Our measures could be refined further, for example by exploring alternative distance measures, different articulatory feature sets, different choices of σ and η in the weighting function, or automatically learned costs and distances. Also, our analysis currently looks at each neighborhood measure as an individual predictor; we could jointly analyze the measures to account for possible correlations. Finally, it may be possible to use neighborhood measures in ASR confidence scoring or even directly in recognition as an additional feature in a discriminative model.

[6] The training set includes other utterances from the same speakers in the STP evaluation utterances. This allows for an additional boost in performance from the speaker adapted acoustic models during recognition. Ideally, the training and evaluation sets should not contain utterances from the same speakers. We allow for this to get word error rates that are more comparable to state-of-the-art results on this corpus.

References

T. M. Bailey and U. Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44(4):568–591.

C. P. Browman and L. Goldstein. 1992. Articulatory phonology: An overview. *Phonetica*, 49(3-4):155–180.

L. Deng and D.X. Sun. 1994. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *The Journal of the Acoustical Society of America*, 95(5):2702–2719.

E. Fosler-Lussier and N. Morgan. 1999. Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, 29(2):137–158.

E. Fosler-Lussier, I. Amdal, and H-K. J. Kuo. 2005. A framework for predicting speech recognition errors. *Speech Communication*, 46(2):153–170.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. of ICASSP*.

S. Goldwater, D. Jurafsky, and C. D. Manning. 2010. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

S. Greenberg, J. Hollenback, and D. Ellis. 1996. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In *Proc. of ICSLP*.

U. Hahn and T. M. Bailey. 2005. What makes words sound similar? *Cognition*, 97(3):227–267.

F. E. Harrell Jr. 2012. RMS: Regression Modeling Strategies. R package version 3.5-0.

J. Hirschberg, D. Litman, and M. Swerts. 2004. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43(1):155–175.

P. Jyothi and E. Fosler-Lussier. 2009. A comparison of audio-free speech recognition error prediction methods. In *Proc. of Interspeech*.

P. Jyothi, K. Livescu, and E. Fosler-Lussier. 2011. Lexical access experiments with context-dependent articulatory feature-based models. In *Proc. of ICASSP*.

P. Jyothi, E. Fosler-Lussier, and K. Livescu. 2012. Discriminatively learning factorized finite state pronunciation models from dynamic Bayesian networks. In *Proc. of Interspeech*.

2011. Kaldi. `http://kaldi.sourceforge.net/`.

K. Livescu and J. Glass. 2004. Feature-based pronunciation modeling with trainable asynchrony probabilities. In *Proc. of ICSLP*.

K. Livescu. 2005. Feature-based Pronunciation Modeling for Automatic Speech Recognition. *PhD Dissertation, MIT EECS department*.

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

P. A. Luce. 1986. Neighborhoods of words in the mental lexicon. *Research on Speech Perception*, (Technical Report No. 6.).

W. D. Marslen-Wilson. 1987. Functional parallelism in spoken word-recognition. *Cognition*, 25(1):71–102.

V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein. 2011. Articulatory information for noise robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1913–1924.

R. M. Nosofsky. 1986. Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39.

D. Povey, A. Ghoshal, et al. 2011. The Kaldi speech recognition toolkit. *Proc. of ASRU*.

R Development Core Team. 2005. R: A language and environment for statistical computing. *R foundation for Statistical Computing*.

M. Richardson, J. Bilmes, and C. Diorio. 2003. Hidden-articulator Markov models for speech recognition. *Speech Communication*, 41(2-3):511–529.

R. A. Scarborough. 2012. Lexical confusability and degree of coarticulation. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*.

T. Shinozaki and S. Furui. 2001. Error analysis using decision trees in spontaneous presentation speech recognition. In *Proc. of ASRU*.

1996. The Switchboard Transcription Project. `http://www1.icsi.berkeley.edu/Speech/stp/`.

M. S. Vitevitch and P. A. Luce. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, 40(3):374–408.

T. Yarkoni, D. Balota, and M. Yap. 2008. Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5):971–979.