INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

The effect of phoneme distribution on perceptual similarity in English

Emma O'Neill, Julie Carson-Berndsen

University College Dublin, Ireland
emma.l.oneill@ucdconnect.ie, julie.berndsen@ucd.ie

Abstract

This paper investigates the extent to which native speaker perceptions regarding the similarity between phonemes of English are influenced by their distributional properties. A similarity hierarchy model based on the distribution of consonantal phonemes in the English language was generated by creating phoneme-embeddings from contextual information. We compare this to similarity models based on phonological feature theory and on native speaker perception. Characteristics of the perception-based model are shown to appear in the distribution-based model whilst not being captured by the feature-based model. This not only provides evidence of similarity perceptions being influenced by distributional properties but is an argument for incorporating distributional information alongside phonological features when modelling perceptual similarity.

Index Terms: phonology, phoneme similarity, distribution, perception

1. Introduction

The concept of phonological similarity is an integral part of language analysis. Studies into the sound patterns of languages often refer to natural classes where phonemes form groupings based on having some characteristic in common. Phonological models that use constraint ranking rely on judgements regarding the faithfulness (or similarity) of an output to its input [1]. Often phonemes or natural classes are discussed in terms of the phonological features they exhibit such as those proposed by Chomsky and Halle [2]. Modelling phoneme similarity with respect to such features is highly dependent on exactly which features are used to distinguish between phonemes and the method in which the feature differences are calculated. Bailey and Hahn [3] compared a number of similarity models including a feature-based model proposed by Frisch [4]. They concluded that their simple model involving counts of feature mismatches across 4 features was the best overall model for predicting similarity judgements. This model is discussed in Section 3 and is used for comparative purposes in this work.

Despite Bailey and Hahn's [3] claims about the superiority of their model, it has been noted that a purely feature-based description of similarity between phonemes can contrast with observed phoneme groupings which are thought to be driven by physiological and perceptual factors [5]. Furthermore, Gallagher and Graff [6] state that perceptual similarity "does not necessarily coincide with natural classes or feature values". In other words, there exists a disparity between native speakers' perceptions of phoneme similarity and similarity as defined in terms of phonological features.

Perceptual similarity relates to a native speaker's intuition regarding the sound system of their language. Modelling this relies on examining how speakers produce or recognise these sounds. One particular approach looks to what is deemed acceptable in the context of rhyming words. A syllable can be broken down into its onset, nucleus, and coda; the consonants occurring before the central vowel, the vowel itself, and the consonants following, respectively. This vowel and following consonant cluster is known as the rhyme and when two words end with the same set of sounds they are considered perfect rhymes (*line* and *mine* for example). However, in the world of poetry or lyrics, writers often employ half-rhymes, words with similar but not identical rhymes in the final syllable (*line* and *time* for instance) [7] [8] [9]. Half-rhymes are a convenient mechanism for investigating how similar two sounds are perceived to be since there exists a gradient in the half-rhyme acceptability that a pair of words exhibit (most English speakers would consider *line* and *time* to be more acceptable than say *line* and *light* despite both pairs of words only differing in a single consonant). Thus we can gather that N and M are more perceptually similar than N and T.

Another way of gathering similarity judgements lies in the idea that similarity is a function of confusability [6]. Thus, when a speaker mistakenly identifies one phoneme for another, we can consider the two must be similar. A confusion matrix capturing the frequencies of phoneme identification errors, like that reported in Cutler et al. [10], can be used to generate a model of perceptual similarity as demonstrated in [7]. This is the method we use in this paper to generate the perception-based similarity hierarchy discussed in Section 4.

While there has been some level of investigation into the differences between perceptual and feature-based models of similarity and their respective performances at predicting observed phonological usage [3] [7], there has been little study into what causes this difference. This paper examines one potential source of this discrepancy; namely, how distributional patterns of phonemes influence a native speaker's perception of phoneme similarity.

There exists evidence to suggest that perception and distribution are linked. For instance, in a recent study by Scharenborg et al. [11] it was demonstrated that mental representations of sounds can adapt based on exposure to deviant input. A sound halfway between R and L was presented to listeners in either an R-context or an L-context and the listeners adapted the particular phoneme representation accordingly. Exposure to particular sounds in particular contexts can affect how a phoneme is perceived. That being the case it is natural to look at how the everyday distribution of phonemes has impacted judgements of similarity.

Section 2 describes the process of generating a similarity hierarchy of English consonants based on their distributional properties. Sections 3 and 4 discuss the hierarchies produced from existing feature- and perception- based models taken from Bailey and Hahn [3] and Cutler et al. [10] respectively. In Section 5 we present evidence of how particular aspects of the distribution-based model can be seen to have influenced perception in areas where it does not align with the more traditional feature specification.

http://dx.doi.org/10.21437/Interspeech.2019-3042

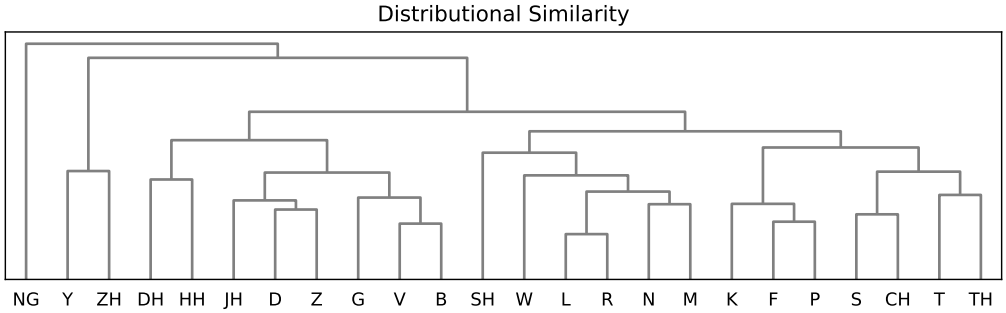

Figure 1: *Similarity hierarchy of the distribution-based model.*

2. The distribution-based model

Phoneme embeddings have recently emerged as an area of interest for speech segmentation and for the determination of sound analogies [12] [13] [14]. It has also been demonstrated that the use of speech data and factors outside of traditional phonological features result in embeddings that better capture phonological similarity [15] [16]. To model distributional similarity we generate phoneme embeddings from their observed environments. This approach is motivated by the success of context-based word-embeddings in capturing semantic similarity between words; in particular the work of Mikolov et al. and their Word2Vec model [17]. In our case, rather than modelling words based on the sentences in which they occur and working under the assumption that semantically similar words appear in similar contexts, we model phonemes based on their syllabic environments assuming that distributionally similar phonemes will occur in similar contexts.

2.1. The data

To generate the phonemic environments with which to train the model we use a phonemically translated text-corpus. Each word in the Brown Corpus [18] was translated to its ARPAbet form using the CMU dictionary[1], resulting in approximately 1 million strings of English phonemes. For words with multiple possible pronunciations one was chosen at random and for words which did not appear in the dictionary a grapheme-to-phoneme tool was used[2]. By doing so we prevent over exposure to words with many pronunciation possibilities which would impact the frequency effects.

The strings of phonemes were then split into syllables using a syllabification tool created for this purpose.The syllabification tool uses the phonotactic constraints of English, including the Sonority Sequencing Principle [19] and the Maximum Onset Principle [20], to obtain a corpus of phoneme strings. This corpus contained approximately 1.5 million syllables with 4 million phoneme tokens, 2.5 million of which were consonants. This syllabification process is illustrated in Figure 2.

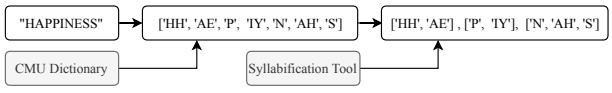

Figure 2: *Syllabification process for the word "happiness"*

The reason we build this model at the syllable level is because this is where the phonotactic rules of English dictate the possible sequences of phonemes. Native speakers possess intuitive knowledge regarding well-formed and ill-formed words in their language and particular combinations of sounds are not permitted. Examining the impact of distributional properties means looking at where the environment of a phoneme is constrained. Across syllable boundaries there are minimal restrictions on phoneme sequences (although there are effects on phonetic realisations which are beyond the focus of this work).

2.2. The model

To generate the phoneme embeddings a simple skip-gram model was used which seeks to maximise the probability of the phonemic environment given the target phoneme. This produced multi-dimensional vector representations of all 39 phonemes in the vocabulary although later analysis will concentrate solely on the 24 consonantal phonemes. The skipgram model was implemented using gensim's Word2Vec tool [21] generating 20 dimensional vector representations using a window size of 1 phoneme either side of the target. This was considered an appropriate context window given the relative size of syllables compared to that of sentences used in word embeddings where the default window size is usually 5-10 words.

One important difference between our phoneme-embedding model and that of a typical word-embedding model lies in the vocabulary of the data set. While a word corpus exhibits a great number of items that occur extremely infrequently, the syllable corpus consists of 1.5 million syllables constructed from a small fixed set of 39 phonemes. As a result, methods used in word-embedding generation to combat the sparsity of items are not used for phoneme-embedding generation. In particular this means that there is no subsampling of frequently observed phonemes and we do not make use of negative sampling where 'false' environments are randomly generated to supplement the limited number of environments observed for infrequent items.

To visualise the similarities captured by this model, the Euclidean distance between the phoneme-embeddings was calculated and the Ward clustering method [22] was used to generate a similarity hierarchy. The cosine distance is typically used for word-embeddings since it eliminates vector magnitude effects and thus words occurring in similar environments, but with differing frequencies, are still considered similar. However, when looking at the distributional properties of phonemes it is desirable to preserve these frequency effects and for this reason the Euclidean distance is used instead. The resulting similarity hierarchy can be seen in Figure 1 where phonemes are clustered according to how distributionally similar they are.

[1] https://github.com/cmusphinx/cmudict
[2] https://github.com/cmusphinx/g2p-seq2seq

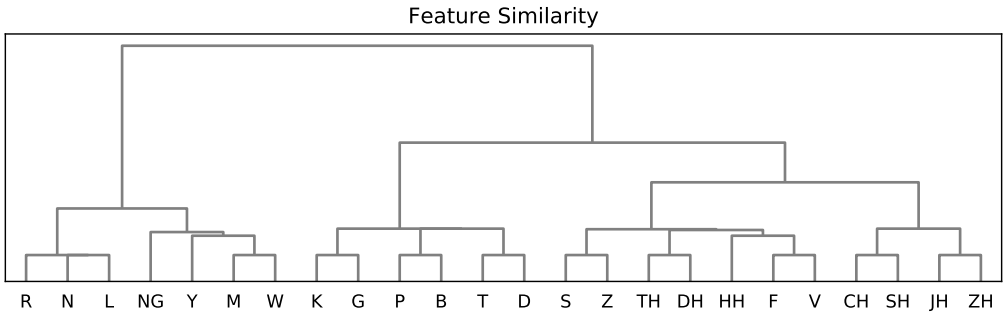

Figure 3: *Similarity hierarchy of the feature-based model.*

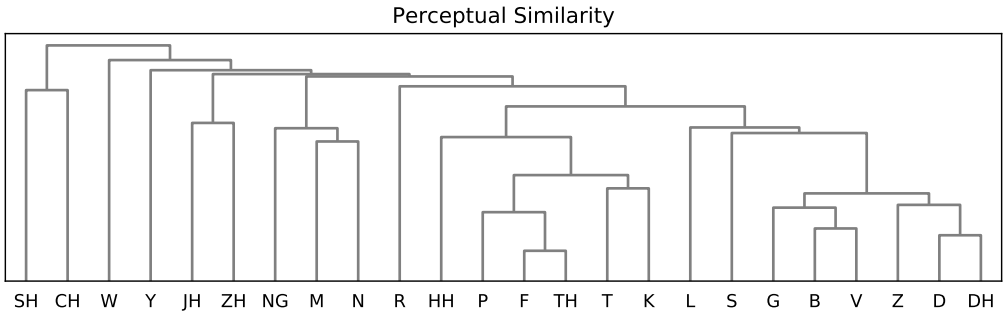

Figure 4: *Similarity hierarchy of the perception-based model.*

3. The feature-based model

There are a number of suggested models regarding which features should be used to distinguish between phonemes and the significance of each. Work by Bailey and Hahn [3] found that the "best measure of phoneme similarity [...] is based on simple counts of the number of major articulatory features [...] in which the two phonemes fail to match". In light of this we used their feature classification in the feature-based model for analyses in Section 5.

The model uses 4 equally weighted distinguishing features;

- Place of articulation: labial, dental, alveolar, palatal, velar, or glottal.
- Manner of articulation: stop, fricative, nasal, glide, lateral, rhotic, or affricate.
- Sonority: sonorant or obstruent.
- Voicing: voiced or voiceless.

Using the distance matrix reported in [3] the Ward clustering method was again applied to produce a similarity hierarchy as shown in Figure 3. Here we see the first distinguishing feature being that of sonority with the most fine-grained differences typically relating to the voicing property. This similarity hierarchy is in agreement with most theoretically based descriptions of phonemes and even resembles the groupings seen in the IPA [23].

4. The perception-based model

As discussed in Section 1, perceptual similarity can be considered a function of confusability [7]. Thus the perception-based model was generated using data from a phoneme confusability experiment carried out by Cutler et al. [10]. As part of this work, native speakers were asked to identify phonemes in consonant-vowel or vowel-consonant pairings at differing levels of background noise. The background noise induced errors in the phoneme identifications and we can assume that listeners are more likely to mistake a phoneme for one which is considered alike. As a result, the frequency with which a phoneme is mistaken for another is directly correlated with how perceptually similar they are.

The confusion matrices of the consonant sounds presented in Cutler et al. [10] were collated and used as phoneme-embeddings. The similarity hierarchy shown in Figure 4 was then generated using the same Euclidean distance metric and Ward clustering method as with the other models. Later analysis in this paper assumes this hierarchy to be representative of a native speaker's perception of phoneme similarity though it should be noted that confusability may not perfectly represent perceptual similarity. Future work in this area will seek to gather similarity judgements through more task-specific experimentation.

5. Discussion

In this section we examine particular attributes of the perception-based similarity hierarchy which appear to have been influenced by the distributive properties of phonemes and which the feature-based model failed to capture.

5.1. The nasal sounds: M, N, and NG

Within the perception-based model (Figure 4 we see the clustering of the nasal sounds: M, N and NG. Of the three sounds NG appears to be the more dissimilar of the three as evidenced by its leaf node being further away from the other two. This judgement also aligns with half rhyme usage where M/N rhymes are much more common than M/NG or N/NG [7]. This cluster is

absent in the feature-based model (Figure 3) since these sounds group more closely with phonemes that share their place of articulation rather than their manner of articulation (M is most similar to W another labial sound while N is considered more similar to the other alveolar sonorants R and L). If we were to more heavily weight the manner of articulation over the place of articulation we would expect to see the cluster of nasal sounds in this hierarchy. However, this is unlikely to exhibit the slight distinction between NG and the other two nasal sounds that we see in the perception-based model and in half-rhyme usage. Conversely, this distinction between M and N on the one hand and NG on the other is magnified in the distribution-based model (Figure 1). This model judges M and N to be very similar to each other while NG is extremely removed. It is fair to say that this is a result of the unique property of NG only being found in a syllable coda and never in an onset. This distributional property is likely the reason why there is greater distance between NG and the other nasal sounds in the perception-based model.

5.2. The palatal sounds: SH, CH, ZH, and JH

Within the feature-based model (Figure 3) there is a branch containing the pairings of palatal sounds: CH and SH, and JH and ZH. These pairings are also seen in the perception-based model (Figure 4) but the degree of similarity between the phoneme in each pair and between the pairs themselves is much less (as evidenced by the length of the branches connecting them). In the perception-based model the palatal sounds appear not to constitute a distinct clustering but instead are rather dissimilar phonemes in relation to the entire vocabulary. The average rate of occurrence in the dataset for a consonant is 4%. However, the palatal sounds fall significantly below this value with each of them making up less than 1% of the total number of phonemes in the syllable corpus. The frequency with which a particular phoneme is encountered appears to impact the strength of mental associations between it and other phonemes. The palatal sounds being relatively 'rare' to a speaker seem to result in the similarity between them and their closest neighbours being much weaker than say that of the more frequent T and K.

5.3. The voicing contrast

Perhaps the most prominent characteristic of the perception-based model (Figure 4) is the distinct separation between voiced and voiceless phonemes. The two main clusters referred to in this subsection are the branch containing the voiced sounds [G, B, V, Z, D, and DH] and the branch containing the voiceless sounds [HH, P, F, TH, T, and K]. These consist of the stops and fricatives of English minus the palatal fricatives discussed in subsection 5.2. In the feature-based model (Figure 3) voicing is typically the most fine-grained distinguishing feature between a pair of phonemes that have all other features in common. As a result we often see a phoneme's most similar neighbour being its voicing counterpart: P and B, K and G, F and V etc. This is a characteristic common not only to this particular feature-based model but to many including even the standard IPA chart of phonemes where voiced and voiceless consonant sounds are paired up based on common place and manner of articulation. The perception-based model (Figure 4) implies that a difference in voicing is much more distinguishing than is captured by the feature-based model (Figure 3) and this difference is also seen in the distribution-based model (Figure 1). Here we see two clusters almost identical to the two seen in the perception-based model. In particular there is the branch containing voiced sounds [JH, D, Z, G, V, and B] and the branch containing voiceless sounds [K, F, P, S, CH, T, and HH]. The voicing of stops and fricatives is evidently an important aspect of the sound in terms of its phonemic environment and the phonotactic rules surrounding it. In turn this has led to a significant perceptual distinction between these groups of sounds.

6. Conclusions

This paper has demonstrated that native speaker perceptions of phoneme similarity are not fully captured by a phonological feature-based model and that there is evidence to support the idea that perceptions are influenced by the distributional features of phonemes. Distinct areas where the perceptual similarities differ from those of the feature-based model were highlighted and such differences were shown to be explainable by a distribution-based similarity model. Future work will look to carry out more task-specific experimentation for gathering perceptual judgements beyond those of confusability and to combining the distribution-based model with acoustic information with the aim of generating a purely data-driven model that best captures phonological similarity.

7. References

[1] A. Prince and P. Smolensky, *Optimality Theory: Constraint interaction in generative grammar*. John Wiley & Sons, 2008.

[2] N. Chomsky and M. Halle, *The sound pattern of English*. New York: Harper & Row, 1968.

[3] T. M. Bailey and U. Hahn, "Phoneme similarity and confusability," *Journal of Memory and Language*, vol. 52, no. 3, pp. 339–362, 2005.

[4] S. Frisch, "Similarity and frequency in phonology," Ph.D. dissertation, Northwestern University, 1996.

[5] J. Mielke, "A phonetically based metric of sound similarity," *Lingua*, vol. 122, no. 2, pp. 145–163, 2012.

[6] G. Gallagher and P. Graff, "The role of similarity in phonology," *Lingua*, vol. 2, no. 122, pp. 107–111, 2012.

[7] S. S. Johnsen, "Rhyme acceptability determined by perceived similarity," in *Paper presented at the 29th West Coast Conference on Formal Linguistics*. University of Arizona, 2011.

[8] A. M. Zwicky, "Well, this rock and roll has got to stop. junior's head is hard as a rock," in *Papers from the... Regional Meeting. Chicago Ling. Soc. Chicago, Ill.*, no. 12, 1976, pp. 676–697.

[9] S. Kawahara, "Half rhymes in japanese rap lyrics and knowledge of similarity," *Journal of East Asian Linguistics*, vol. 16, no. 2, pp. 113–144, 2007.

[10] A. Cutler, A. Weber, R. Smits, and N. Cooper, "Patterns of English phoneme confusions by native and non-native listeners," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3668–3678, 2004.

[11] O. Scharenborg, S. Tiesmeyer, M. Hasegawa-Johnson, and N. Dehak, "Visualizing phoneme category adaptation in deep neural networks," *Proceedings of Interspeech, Hyderabad, India*, 2018.

[12] J. Ma, Ç. Çöltekin, and E. Hinrichs, "Learning phone embeddings for word segmentation of child-directed speech," in *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, 2016, pp. 53–63.

[13] A. Parrish, "Poetic sound similarity vectors using phonetic features," in *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2017. [Online]. Available: https://aaai.org/ocs/index.php/AIIDE/AIIDE17/paper/view/15879/15227

[14] M. P. Silfverberg, L. Mao, and M. Hulden, "Sound analogies with phoneme embeddings," *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pp. 136–144, 2018.

[15] Y.-A. C. J. Glass, "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech," *Proceedings of Interspeech, Hyderabad, India*, 2018.

[16] Y. Do and R. K. Y. Lai, "Measuring phonological distance in a tonal language: An experimental and computational study with cantonese," *Proceedings of the Society for Computation in Linguistics*, vol. 2, no. 1, pp. 371–372, 2019.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space. corr abs/1301.3781 (2013)," *arXiv preprint arXiv:1301.3781*, 2013.

[18] W. N. Francis and H. Kučera, *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown)*. Providence, Rhode Island: Brown University, 1964.

[19] E. O. Selkirk, "On the major class features and syllable theory," in *Language Sound Structure*, M. Aronoff and R. Oehrle, Eds. Cambridge, Massachusetts: The MIT Press, 1984.

[20] D. Kahn, *Syllable-based generalizations in English phonology*. Routledge, 1980.

[21] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[22] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[23] International Phonetic Association and others, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.