A cognitively grounded measure of pronunciation distance

Martijn Wieling1,3, Jelke Bloem2, John Nerbonne3, R. Harald Baayen1,4

1University of Tübingen, 2University of Amsterdam, 3University of Groningen, 4University of Alberta

wieling@gmail.com, j.bloem@uva.nl, j.nerbonne@rug.nl, harald.baayen@uni-tuebingen.de

1 Introduction

Obtaining a suitable difference measure (i.e. distance) between two pronunciations is important, not only for dialectologists who are interested in finding the relationship between different dialects (e.g., Heeringa, 2004), but also for language researchers investigating the relationship between the world’s languages (e.g., Bakker et al., 2009), or those investigating second language acquisition (e.g., Flēge et al., 2006). Obtaining distances between word pronunciations enables quantitative analyses in which the effect of various factors can be investigated.

A commonly used automatic measure of pronunciation distance is the Levenshtein distance (Levenshtein, 1965) which calculates the minimal number of insertions and deletions to transform one phonetically transcribed string into the other. While improvements have been proposed to make the method phonetically sensitive (Wieling et al., 2012), it suffers from two important drawbacks. The first is that there is no cognitive basis for using the Levenshtein distance as a pronunciation distance measure. The second is that the Levenshtein distance does not allow asymmetric distances (generally characterizing perceptual pronunciation distances; Gooskens and Heeringa, 2004).

Here we propose a new method, Naive Discriminative Learning (NDL; Baayen et al., 2011), which does not suffer from these drawbacks. The idea behind this approach (grounded in human learning theory; Rescorla and Wagner, 1972) is that we model how well a listener understands meaning when listening to a speaker with a certain accent. In this model, the past experience (i.e. exposure to speech) of a listener shapes how well a word’s phonetic cues activate meaning by means of association strengths. When a cue is present together with a certain word (representing meaning), their association strength increases, whereas it decreases when the cue is present but the word (meaning) is not.

After determining the association strengths between all cues and meanings in the network of an adult listener (Danks, 2003; using the R-package ‘ndl’), the activation of a meaning for a specific set of cues is calculated by summing the corresponding association strengths. For example, when the listener model is based on native American English (AE) speech, we can compare the activation of a certain meaning for a set of cues on the basis of a native as opposed to a non-native AE pronunciation. Presumably the non-native cues (through lower association strengths) will give rise to a lower activation of the meaning compared to fully native cues. If we then calculate the difference between these activations, we obtain a measure of distance between the two pronunciations of a word.

2 Material

The Speech Accent archive (Weinberger and Kunath, 2011) is digitally available at http://accent.gmu.edu and contains a large sample of speech samples in English from people with various language backgrounds. Each speaker reads the same paragraph of text in English (containing 69 words, of which 55 are unique).
All speech samples are transcribed according to the International Phonetic Alphabet, and the associated audio files are available. In 2010, we extracted all available 989 transcribed samples and their audio from the Speech Accent Archive. In this study, we use a subset consisting of all 115 native U.S.-born English speakers (used as the native reference pronunciations) and 286 mostly non-native speakers for whom we obtained foreignness ratings.

3 Methods

To obtain the NDL-based network of association strengths representing a native AE listener, we randomly selected 58 native AE speakers whose pronunciations were converted to cues (i.e. trigrams of sound segments, including markers representing word boundaries) for the corresponding meanings. As the association strength between a cue and a meaning will obviously depend on the relative frequency with which they co-occur, we extracted word frequency information from the Google N-Gram Corpus (Brants and Franz, 2009). We then constructed the model, yielding a network of association strengths (representing a native AE listener).

Using this network, we first determined the activation of each meaning when supplying the phonetic trigram cues of the remaining 57 native AE speakers. These activations were averaged (across speakers) in order to estimate how well an average native AE speaker is understood by our simulated native AE listener for each meaning separately. In similar fashion, we calculated how well each of the 286 speakers is understood by our simulated native AE listener (for each meaning). To determine the NDL-based pronunciation distance between each individual speaker and the average native AE speaker, we simply calculated the difference between their activations averaged across all meanings.

4 Results

To determine how well these NDL-based pronunciation distances matched perceptual distances we developed a questionnaire in which participants listened to 50 different speech samples and rated their native-likeness (on a scale from 1 to 7). As the questionnaire was advertised in a post on Language Log by Mark Liberman, more than 1100 native AE speakers participated, resulting in at least 50 ratings per speech sample (Cronbach’s alpha: 0.85). The Pearson correlation between the perceptual native-likeness ratings and log-transformed NDL-based pronunciation distances was $r = -0.82$ ($p < 0.001$). These results are comparable to those using the Levenshtein distance (Wieling et al., submitted), which is also illustrated by the high correlation between the two types of computational distances: $r = 0.89$ ($p < 0.001$).

5 Discussion

The high correlation between the perceptual native-likeness ratings and NDL-based pronunciation distances indicates that our new measure indeed captures pronunciation distances. While the Levenshtein distance offers comparable performance and is computationally efficient (it takes about 10 seconds, compared to 50 seconds for the complete NDL procedure), it has no cognitive basis supporting a link with perceptual pronunciation distances, and it does not allow asymmetric distances (such as those reported by Gooskens and Heeringa, 2004). NDL does not suffer from these drawbacks, and as it also allows for the inclusion of non-segmental cues (such as intonation markers), it is a promising alternative to the Levenshtein distance (which does not allow the inclusion of non-segmental information).
References


