A Phonologically Calibrated Acoustic Dissimilarity Measure

Greg Kochanski, Ladan Baghai-Ravary, and John Coleman

University of Oxford Phonetics Lab

One of the most basic comparisons between objects is to ask how similar they are. Linguistics and phonology are founded on this question. The classic definitions of phonemes and features involve contrast between minimal pairs. A minimal pair of words requires that there be two sounds that are dissimilar enough for the words to be considered different. Otherwise we wouldn't speak of a minimal pair of words but rather of a single word with two meanings.

Likewise, phonetic similarity is needed to group together separate instances into a single word or sound. Without some intuition about which sounds are so similar that they should be treated as instances of the same linguistic object, the field would be no more than a collection of trillions of disconnected examples.

So, it is important to have a measure of dissimilarity between sounds. Some exist already: e.g. measures of dissimilarity between the input and output of speech codecs have been used as a way of quantifying their performance (e.g. Gray, Gray, and Masuyama 1980). Cepstral distance has been frequently used, and the Itakura-Saito divergence is also widely used.

But, none of these have been explicitly calibrated against the differences in speech that are important to human language. In this paper, we do so.

As our test data, we collected a corpus of speech where many short phrases were recorded several times each. We paired the phrases and divided them into two classes: where the pairs were recorded from the same text *vs.* pairs from different texts. We then computed distances within all the pairs via a dynamic time-warping algorithm and constructed two histograms: same text and different text. From these histograms, we computed a numerical measure of how much they overlapped each other (it is effectively a t-statistic).

We can then compare different algorithms and/or variations on algorithms. We explored variants on the Itakura-Saito distance with an parametrized pre-filter, and also a Euclidean distance computed on an approximation of the perceptual spectrum. In the latter, we included parameters to individually scale all the components of the vector.

Within each algorithm, we used a simulated annealing algorithm to vary the parameters and find the minimum overlap. We found that the maximization made a dramatic difference in both cases. In both cases, the two histograms started out strongly overlapped, so that distances between different texts were often smaller than distances between utterances recorded from identical texts. However, after the maximization, the histograms were well separated: large distances were reliably associated with different texts, and vice versa.

We showed that an optimization procedure can tune a measurement of acoustic distance so that it corresponds well with the linguistic same-text/different-text dichotomy. We suggest that this technique can be valuable for quantifying similarity and dissimilarity in phonetics and phonology.