Validating and using the PMI-based Levenshtein distance as a measure of foreign accent strength

MARTIJN WIELING
University of Groningen / Tübingen

In this study, we use and validate an adaptation of the Levenshtein distance (Levenshtein, 1965) to determine pronunciation distances between accented speech. The Levenshtein distance measures the minimum number of insertions, deletions and substitutions to transform one (phonetic) string into the other. Our adaptation (Wieling et al., 2009) uses automatically obtained, sensitive linguistically sound (Wieling et al., 2012) segment distances (based on how frequent sound segments correspond in substitutions).

We apply this adapted (i.e. PMI-based) Levenshtein distance algorithm to data from the Speech Accent Archive (Weinberger & Kunath, 2011). The Speech Accent archive contains more than 1000 transcribed speech samples in English from people with various language backgrounds and is digitally available at http://accent.gmu.edu. Each speaker reads the same paragraph of 69 words in English. By calculating the average PMI-based Levenshtein distance between all native American English speakers in the dataset (115 in total) and each individual speaker, we obtain a computational measure of pronunciation distance between a speaker's pronunciation and 'average' American English speech.

To assess how well our automatic method matched human judgments of how native-like a certain accent sounds, we conducted a perception experiment. On the basis of the native-likeness judgments of more than 1100 native American English participants we determined the average native-likeness for 286 samples of native and non-native English speech. As the correlation between the average native-likeness judgments and the (log-transformed) PMI-based Levenshtein distances was r = -0.8, we conclude that our automatic procedure is a valid method to measure foreignness of transcribed speech (Wieling et al., forthcoming).

Given the wealth of data, it is interesting to investigate the accent differences. For example, Figure 1 shows the visualization of the pairwise pronunciation distances between some of the world's English accents.

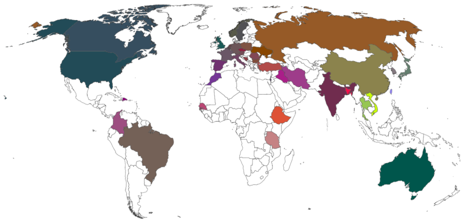

Figure 1: MDS plot of the pairwise pronunciation distances averaged per country. Similar colors indicate similar accents. Only countries were included with at least 5 speakers.