Measuring foreign accent strength using an acoustic distance measure

Martijn Bartelds[1], Caitlin Richter[2], Mark Liberman[2] and Martijn Wieling[1]

[1]University of Groningen, [2]University of Pennsylvania

Introduction

Transcribed speech is often used to study and compare pronunciations (Nerbonne and Heeringa, 1997). Transcribing speech is, however, time consuming and labor intensive (Novotney and Callison-Burch, 2010). The set of discrete symbols used in transcriptions is also unable to capture all the acoustic details relevant for studying accented pronunciations (Cucchiarini, 1996). An acoustic-only method to study pronunciation differences is therefore potentially useful. The approach we will take compares the speech of non-native accented American-English speakers to native American-English speakers using Dynamic Time Warping (DTW). Word-level acoustic differences are computed after applying speaker-based cepstral mean and variance normalization to the feature representations to generalize across speakers. To assess whether the acoustic distance measure is a valid native-likeness measurement technique, we compare the acoustic distances to a collection of human native-likeness judgments collected by Wieling et al. (2014) to evaluate a phonetic transcription-based method.

Data

We use data from the Speech Accent Archive, which contains speech samples from both native and non-native American-English speakers (Weinberger, 2015). Every speaker made a voice recording of the same standard 69-word paragraph. Speech samples from 280 non-native American-English speakers were extracted as our target data set, and 115 reference speech samples (from U.S.-born L1 speakers of English) served as our reference data set. These data are similar to those used in the study of Wieling et al. (2014).

Methods

To include only comparable segments of speech, we automatically time-align the speech samples with a word-level orthographic transcription using the Penn Phonetics Lab Forced Aligner (Yuan and Liberman, 2008). After the forced alignment procedure, we automatically segment each speech sample in the target and reference data set into 69 word-level speech samples. For each segmented speech sample, we compute a numerical feature representation based on Mel-frequency cepstral coefficients (MFCCs). A total of 39 coefficients is computed at each 10 ms step per speech sample, to represent the most important phonetic information embedded within each 25 ms windowed frame. The MFCC feature representation per segmented speech sample is obtained by concatenating its corresponding vectors of 39 coefficients computed for each of the windowed frames. Speaker-based cepstral mean and variance normalization is used to reduce the influence of noise by applying a linear transformation to the coefficients of the MFCC feature representations. Weighting of vowel-identified frames (coefficient ranged between -1 and 2) is explored to enhance relevant phonetic content. The final speaker pronunciation distances are obtained by first calculating the acoustic distance for each of the 69 words pronounced by a non-native speaker of American-English and a single native speaker of American-English in the reference data set. We subsequently average these word-based distances to measure the between-speaker acoustic distance. The difference between the pronunciation of a non-native speaker and native American-English in general, is determined by calculating the between-speaker acoustic distances compared to all 115 native American-English speakers, and subsequently averaging these. We compute these acoustic distances for all foreign-accented speech samples by applying this same procedure to each of the 280 non-native speakers of American-English in the target data set. To evaluate our measure, the correlation between the native-likeness ratings and the acoustic distances is computed. We evaluate the impact of the (size of the) set of reference speakers, by calculating the correlation for successively smaller subsets of reference speakers (75, 50, 25 or 10).

Results

The correlation between the native-likeness ratings and the acoustic distances computed using our acoustic method is $r = -0.71$ ($p < 0.0001$), and therefore accounts for about half of the variance in the native-likeness ratings ($r^2 = 0.50$). As the set of reference speakers might affect the correlation, we evaluated the impact of reducing the set of reference speakers. The correlation remains comparable, irrespective of the (size of the) reference set (i.e., $-0.68 \leq r \leq -0.72$). To assess whether language variation within the set of reference speakers might be important, we computed the acoustic distances using as our reference set ($N = 14$) only the native American-English speakers who originated from the western half of the U.S. and the English-speaking part of Canada. These areas are characterized by less dialect variation compared to the eastern half of the U.S. (Boberg, 2010). Again, this did not substantially affect the correlation, as it remained similar ($r = -0.70$). Weighting the vowel-identified frames did not result in a significantly improved correlation with human perception ($p > 0.05$).

Compared to the transcription-based method of Wieling et al. (2014), who used a modified Levenshtein distance that included automatically determined linguistically-sensible segment distances, and reported a correlation of $r = -0.77$, the performance of our measure is, however, significantly lower (using the modified z-statistic of Steiger (1980): $z = 2.10$, $p < 0.05$).

Discussion

We have created an acoustic-only approach for calculating pronunciation distances between utterances of the same word by different speakers. We have evaluated the measure by calculating how different the speech of non-native speakers of American-English is from native American-English speakers, and by comparing our computed results to human judgments of native-likeness. While our method is somewhat outperformed ($r = -0.71$ vs. $r = -0.77$) by the transcription-based method of Wieling et al. (2014), our measure does not require phonetic transcriptions, whose production is time consuming and prone to errors. Given that our method is fully automatic, the trade-off in performance may be worthwhile.

References

Boberg, C. (2010). The English Language in Canada: Status, History and Comparative Analysis. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511781056.

Cucchiarini, C. (1996). Assessing transcription agreement: methodological aspects. Clin. Linguist. Phonet. 10, 131–155. doi: 10.3109/02699209608985167.

Nerbonne, J., and Heeringa, W. (1997). "Measuring dialect distance phonetically," in *Computational Phonology. Third Meeting of the ACL Special Interest Group in Computational Phonolby* (Stroudsburg, PA: Association for Computational Linguistics (ACL)), 11–18.

Novotney, S., and Callison-Burch, C. (2010). "Cheap, fast and good enough: automatic speech recognition with non-expert transcription," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Los Angeles, CA: Association for Computational Linguistics), 207–215.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. Psychol. Bull. 87:245. doi: 10.1037/0033-2909.87.2.245.

Weinberger, S. (2015). *Speech Accent Archive*. George Mason University. Retrieved from http://accent.gmu.edu.

Wieling, M., Bloem, J., Mignella, K., Timmermeister, M., and Nerbonne, J. (2014). Measuring foreign accent strength in english: validating levenshtein distance as a measure. *Lang. Dyn. Change* 4, 253–269. doi: 10.1163/22105832-00402001.

Yuan, J., and Liberman, M. (2008). Speaker identification on the scotus corpus. *J. Acoust. Soc. Am.* 123:3878. doi: 10.1121/1.2935783.