Measuring distances acoustically

Wilbert Heeringa, Vincent van Heuven, Charlotte Gooskens

April 9th, 2025

Approach

The acoustic distance measure used in LED-A is an adapted version of the measure developed by Bartelds at al. (2020), who developed a distance measure which they used for assessing foreign accent strength in American-English. The speech of non-native American-English speakers was compared to a collection of native American-English speakers. The authors found a strong correlation between the acoustic distances and human judgments of native-likeness provided by more than 1,100 native American-English raters ($r = -0.71$, $p < 0.0001$).

Trim silence

The procedure that we used for comparing the acoustics samples of the realizations of two words is as follows. First, leading and trailing silence is trimmed. This is done by using the Praat function `Sound: To TextGrid (speech activity)…` As to the parameters of this function the default values as provided in Praat are used.

Change gender

In LED-A the user can optionally upload an Excel table that indicates the gender of each speaker. When the table has been uploaded, the male voices will be changed into female voices using the Praat function `Sound: Change gender…` is used. As formant shift ratio the default value of 1.2 is used. As new pitch median we use the average pitch of the word sample multiplied by 2. The default values of the pitch range factor (1.0 = no change) and the duration factor (1.0) are not changed.

Get MFCC representation

Then from the trimmed samples a representation based on Mel-frequency cepstral coefficients (MFCCs) is calculated. A MFCC representation consists of a series of frames where each frame normally includes 12 MFCC coefficients. The 12 parameters are related to the amplitude of the frequencies. MFCCs are popular due to their greater invariance to physical differences between speakers (Davis and Mermelstein 1980).

In order to calculate de MFCC coefficients, the function `melfcc` from the R package tuneR is used. The help text of this function the author writes that the calculation of the MFCCs inlcudes the following steps:

1. Preemphasis filtering
2. Take the absolute value of the STFT (usage of Hamming window)
3. Warp to auditory frequency scale (Mel/Bark)
4. Take the DCT of the log-auditory-spectrum
5. Return the first 'ncep' components

Tweaking the parameters of the MFCC function

Gooskens and Heeringa (2004) validated dialectometric measurements by correlating them with the results of a perception experiment. Recordings of translations of the fable "The North Wind and the Sun" of 15 local dialects in Norway[1] were presented to groups of Norwegian high school pupils in the same locations as where the dialects are spoken. The pupils in each location were familiar with their own dialect and had lived most of their lives in the place in question. The 15 dialects were presented in a randomized order. While listening to the dialect recordings the pupils rated each of the 15 dialects on a scale from 1 (similar to native dialect) to 10 (not similar to native dialect). Since the pupils judged in each of the 15 locations the linguistic distances between their own dialect and the 15 dialects, a 15×15 distance matrix was obtained.

As a basis for the acoustic measurements the 15 recordings of the fable 'The North Wind and the Sun' were split in separate word samples. The Norwegian translations of the fable consists of 58 difference words. If the same word appears more than once in a text, we selected only the first occurrence. For most varieties we got samples for all 58 words. Due to the free translation of some phrases for certain varieties a few of the expected words were missing.[2] The recordings of the varieties of Larvik, Bø, Herøy and Bodø were pronounced by male speakers, the other recordings were pronounced by female speakers.

Using the methodology as presented above we measured the acoustic distances among the 15 varieties and correlated them with the perceptual distances. We experimented with the parameters of the function `melfcc` and kept the settings that gave the highest correlation. The values of the parameters the we found in this ways are given in Table 1. Note that numcep=9 (instead of 12).

[1] The recordings were taken from https://www.hf.ntnu.no/nos/.
[2] For Herøy there are 56 samples, for Lesja 57 samples, for Stjørdal 56 samples, for Trondheim 57 samples and for Verdal 57 samples.

parameter	value	description
samples		Object of Wave-class or WaveMC-class. Only the first channel will be used.
sr	Sample rate of wav file	Sampling rate of the signal.
wintime	0.025	Window length in sec.
hoptime	0.01	Step between successive windows in sec.
numcep	**9**	Number of cepstra to return.
lifterexp	0.6	Exponent for liftering; 0 = none.
htklifter	FALSE	Use HTK sin lifter.
sumpower	TRUE	Ifsumpower = TRUE the frequency scale transformation is based on the powerspectrum, if sumpower = FALSE it is based on its squareroot (absolute value of the spectrum) and squared afterwards.
preemph	0.97	Apply pre-emphasis filter [1 -preemph] (0 = none).
dither	FALSE	Add offset to spectrum as if dither noise.
minfreq	**50**	Lowest band edge of mel filters (Hz).
maxfreq	**5000**	Highest band edge of mel filters (Hz).
nbands	**32**	Number of warped spectral bands to use.
bwidth	1	Width of spectral bands in Bark/Mel.
dcttype	t2	Type of DCT used - 1 or 2 (or 3 for HTK or 4 for feacalc).
fbtype	mel	Auditory frequency scale to use: "mel", "bark", "htkmel", "fcmel".
usecmp	FALSE	Apply equal-loudness weighting and cube-root compression (PLP instead of LPC).
modelorder	NULL	If modelorder > 0, fit a linear prediction (autoregressive-) model of this order and calculation of cepstra out of lpcas.

Table 1. Parameters used in the R function melfcc. The values highlighted in red are different from the default values.

Standardize MFCC coefficients

The quality of the MFCC feature representation is highly influenced by the presence of noise in the speech samples (Ganapathy et al. 2011, Shafik et al. 2009). The effect of noise can be reduced by standardizing the MFCC coefficients. Individually for each of the MFCC coefficients the mean and standard deviation are calculated over those coefficients in the course of the time. Subsequently, the mean is removed from the coefficients, and the resulting values are divided by the standard deviation. This standardization procedure is applied to each word sample individually.

Dynamic time warping

The acoustic word distance between the pronunciation of the user and the pronunciation of the reference speaker is computed using the dynamic time warping (DTW) algorithm (Galbally & Galbally 2015). The frames of the two respective MFCC representations are aligned to each other. Every frame in the one representation must be matched with one or more frames from the other representation, and vice versa. In order to find a logical match of the frames in the one representation with the frames in the other representation, frames are compared to each other. Bartelds at al. (2020) use the Euclidean distance. We calculate 1 minus Pearson's correlation as distance between two frames, which gives easy to interpret distances between 0 and 1, while we found it functioning well. The DTW algorithm matches the frames so that the overall distance between the two sequences of frames is minimized.

Normalize DTW distance

Bartelds at al. (2020) normalizes the DTW distance by dividing it by the sum of the lengths of the two representations. Instead we normalize by dividing by the length of the alignment, which we judge to be more precise. Since the frame distance varies between 0 and 1, the normalized distance will vary between 0 and 1 as well.

References

Bartelds, Martijn & Richter, Caitlin & Liberman, Mark & Wieling, Martijn (2020). A New Acoustic-Based Pronunciation Distance Measure, *Frontiers in Artificial Intelligence* 3, doi: 10.3389/frai.2020.00039.

Davis, S. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust. Speech Signal Process* 28, 357–366, doi: 10.1109/TASSP.1980.1163420.

Galbally, Javier & Galbally, David (2015). A pattern recognition approach based on DTW for automatic transient identification in nuclear power plants, Annals of Nuclear Energy 81, 287–300, doi: 10.1016/j.anucene.2015.03.003.

Ganapathy, S. & Pelecanos, J. & Omar, M.K. (2011). Feature normalization for speaker verification in room reverberation, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Prague)*, 4836–4839, doi: 10.1109/ICASSP.2011.5947438.

Gooskens, Charlotte & Heeringa, Wilbert (2004). Perceptive Evaluation of Levenshtein Dialect Distance Measurements using Norwegian Dialect Data, *Language Variation and Change*, 16(3), 189-207.

Heeringa, Wilbert & Gooskens, Charlotte (2003). Norwegian Dialects Examined Perceptually and Acoustically. *Computers and the Humanities*, 37(3), 293-315.

Heeringa, Wilbert & Van Heuven, Vincent & Van de Velde, Hans (2022). *LED-A: Levenshtein Edit Distance App* [Computer program]. Retrieved 2 January 2023 from https://www.led-a.org.

Heeringa, Wilbert & Johnson, Keith & Gooskens, Charlotte (2009). Measuring Norwegian Dialect Distances using Acoustic Features. *Speech Communication* 51(2), 167-183.

Shafik, A. & Elhalafawy, S.M. & Diab, S. & Sallam, B.M. & El-Samie, F.A. (2009). A wavelet based approach for speaker identification from degraded speech, *International Journal of Communication Networks and Information Security* 1(3).