Deep acoustic representations for clustering Dutch dialect pronunciations

Martijn Bartelds and Martijn Wieling

University of Groningen
The Netherlands

{m.bartelds, m.b.wieling}@rug.nl

Generally, dialectometric methods have focused on computing the string edit distance to compare phonetic transcriptions of speech samples (e.g., Nerbonne et al., 1999). The process of manually transcribing speech samples is, however, time-consuming and labor-intensive (Hakkani-Tür et al., 2002; Novotney and Callison-Burch, 2010). Another limitation of this procedure is that transcriptions can not fully capture all acoustic details of human speech, since often a limited set of transcription symbols is used (Liberman, 2018).

Alternatively, deep acoustic models automatically learn linguistic information on the basis of large audio corpora by taking the complete audio signal into account. While these acoustic models are primarily developed for automatic speech recognition (e.g., `wav2vec 2.0`; Baevski et al., 2020, `XLSR-53`; Conneau et al., 2020), they might learn information that is useful for other tasks. In this study, we therefore investigate whether we can use deep acoustic models to develop an acoustic distance measure for investigating differences between Dutch dialect pronunciations. Specifically, we use a deep acoustic `wav2vec 2.0` model pre-trained and fine-tuned on Dutch, and a multilingual `XLSR-53` model fine-tuned on Dutch. In addition, we compare the deep acoustic models to the adjusted Levenshtein distance algorithm of Wieling et al. (2012).

We extract data from the Goeman-Taeldeman-Van Reenen-Project (Goeman and Taeldeman, 1996), which contains audio recordings and phonetic transcriptions of hundreds of words for 613 dialect varieties in the Netherlands and Flanders. The metadata with time stamps to segment the recordings into words was only available for a small subset of the data. We therefore use recordings and phonetic transcriptions for 106 Netherlandic dialect varieties for which we have dialect pronunciations recordings of the same 10 words.

We extract acoustic representations from the deep acoustic models, and compare representations of the same word for every pair of locations using dynamic time warping (Senin, 2008). Similarly, the adjusted Levenshtein distance algorithm is used to compare the phonetic transcriptions. We average word-based distances between two locations to obtain a single pronunciation distance score, and do this for each pair of locations. The resulting distance matrices are subsequently clustered[1] into four groups and compared (using the spatially-sensitive `CDistance` score of Coen et al., 2010) to a gold standard clustering, distinguishing the three officially recognized regional (minority) languages spoken in the Netherlands (i.e. Frisian, Low Saxon, and Limburgish) and Dutch.

[1]While there are many clustering algorithms, we only included the ones available in *Gabmap* (Nerbonne et al., 2011), and for each approach we selected the clustering algorithm resulting in the highest cophenetic correlation coefficient (Sokal and Rohlf, 1962).

Using this approach, we obtain `CDistance` scores (lower is better) of 0.34 (Dutch `wav2vec 2.0` using WPGMA (Weighted Pair Group Method using Arithmetic averages) clustering), 0.20 (fine-tuned `XLSR-53` using complete link clustering), and 0.46 (adjusted Levenshtein distance using UPGMA (Unweighted Pair Group Method using Arithmetic averages) clustering). Consequently, we find that the fine-tuned `XLSR-53` model can be most effectively used to distinguish between language varieties in the Netherlands.

Combined with earlier work of Bartelds et al. (2021), which showed that these deep acoustic models were also superior in distinguishing accented speech, our results suggest that our approach is a suitable alternative to dialectometric analysis requiring (time-consuming) phonetic transcriptions. Importantly, our analysis appears to be effective even when only few audio samples are available.

References

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Bartelds, M., de Vries, W., Sanal, F., Richter, C., Liberman, M., and Wieling, M. (2021). Neural Representations for Modeling Variation in Speech. *arXiv preprint arXiv:2011.12649*.

Coen, M. H., Ansari, M. H., and Fillmore, N. (2010). Comparing Clusterings in Space. In *Proc. of ICML*.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2020). Unsupervised Cross-lingual Representation Learning for Speech Recognition. *arXiv preprint arXiv:2006.13979*.

Goeman, A. and Taeldeman, J. (1996). Fonologie en morfologie van de Nederlandse dialecten; een nieuwe materiaalverzameling en twee nieuwe atlasprojecten. *Taal en Tongval*, 48:38–59.

Hakkani-Tür, D., Riccardi, G., and Gorin, A. (2002). Active learning for automatic speech recognition. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3904. IEEE.

Liberman, M. (2018). Towards progress in theories of language sound structure. In Brentari, D. and Lee, J. L., editors, *Shaping phonology*. University of Chicago Press.

Nerbonne, J., Heeringa, W., and Kleiweg, P. (1999). Edit distance and dialect proximity. *Time Warps, String Edits and Macromolecules: The theory and practice of sequence comparison*, 15.

Nerbonne, J., Rinke, C., Charlotte, G., Peter, K., and Therese, L. (2011). Gabmap - a web application for dialectology. *Dialectologia: revista electrònica*, pages 65–89.

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*, pages 207–215. Association for Computational Linguistics.

Senin, P. (2008). Dynamic Time Warping Algorithm Review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, 855(1-23):40.

Sokal, R. R. and Rohlf, F. J. (1962). The Comparison of Dendrograms by Objective Methods. *Taxon*, 11(2):33–40.

Wieling, M., Margaretha, E., and Nerbonne, J. (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, 40(2):307–314.