TOWARDS A SPEAKER LOCALIZATION FROM SPONTANEOUS SPEECH: NORTH-SOUTH CLASSIFICATION FOR SPEAKERS OF CONTEMPORARY GERMAN

Geographical regression analysis based on phonetic features aims to locate the origin of an unknown speaker by relating phonetic features derived from a small speech sample to longitude/latitude coordinates. In this paper we present results from a preliminary experiment in which we classify only North/South distinction based on openSMILE ([1]) features derived from the German "German Today" ([2]) corpus using Random Forests. The aim of the present study is to test the feasibility of a data-driven approach to geolocalization with Random Forests, to evaluate the German phone classes that carry geoinformation, and to evaluate which phonetic features contribute to such a binary classification of speakers belonging to the North or South. It is worth noting, that those North/South labels are based on the geographic centroid of the recording sites and not a dialectology-driven North/South boundary. This is important as the method should be a bottom-up data-driven approach in localization, where no further linguistic or meta-information is provided.

The corpus was recorded in locations distributed over Germany, Austria, Switzerland, in small parts of South Tyrol (Italy) and Luxembourg and is the biggest corpus of contemporary German that is currently available. In each of the 165 locations up to two male and two female students were recorded. The subjects needed to be aged between 16 and 20, to be born and raised in the area of the recording, and have at least one parent that grew up in the recording region as well. The data used consists of 640 speakers (328 female, 312 male). The grid over the corpus area is not equally spread, but [2] points out that no important classically defined dialect area was left out. The average distance of the recording sites to their closest neighbor is 41.12 *km*, the two closest ones are 16.76 *km* apart, the biggest gap is 72.91 *km* and distances have a standard deviation of 11.48 *km*.

Random Forests were used for the classification as the literature suggests, that, despite the free lunch theorem, Random Forests outperform other methods in many settings [3]. A big advantage of Random Forests are the few, insensitive hyper-parameters to tune (e.g. for number of trees in the tree). Due to its nature, Random Forests are easily and efficiently parallelizable, which makes them a good choice in case many classifiers have to be learned (like in the present study a Random Forest for each of the 43 phonemes is necessary).

It turns out that with the voiced fricative /z/ alone it is possible to correctly classify 81.72% of the speakers, which confirms the often reported North/South voicing contrast in Germany of /z/ in many positions (e.g. intervocalic [4]). Therefore it is also not surprising that features associated with voicing, e.g. spectral roll off, voicing probability, zero and mean crossing rate are within the top features. A fusion of all phone classifications yields a slightly better accuracy of 85.16%. Surprisingly, all phone class were found to contribute to the classification which is very promising for the more difficult regression problem. However, we identify a number of features that do not contribute to the geolocalization and will therefore be discarded in future experiments.

References

[1] EYBEN, F., M. WÖLLMER, and B. SCHULLER: *openSMILE: the munich versatile and fast open-source audio feature extractor*. In *Proceedings of the International Conference on Multimedia*, MM '10, pp. 1459–1462. ACM, New York, NY, USA, 2010.

[2] BRINCKMANN, C., S. KLEINER, R. KNÖBL, and N. BEREND: *German today: an areally extensive corpus of spoken standard german.* In *Proceedings 6th International Conference on Language Resources and Evaluation (LREC). Marrakesch, Marokko*. 2008.

[3] FERNÁNDEZ-DELGADO, M., E. CERNADAS, S. BARRO, and D. AMORIM: *Do we need hundreds of classifiers to solve real world classification problems?* *Journal of Machine Learning Research*, 15(1), pp. 3133–3181, 2014.

[4] KÖNIG, W.: *Atlas zur Aussprache des Schriftdeutschen in der Bundesrepublik Deutschland: Text*, vol. 1 (Text). M. Hueber Verlag, 1989.