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

Thomas Kisler, Florian Schiel

Bavarian Archive for Speech Signals,
Institute of Phonetics and Speech Processing,
Ludwig Maximilian University Munich, Germany
[kisler\schiel]@bas.uni-muenchen.de

Abstract: Geographical regression analysis based on phonetic features aims to locate the origin of a 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 using Random Forests, we classify speakers into two geographical regions North/South based on openSMILE [1] features derived from the German "German Today" [2] corpus. The aims are 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 from the North or South. Results show that based on 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 most positions [3]. We identify a number of features that do not contribute to the geolocalization and will therefore be discarded in future experiments.

1 Introduction

It is commonly assumed that linguistic features of speakers vary according to where they spent their childhood and where they reside. In dialectology it has been shown that variation of certain linguistic features correlates with geographic distribution. The border of two variants can be visualized as a line on a map, indicating the position where the change occurs (isogloss). The local accumulation of isoglosses allows dialect areas to be defined. However, dialect areas are not as homogeneous as dialect maps suggest, since the exact geographical position of a dialect boundary is open for discussion [4]. For instance, moving from one point in Germany to another, one would expect to find places where sudden changes in a single linguistic feature occur, but that the overall dialectal variation (bundles of features) would change semi-continuously along this path, as single differences accumulate. In this study we assume that acoustic features show a regional distribution like other linguistic features, and exploit this information to determine the origin of speakers.

We concentrate on acoustic features, more precisely, features that can be calculated from the speech signal and an automatically calculated alignment to phone classes, without any additional linguistic or meta information. The reason that we constrain our paradigm to this sort of feature is to allow any future applications of our method to work fully automatically with only the speech signal as input. Within this paradigm we address three major questions:

1. Is it possible to predict the geographic coordinates (longitude and latitude) of the speaker's origin from a short speech sample, and what is the average geographical accuracy?

2. Is it possible to automatically cluster a large geographic area into regions in which speakers display similar phonetic behavior based on the speech signal sampled from a large and geographically distributed number of speakers? Do the resulting clusters resemble traditional dialect areas, or – if not – why do they deviate? Do these data-driven clusters out-perform traditional dialect areas in dialect classification?

3. If either questions 1 or 2 can be successfully answered, can this sort of data-driven geographic referencing be utilized to improve automatic speech recognition?

In this paper we concentrate only on the first question. To test the feasibility of this idea we conducted an experiment to predict whether a test speaker originates from the northern or southern part of the corpus' geographic space. The results of the experiment have the potential to illuminate several issues: First, if they suggest that such a (relatively) simple two-way prediction is not possible, then the much more challenging regression/cluster analysis into geographic coordinates is unlikely to work at all. Secondly, if the prediction does work, it would then be possible to investigate which phonetic features are contributing the most information for this prediction, and which phonetic sound classes carry these features. If for instance it turns out that voiceless fricatives do not contribute to this prediction at all, they might be omitted in future experiments, thus reducing modeling effort and noise in the predictor system.

Geographic regression analysis based on phonetic features alone has not been investigated before. A related area of research is dialect classification of a speech sample, where the geographic distribution of features is of less interest than the correct assignment of a dialect label to an apriori defined geographic area. The prediction of dialect classes was tested on the German RVG1 corpus by Stadtschnitzer et al. [5]. The authors attempted to predict dialect membership of test speakers to a fixed set of 9 large German dialect areas using phonetic features alone without regard to the phone class. They used the speaker identification tool kit ALIZE [6] (by labelling various speakers from one dialect group as one speaker), and found that prediction accuracy was just about chance. It is interesting to note that the authors also tested a different approach based on a phonemic 4-gram model, which yielded results well above chance on the same data set. This suggests that 'higher' linguistic features (such as phonological, lexical) out-perform phonetic features with regard to dialect classification.

Regarding languages other than German, several studies of dialect/accent classification have been conducted based on read [7, 8, 9, 10, 11] and spontaneous speech [10, 11, 12]. For the sake of brevity, we concentrate on studies that included spontaneous speech as investigated in the present study.

Brown [10] applied a variant of the Accent Characterisation by Comparison of Distances in the Inter-segment Similarity Table (ACCDIST) system on read and spontaneous speech in a 4-way dialect discrimination task on speech from the English/Scottish border, and reported a drop of about one third in classification accuracy on spontaneous (52.5%) compared to read speech (86.7%), which indicates that accent recognition for spontaneous data is more challenging. A slight increase in accuracy was found when the phoneme context was discarded, probably because the number of observations for each class increased. The features used in the system were the first 12 Mel-scale cepstral coefficients (MFCCs) extracted at the vowel midpoint.

Woehrling et al. [11] conducted a dialect identification experiment on French dialect regions. The study used both read speech (3 minutes) as well as spontaneous speech (10-15 minutes). Aside from phonetic features such as formant frequencies and voicing, other linguistic features were analyzed such as pronunciation variants (derived from an automatic phoneme alignment) as well as several prosodic features mainly derived from duration measurements and fundamental frequency contours. The best classification rate of 85% was reported for a speaker classification into three major dialect regions using Support Vector Machines (SVMs) (82% on

5 classes). It is interesting to note that although SVMs yielded the best results, the authors favored a decision tree as classificator since these allow better insight into which features are most useful for the classification problem.

Biadsy et al. [12] implemented a dialect classifier on 4 different variants of Arabic. Using a sophisticated SVM technique, exploiting the phone context of inspected data, the authors reported an average equal error rate of 4.9% in binary classifiers, i.e. classifiers that discriminate between two of the four variants. It is noteworthy here that (in contrast to [10]) the authors stress the importance of the phone context: speech signals firstly undergo a phone recognition stage that allows the later pairing of data from the same phone type in the discrimination task.

The outline of the remaining paper is as follows: the next section describes the data set derived from the German corpus of spoken language "German Today" provided by the Institut für Deutsche Sprache, Mannheim, Germany. Section 3 describes the methodology of the experiment, followed by a discussion of the results together with their implications regarding the discriminative power of certain phonetic features and the phonetic classes that contribute most to the North/South classification.

2 Data

2.1 Corpus and phonetic segmentation

Training and test data were taken from the German speech corpus "German today" [2]. The corpus was recorded in locations distributed over Germany, Austria, Switzerland, and a few sites located in South Tyrol (Italy) and Luxembourg (denoted as "corpus area" in the following). In each location, up to two male and two female students of the local high school (Gymnasium) were recorded, aged between 16 and 20, and born and raised in the area; at least one parent had to be from the region as well. Due to technical issues we used a subset of speakers in this study, which consists of 640 speakers (328 female, 312 male) from 165 locations. Speakers performed a map task in pairs [2] resulting in semi-spontaneous speech, where certain words (e.g. objects on the map, measurements) occur more frequently than others.

Recordings were orthographically transcribed by human annotators. WebMAUS [13] was applied to segment and label recordings into word and phoneme segments. The Munich AUtomatic Segmentation (MAUS) was applied in forced-alignment mode to prevent it from changing the canonical transcript to model speech variants based on the signal for two reasons: a) MAUS has not been trained on dialect data, and therefore might not be able to model dialectal variations over the whole corpus area equally well, and b) we aim to compare dialect differences at the signal level, not at the phonological level.

The result of this procedure was a collection of speaker/location labeled phoneme segments of 43 different German phonemes [1]

2.2 Acoustic features

In dialect classification, among other linguistic features, several standard acoustic features have been employed frequently, such as MFCCs [10, 15, 16, 17, 18], also in combination with the closely related Shifted-Delta Cepstral (SDC) [19, 20]. Other acoustic features include signal energy [16, 21, 22], formants and auditory filterbanks [9], duration [21] and fundamental frequency [22].

Since we were interested in investigating which features carry information regarding the origin of a speaker, we started off with a large set of features including those, that might be

[1] A subset of the German phoneme set of WebMAUS that ultimately is based on [14].

redundant and/or irrelevant. This approach is inspired by the large feature sets used for paralinguistic challenges, like Schuller et al. [23].

Given that Brown [10] reported that adding phoneme context slightly decreased classification performance, we did not use any context-dependent features. The following features were extracted every 10*ms* from a window of 20*ms* width, and then averaged over the 20% midpoint centered region of each phonetic segment applying the openSMILE software package [1]: Phoneme duration, RMS and log energy, raw and smoothed fundamental frequency (F0), MFCCs, Zero Crossing Rate (ZCR), Mean Crossing Rate (MCR), Voicing Probabilities (VPs), Harmonics-to-Noise Ratio (HNR), chroma features, jitter and shimmer, Perceptual Linear Predictive (PLP) and RASTA-PLP coefficients, Linear Predictive Coefficients (LPC), Line Spectral Pairs (LSP), auditory spectra, spectral features (arbitrary band energies, roll-off points, centroid, entropy, maxpos, minpos, variance (=spread), skewness, kurtosis, slope), formant frequencies and bandwidths, Semi-tone Spectra (STS). Additionally, we calculated velocity (Δ) and acceleration ($\Delta\Delta$) of each feature, which resulted in a total set of 737 features.

3 Method

3.1 Division of speakers

As stated above, we examined a binary classification of the corpus area into North and South. The area was split at the midpoint which is defined as the mean of positions of all 640 speakers in the corpus (50.01903E, 10.41484N). All speakers originating below or at the same altitude as the midpoint were put into the "South" class and the remainder in the "North" class. This division resulted in 334 speakers in the "South" class (159 male, 175 female) and 306 speakers in the "North" class (153 male, 153 female).

3.2 Random Forests

As a predictor we chose Random Forests (RFs), a non-linear classification technique originally proposed by Breiman [24]. A comparative study on real world classification problems revealed that RFs were in many cases the best classifier, even outperforming the more complex and slower SVM [25]. RFs (like SVMs and decision trees) support both classification and regression problems with minor changes in the splitting criterion.

Furthermore, RF are very efficiently parallelizable and are reported to be insensitive to their hyperparameters [24, 26]). Nevertheless, we evaluated the two most important hyperparameters: *number of trees* and *mtry*, the number of features considered at each split. We generated fully grown trees and used the default value of 1 for classification for minimal node size. The R package "Random Forest Generator" (ranger) was used for RF training [27].

3.3 Data Partition

As outlined in Section 1, one of the aims of this experiment was to determine which phonemes can be successfully used to distinguish between North/South. To address this aim, we trained a RF for each phoneme independently and compared their predictive power by calculating the accuracy, precision and recall for each phoneme. We applied a standard Leave-25%-speaker-out cross-validation. The four speakers recorded in each corpus location were randomly assigned to four location and gender balanced speaker groups. Since four speakers did not exist for all locations, this resulted in slightly unbalanced sets.

3.4 Evaluation of features

The Variable Importance (VI) is a measure that evaluates the contribution of a feature to the classification/regression problem [24]. The VI is based on the Gini index [26], which is used to split nodes during the growth of the decision trees, and the resulting decrease of impurity is used to estimate the importance of a variable [27]. The VI has certain problems, e.g. features ranked high in VI "mask" correlated features. Moreover, the Gini index, on which it is based prefers features with many outcomes over ones with few [28]. Nevertheless, the VI has been used to evaluate the importance of features, e.g. in [26, 29]. In our case, we used the VI to select those features that did not contribute to the prediction model at all.

4 Results

4.1 RF parametrization

As indicated in Sec. 3.2 we first tested the influence of the two main RF parameters: we varied *mtry* between $\sqrt{p} = 27$, 100 and $p/3 = 245$[2] with a fixed number of 100 trees, and we varied the *number of trees* between 100, 150 and 250 with a fixed $mtry = \sqrt{p}$ (27). In each case we calculated the average accuracy of the 5 highest ranking phonemes. For the three different *mtry* values $\sqrt{p}$, 100 and $p/3$ (with ntree = 100) the accuracies are 0.7648, 0.7598 and 0.7638 respectively. For the three values for number of trees 100, 150 and 250 (with $mtry = \sqrt{p}$) the accuracies were 0.7648, 0.7607 and 0.7619 respectively. Based on these values, it seems that the RFs classification was not very sensitive to these parameters, which agrees with previous findings (e.g. [24, 26]). Since the lowest values in both parameters lead to slightly better results and are faster to train, we applied only these in the following experiments.

Table 1 – Classification accuracy, precision and recall for the 5 top ranking phonemes; ordered by accuracy and rounded to 4 decimals.

Phoneme	Accuracy	Precision	Recall
/z/	0.8172	0.8703	0.7635
/t/	0.7516	0.7676	0.7515
/b/	0.7500	0.7719	0.7395
/øː/	0.7412	0.7979	0.6818
/ɐ/	0.7391	0.7147	0.8323

4.2 Binary classification

The performance metrics are accuracy, precision and recall. They are defined as:

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n}, \quad Precision = \frac{t_p}{t_p + f_p}, \quad Recall = \frac{t_p}{t_p + f_n}$$

where t_p denotes the true positives, t_n the true negatives and f_p and f_n the false positives and false negatives respectively.

[2] $\sqrt{p}$ and $p/3$ are the default values for classification and regression tasks respectively of the original R programming environment [30] implementation of Breiman's algorithm [31].

Table 1 shows the results for the 5 best phonemes ranked according to accuracy. It is worth mentioning that all phonemes (including the phonemes not shown in Table 1) predict the correct geographic class above chance; the worst accuracy of 0.6397 is achieved with the phoneme /ʏ/.

Table 2 – The top 10 features for the best performing phonemes /z/ and /t/ ranked by VI (column 1 and 2), and ranked by their mean over all phonemes (cf. 3.4). Abbreviations, if not mentioned before, are Auditory Spectrum (AS), Rasta Filtering (RF), Voicing Candidate (VC), Voicing Final Unclipped (VU), Spectral Roll Off 25% (Spec. RO). If a feature consists of a vector, its feature index is given in parentheses.

/z/	**/t/**	**Mean**
VU	LPC (7)	AS (1)
VC (0)	AS RF (1)	STS (20)
ZCR	LSP (4)	MFCC (3)
AS (2)	AS RF (19)	AS (0)
MFCC (8)	AS RF (20)	MFCC (8)
AS (13)	AS RF (18)	MFCC (5)
AS (14)	MFCCΔΔ (1)	STS (8)
AS (1)	LPC (0)	STS (17)
Spec. RO	AS RF (21)	AS RF (1)
MCR	AS RF (16)	MFCC (6)

It is not surprising that the phoneme /z/ shows the best accuracy, as it is often reported that southern varieties of German devoice standard German /z/ in most positions [3]. It is also not surprising that features associated with voicing, e.g. spectral roll off, voicing probability and measures of the periodicity of the signal, like ZCR and MCR, are within the top features (c.f. Table 2, left column).

4.3 Noise features

The following *81 features* showed a VI of 0.0 in all phonemes (feature index in brackets):

- Linear Predictive Coefficients (LPC) (3,4,5,6,7) (*5 features*)
- Semi-tone Spectra (STS) (0-7,9-11,13,14,16,18,19,21,23,26,30,94,95) (*22 features*) including the respective Δ and ΔΔ features (*54 features*).

5 Discussion and future work

We have shown that all German phoneme types available in the corpus can be used to classify North/South above chance. These results suggest that even a small subset of phoneme types could be sufficient for regression/cluster analysis, which might enable applications such as automatic speech recognition to estimate speaker origin from the first few spoken words. Classification schemes based on linguistic features such as dialectal word forms or phonemic n-grams (cf. [5]) are expected to require much more input data from the target speaker.

A total of 81 ($\approx 11\%$ of the set of 737) features were found in the openSMILE feature set, that do not contribute to the classification task at all. In further tests these can be omitted to save processing time.

In future experiments, we will examine the East/West classification and the possibility of estimating the speaker origin in a two-step-regression. Pre-tests indicate that the East/West classification is harder to achieve than the North/South distinction.

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

[4] WIESINGER, P.: *Die Einteilung der deutschen Dialekte*. *Dialektologie*, 2. Halbband, pp. 807–900, 1983.

[5] STADTSCHNITZER, M., C. SCHMIDT, and D. STEIN: *Towards a localised german automatic speech recognition*. In *Speech Communication; 11. ITG Symposium*, pp. 1–3. 2014.

[6] LARCHER, A., J.-F. BONASTRE, B. G. FAUVE, K.-A. LEE, C. LÉVY, H. LI, J. S. MASON, and J.-Y. PARFAIT: *Alize 3.0-open source toolkit for state-of-the-art speaker recognition*. In *Interspeech*, pp. 2768–2772. 2013.

[7] HANANI, A. and M. J. RUSSEL: *Human and computer recognition of regional accents and ethnic groups from british english speech*. *Computer Speech and Language*, (27), pp. 59–74, 2012.

[8] NAJAFIAN, M., S. SAFAVI, P. WEBER, and M. RUSSELL: *Identification of british english regional accents using fusion of i-vector and multi-accent phonotactic systems*. In *Proceedings of Odyssey*. 2016.

[9] HUCKVALE, M.: *ACCDIST: a metric for comparing speakers' accents*. In *Proceedings of Interspeech*, pp. 29–32. 2004.

[10] BROWN, G.: *Automatic recognition of geographically-proximate accents using content-controlled and content-mismatched data*. In T. S. C. FOR ICPHS 2015 (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK, 2015.

[11] WOEHRLING, C., P. B. DE MAREÜIL, and M. ADDA-DECKER: *Linguistically-motivated automatic classification of regional french varieties.* In *Proceedings of Interspeech*, pp. 2183–2186. 2009.

[12] BIADSY, F., J. HIRSCHBERG, and M. COLLINS: *Dialect recognition using phone-gmm-supervector-based SVM kernel*. In *Proceedings of Interspeech*. 2010.

[13] KISLER, T., U. REICHEL, and F. SCHIEL: *Multilingual processing of speech via web services*. *Computer Speech & Language*, 2017.

[14] WELLS, J. C. ET AL.: *Sampa computer readable phonetic alphabet*. *Handbook of standards and resources for spoken language systems*, 4, 1997.

[15] DAVIS, S. and P. MERMELSTEIN: *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4), pp. 357–366, 1980.

[16] ARSLAN, L. M. and J. H. L. HANSEN: *Language accent classification in american english*. *Speech Commun.*, 18(4), pp. 353–367, 1996.

[17] HANSEN, J. H. L., U. H. YAPANEL, R. HUANG, and A. IKENO: *Dialect analysis and modeling for automatic classification*. In *Proceedings of Interspeech*. 2004.

[18] PEDERSEN, C. and J. DIEDERICH: *Accent classification using support vector machines*. In *Computer and Information Science, 2007. ICIS 2007. 6th IEEE/ACIS International Conference on*, pp. 444–449. IEEE, 2007.

[19] HANANI, A., M. RUSSELL, and M. CAREY: *Human and computer recognition of regional accents and ethnic groups from british english speech*. *Computer Speech & Language*, 27(1), pp. 59–74, 2013.

[20] DEMARCO, A. and S. COX: *Iterative classification of regional british accents via i-vector space*. In *Symposium on machine learning in speech and language processing*. 2012.

[21] SINHA, S., A. JAIN, and S. AGRAWAL: *Acoustic-phonetic feature based dialect identification in hindi speech*. *International Journal on Smart Sensing & Intelligent Systems*, 8(1), 2015.

[22] HILLENBRAND, J. and R. T. GAYVERT: *Vowel classification based on fundamental frequency and formant frequencies*. *Journal of Speech, Language, and Hearing Research*, 36(4), pp. 694–700, 1993.

[23] SCHULLER, B., S. STEIDL, A. BATLINER, F. BURKHARDT, L. DEVILLERS, C. MÜLLER, and S. NARAYANAN: *Paralinguistics in speech and language state-of-the-art and the challenge*. *Computer Speech & Language*, 27(1), pp. 4–39, 2013.

[24] BREIMAN, L.: *Random forests*. *Machine learning*, 45(1), pp. 5–32, 2001.

[25] 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.

[26] ARCHER, K. J. and R. V. KIMES: *Empirical characterization of random forest variable importance measures*. *Computational Statistics & Data Analysis*, 52(4), pp. 2249–2260, 2008.

[27] WRIGHT, M. N. and A. ZIEGLER: *ranger: A fast implementation of random forests for high dimensional data in C++ and R*. *arXiv preprint arXiv:1508.04409*, 2015.

[28] LOUPPE, G.: *Understanding random forests: From theory to practice*. Ph.D. thesis, 2014.

[29] LIAW, A. and M. WIENER: *Classification and regression by randomForest*. *R news*, 2(3), pp. 18–22, 2002.

[30] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[31] LIAW, A. and M. WIENER: *Classification and regression by randomforest*. *R News*, 2(3), pp. 18–22, 2002.