SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education
20-21 September 2019, Graz, Austria

Implementing and Evaluating Methods of Dialect Classification on Read and Spontaneous German Speech

Johanna Dobbriner[1], *Oliver Jokisch*[2]

[1]Leipzig University, Institute of Applied Informatics, 04107 Leipzig, Germany
[2]Leipzig University of Telecommunications (HfTL), Institute of Communications Engineering, 04277 Leipzig, Germany

johanna@bioinf.uni-leipzig.de, jokisch@hft-leipzig.de

Abstract

The majority of pronunciation tutoring systems is addressing foreign learners of a language, while some applications, such as the rhetoric or articulation training of actors and managers, are also dedicated to native speakers. In particular, accent-reduction tutoring requires reliable methods of Automatic Dialect Classification (ADC) on both, learners' or reference speech. Beyond that, ADC can support further applications of language and speech technology, e.g. the localization of call-center talks or a forensic analysis. Our contribution describes ADC experiments on different corpora of read and spontaneous German speech, which are not based on prior corpus transcriptions. We started with selected feature combinations and classification methods, which have been already studied on English, Mandarin or Arabic. Based on this, we implemented, trained and evaluated classifiers on German dialect varieties in the corpora "Regional Variants of German 1" (RVG) and "Deutsch Heute" (DH). Our test design is focused on differences among the read and spontaneous speech data. The evaluation indicates, that a three or nine-class dialect discrimination and classification on read and spontaneous speech are utilizing the same basic principles, although the overall results, purely using spectral or correlating features, are less sophisticated and call for further clarifications.

Index Terms: dialect classification, read and spontaneous speech, spectral features, GMM

1. Motivation

Automatic Dialect Classification (ADC) combines principles from both language and speaker identification in order to automatically recognize a regional dialect of a given language from speech samples and in some cases from corresponding transcripts. The differences between dialects are usually less distinctive than between languages in general, which often challenges a discrimination of dialects, even for human listeners. Beyond the small variations, there are usually no fixed borders between dialect realizations, neither phonetic nor prosodic, as people, who live in different places over their lifetime, tend to exhibit a mixture of different dialects.

For many applications, ADC can be useful despite the impediments mentioned. Automatic Speech Recognition (ASR) constitutes a typical example, since the variation within a language can be challenging for a recognizer engine. Another application scenario of ADC is language tutoring for native speakers with the goal to reduce their regional accent. Accent reduction training is especially important for e.g. actors and people like newsreaders, who appear in the media frequently. An automated training for accent reduction would have to rely on ADC to recognize the strength and nature of the trainee's accent to evaluate their progress and to decide on the correct strategy according to the dialect recognized. Since the adjustment of training strategies is supposed to be done instantly, transcripts of the student's speech are unlikely to be available and it is desirable to have a robust ADC system that does not rely on speech transcripts. An automatic, text-independent dialect classification has been researched by several authors, e.g. for English [1, 2, 3, 4], Arabic [5, 6, 7] or Chinese [8, 9, 10].

For the aforementioned languages, there have been different approaches to ADC that can be divided into acoustic/phonetic [11, 12, 13], phonotactic [6, 7, 14] or prosodic [5, 15] with a number of variations and combinations in features, modeling and classification methods [16, 17]. Frequently, the base model to compare to is a GMM-UBM system [1, 4, 18, 19] that takes Mel-Frequency Cepstral Coefficients (MFCC) as features to create a Gaussian Mixture Model (GMM) as an Universal Background Model (UBM), followed by an UBM adaptation to the dialects, which need to be classified.

Text-independent ADC for German however, is a barely researched topic so far. One study focussed on German dialect classification as part of an ASR-System for broadcast speech [20] using the phonotactic and the acoustic approach. In this contribution we build upon the results of a previous paper [21] on German ADC, in which we tested various feature combinations for spontaneous speech from one corpus only. Nevertheless, in language tutoring also read speech needs to be analyzed. Both modes of speech differ in multiple ways, so they are not usually pooled in the same model.

Consequently, we use the findings from our previous experiments and apply the favored algorithmic settings and feature combinations on read speech from two different German corpora of around 500 and 830 speakers to compare the baseline ADC system on both, spontaneous and read speech.The speech data and algorithms are described in the methods section, followed by the experimental setup, results and discussion, and the conclusions.

2. Methods

2.1. Speech data

A number of speech corpora can be found that contain German speech samples and even some of regionally accented speech with a fairly large number of speakers. Two such corpora are "Regional Variants of German 1" (RVG) [22] and "Deutsch Heute" (DH) [23], both of which we used for this contribution.

RVG is a speech corpus within the BAS CLARIN Repository [22] which contains recordings of about 500 speakers from nine different dialect regions in Germany. There are samples

10.21437/SLaTE.2019-10

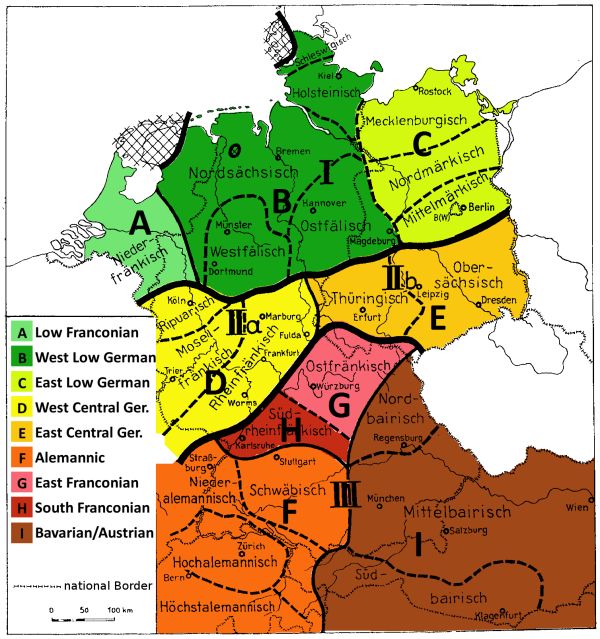

Figure 1: *The map, adapted from [24], shows the dialect regions as given in the corpus [22]. The dialect regions are colored according to the legend.*

of 1 minute of spontaneous speech as well as single numbers, commands and 30 phrases per speaker, recorded by four microphones simultaneously. The corpus is divided into the nine dialect regions shown in Figure 1, a map from 1989 that shows the then current dialect regions of the German language. The map corresponds well with the dialects in the corpus, since RVG was recorded 1996 – 1997.

As spontaneous and read speech differ significantly in many ways (e.g. speed, clarity of pronunciation etc.), the two modes were analyzed in separate models during the experiments. We decided to take the recordings from the first microphone, because the majority of speakers was recorded with it.

DH is the second corpus and was created by the foundation "Institut für Deutsche Sprache" (IDS, Institute of the German Language) in Mannheim for their project "Gesprochenes Deutsch" on variations in contemporary spoken German. The recordings for this corpus were made 2006 – 2009, approximately ten years after RVG, in 195 different locations within the German-speaking area. The recording sites therefore include not only Germany, but also other countries close to Germany such as Switzerland and Austria. DH contains around 830 speakers, most of whom are teenaged students, because the recordings were made mainly in high schools, but also at adult education centers. All speakers were selected for their speech being typical of the area they lived in and for having lived at the town the recording was made all of their life, in order to create a speech database with clearer regional variations.

There are again different parts to this corpus of read and task prompted speech. In our study, we used speech samples from a read text called "Schluss mit dem Gesundheitsterror", because it was recorded for the greatest number of speakers. Contrary to RVG, the speakers in DH were not assigned a dialect in the database, so we did our own assignment using the map from Figure 1 on the recording sites.

The number of speakers per subcorpus and dialect class is presented in Table 1. In RVG the speakers are not distributed evenly over Germany and the dialect regions themselves vary in size significantly, which leads to imbalanced classes in this corpus. The recording sites of DH, on the other hand, are distributed more evenly over the entire region where German is spoken, but the varying size of the dialect regions leads to imbalanced classes here, as well.

Table 1: *Number of speakers per dialect and subcorpus (e.g. RVG-S – spontaneous speech from RVG)*

Region	Dialect	Speakers		
		RVG-S	RVG-R	DH-R
North	A Low Franconian	44	44	20
	B West Low German	103	103	149
	C East Low German	31	31	67
Center	D West Central German	73	73	128
	E East Central German	52	53	76
South	F Alemannic / Swabian	63	63	145
	G East Franconian	19	20	39
	H South Franconian	10	10	26
	I Bavarian / Austrian	100	100	179

2.2. Models, feature analysis and classification

To compare, how a basic ADC system works for spoken versus read German, we computed our models according to the basic GMM-UBM approach, which is comprised of the following steps:

1. Feature extraction
2. Feature processing
3. Computing the UBM
4. UBM adaptation to different dialects
5. Scoring of test samples for each dialect model
6. Classification

Our first step consisted of extracting the Mel-Frequency Cepstral Coefficients (MFCC). We used fairly standard conditions for the extraction: a sampling rate of 8kHz, frame length of 25ms, Hamming-windowing and 10ms frame shift. The result were feature vectors consisting of 12 MFCC and the spectral energy per frame.

The feature vectors were then processed in Step 2 using Voice Activity Detection (VAD) through an energy threshold, RASTA-filtering to remove the spectral components, that changed at a rate different from human speech and Cepstral Mean and Variance Normalization (CMVN). Additionally, delta and double delta, as well as Shifted Delta Cepstra (SDC) were computed from the processed MFCC to incorporate temporal context for each frame. The speech data was then randomly divided into a speaker disjunct training set and a test set.

Step 3 was comprised of accumulating the feature vectors of all training speakers and training the UBM by Expectation-Maximization (EM) for 512 gaussians, which had proven to be successful in prior ADC-research.

Afterwards, the maximum-a-posteriori (MAP) algorithm was used, to adapt the means of the UBM to each dialect by using all speakers of this dialect in the training set. MAP adapts the measure of interest (in our case the means of each gaussian

in the UBM) until the probability of all data used in the adaptation process is maximized in the distributions of the adapted model. All test samples were then scored in every adapted model using log-likelihood as a measure, and the highest score per sample was determined as the corresponding dialect.

Lastly, the weighted accuracy of the model was calculated by dividing all correctly classified test samples per class by the total number of test samples per class and then taking the average accuracy over all classes. We use a weighted accuracy measure, because our classes are imbalanced in their number of speakers as can be seen in Table 1.

3. Experiments

To clarify, whether the GMM-UBM approach is suitable for ADC in read German speech as well, we decided to build upon our prior experiments with spontaneous speech [21]. We took the three combinations of feature processing that worked best over all in these experiments. Always applying VAD, we used Delta only (D), RASTA and SDC (RS) as well as RASTA, CMVN and SDC combined (RCS). Additionally, we switched between a coarse-grained dialect classification, which divided the speakers into just three main regions (low, central and high German) and the fine-grained partition of the RVG corpus into nine dialect regions.

From our baseline experiments, we used the models and results already computed on the spontaneous speech samples of RVG and the first microphone. To vary the conditions as little as we were able and to use as many speakers as possible, we decided to take the RVG read speech samples from the same microphone. Since there were approximately 30 single phrases per speaker and the prior spontaneous classification was done on samples of a longer duration (one minute), we concatenated the 30 files to three files of ten phrases per speaker in RVG. This was also done to increase the number of test samples per dialect, while maintaining a duration that our classifier could cope with. In DH-R, on the other hand, there was only one microphone to use. Also, we had over 800 speakers from the outset and therefore simply took one minute of read speech per speaker as samples for training and testing the ADC system.

For the surveyed models we divided our speakers per corpus randomly into a training set, containing 80 % and a test set with 20 % of the speakers and their associated samples. In RVG we took the dialect classes into account for the division, whereas the DH speakers who had no innate dialect assignment were divided entirely at random. The sets are speaker-disjunct by design, since the division was done by speaker rather than by sample. We then processed the models mentioned above seperately per corpus for two such training/test combinations each. Afterwards we tried the same combinations on a GMM with only 256 gaussians to determine, whether similar accuracies could be achieved in smaller models thereby allowing for lower calculation complexity, which is always useful when dealing with large amounts of data.

To implement our setup, we used the Python toolkit "Sidekit" [25], originally written for the purpose of Speaker Identification, but adaptable for ADC as well. Sidekit enables the entire experimental process from the feature extraction to the classification.

4. Results and Discussion

Figures 2 and 3 summarize our results after training and testing for three and nine dialects. The achieved weighted accuracies are plotted per feature combination and model type. The three main colors (blue, orange and green) represent the subcorpus used. Each of the main colors corresponds to two different columns, associated with the train/test set used (0 or 1) in this model.

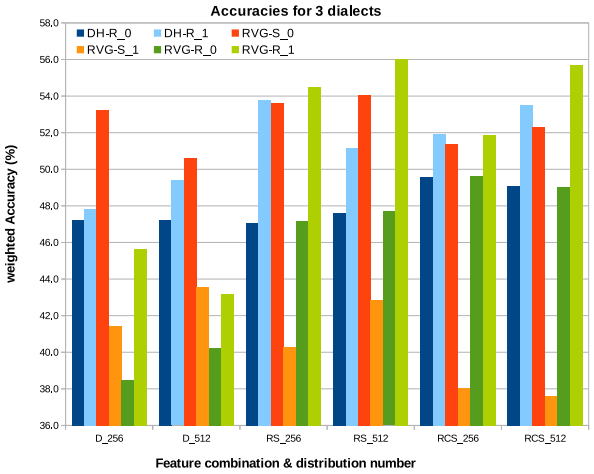

Figure 2: *Weighted accuracies of the 3 dialect models for different GMM distribution numbers and feature combinations on top of MFCC and VAD: R – RASTA, D – Delta/DoubleDelta, S – SDC, C – CMVN. Each color is one subcorpus train/test set, e.g. "RVG-R_0" refers to the read speech from RVG, microphone a , test set number 0.*

In the three dialects classification we see that the achieved accuracies span from 37.6%, which is barely above chance level, up to 56.0%. It is particularly interesting to note, how differently two train/test sets of the same subcorpus classify. This is especially apparent for the spontaneous speech (RVG-S). While part of these differences can certainly be attributed to chance, due to the random division, this does not quite explain the truly dramatic differences as can be seen especially for the spontaneous speech. Another possible explanation could be that more speakers in one test set actually speak something closer to Standard German than their assigned dialect. This would lead to the dialect model being trained on dialect speakers and tested with standard speakers, who would naturally be classified anywhere but that dialect. The same could also happen, if there were many speakers with a very slight or no dialect for training and strong dialect speakers for testing, but as the training set is bigger than the test set, it is more likely to have at least some strong dialect speakers among the training set to balance any wrongly assigned standard speakers.

Notably, the read speech corpora both achieve better classification results when using the SDC features while for RVG-S these models only worked better in one of the two train/test divisions. Another observation to be made from this graph is, that more gaussians in the GMM are actually necessary to achieve better accuracies, since the 256 GMMs generally scored at a lower accuracy then the 512 GMMs.

The nine dialects classification, also stretches over a rather great range from 14.0% to 35.3%. The two train/test sets per corpus are again quite different. As with the 3 dialects, the read speech subcorpus of RVG (RVG-R) reached the highest accuracies, only this time the difference is clearer with RVG-R reaching the highest accuracies over all, RVG-S following and DH-R

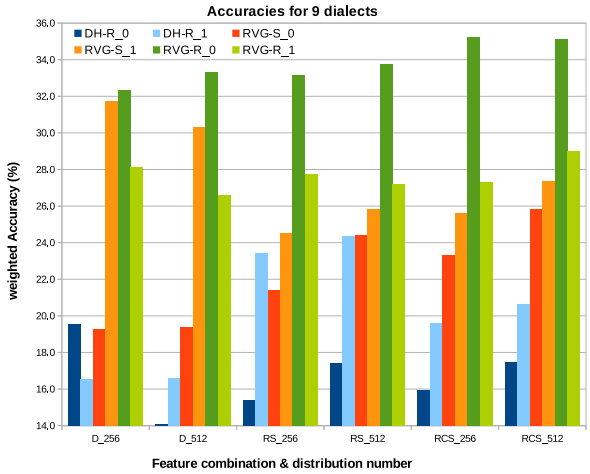

Figure 3: *Weighted accuracies of the 9 dialect models for different GMM distribution numbers and feature combinations on top of MFCC and VAD: R – RASTA, D – Delta/DoubleDelta, S – SDC, C – CMVN. Each color is one subcorpus train/test set, e.g. "RVG-R_0" refers to the read speech from RVG, microphone a , test set number 0.*

scoring the lowest accuracies. For RVG-S, the RCS combination appears to work best, whereas the highest accuracies for DH-R were reached with RS.

Compared to the other German ADC study by M. Stadtschnitzer [20], which used the 9 dialects of RVG-S as well, but employed a phonotactic approach and reached a maximum classification accuracy of 19.2%, it appears that the acoustic approach to ADC may work better for German dialects. Stadtschnitzer's second approach achieving 56.7% accuracy on 4 classes and 77.1% on 2 classes, seems to confirm this hypothesis: His further experiments use a different corpus that is more finely annotated, but smaller, for 3 southern dialects and Standard German as well as an acoustic/spectral approach with MFCC features and a Convolutional Neural Network as classifier. Since a different corpus is used and the classes and number of speakers are different from our experiments, no direct comparison is possible for our 3 dialect classification task. Likely, it would be useful to manually reassign dialects to the speakers in RVG and find out how many standard speakers there are.

Table 2: *Confusion matrix for 9 dialects and RVG-S*

	A	B	C	D	E	F	G	H	I
A	7	1	0	6	1	1	1	0	2
B	0	4	0	4	3	1	0	0	3
C	0	3	4	1	5	2	0	1	1
D	0	5	0	2	0	3	2	0	3
E	0	1	1	0	0	1	0	0	0
F	2	2	1	0	1	3	0	0	1
G	0	3	0	0	1	0	0	0	0
H	0	1	0	1	0	0	1	1	1
I	0	1	1	1	0	2	0	0	9
Σ	9	21	7	15	11	13	4	2	20
Acc(%)	77.8	19.0	57.1	13.3	0.0	23.1	0.0	50.0	45.0

One possible explanation for the accuracies of the DH-R subcorpus, which are rather high for three dialects but the lowest out of all three sub-corpora for nine dialects, is the recording time which was a decade after RVG. Due to the time gap to the RVG recordings and to the map upon which this dialect classification is based, it is possible that while the broad dialect division of low, central and high German is still present, the fine-grained dialect structure has changed in recent years and a different and more current dialect map ought to be applied to this corpus. To directly compare the models of read and spontaneous speech, the results of the best model per mode in a nine-dialect classification are presented by the corresponding confusion matrices in the Tables 2 and 3.

Table 3: *Confusion matrix for 9 dialects and RVG-R*

	A	B	C	D	E	F	G	H	I
A	12	9	0	11	4	1	0	0	6
B	6	12	2	11	8	0	0	0	6
C	0	7	15	0	3	1	3	0	4
D	4	9	1	4	6	6	3	0	11
E	2	0	3	0	5	0	0	0	0
F	0	12	0	2	7	13	0	3	11
G	0	0	0	4	0	6	6	0	1
H	3	14	0	12	0	9	0	3	6
I	0	0	0	1	0	3	0	0	15
Σ	27	63	21	45	33	39	12	6	60
Acc(%)	44.4	19.0	71.4	8.9	15.2	33.3	50.0	50.0	25.0

The first thing to catch the eye is the different number of test samples between the modes, which for RVG-S is only one third of RVG-R. While both sub-corpora consist of a comparable number of speakers, due to our experimental setup described above, we ended up with three read samples and one spontaneous sample for each speaker.

Nonetheless, these matrices show some interesting similarities and differences. As can be gathered from Table 1, the dialects A-C are northern dialects, belonging to low German (thus called, because of the lowlands in northern Germany). In both matrices it is a northern dialect, that was classified with the highest accuracy, A at 77.8% in RVG-S and C at 71.4% in RVG-R. Furthermore, dialect B was frequently confused with just about any other dialect, probably because the region of B is rather large and contains a part of Germany, that is known for speaking Standard German. Therefore, speakers of Standard German from other regions might be wrongly recognized as B and vice versa. It is also possible, that the classifiers found too few distinct characteristics in dialect B, so its speakers were assigned any dialect almost randomly.

Next in their accuracies are, again in both modes of speech, the southern dialects known as high German because of the mountainous landscape in the south of Germany. These dialects are rather distinct for human listeners, Bavarian especially, which corresponds to dialect I, but also the Franconian and Alemannic and Swabian dialects that make up the remaining three. In RVG-R, dialects G and H were recognized surprisingly well, considering their low number of speakers and the small region, whereas in RVG-S the number of speakers is so low, that the results may as well be random. At 45.0% accuracy, dialect I was well recognized in RVG-S, but frequently confused with F (Alemannic/Swabian) and D in read speech. The confusion with F is understandable insofar as human listeners tend to confuse the southern dialects for one another as well, if they do not live there themselves, and speakers with a very light Bavarian dialect may be confused for another dialect like Franconian

or even low German, if they speak standard German. This phenomenon of weakened dialects is somewhat expected in read speech where readers do not choose their own wording, thus avoiding dialect specific words, and concentrate more on pronunciation than they would in spontaneous speech.

The worst results in classification are to be found in the two dialects of central Gemany in both RVG-R and RVG-S. This result is rather unexpected, since dialect E, at least, is a strong dialect and commonly easy to recognise as east German or Saxon by human listeners. In RVG-S though, it was never classified correctly and even in RVG-R it was confused primarily for dialects B and F. Possibly there are mostly speakers with a slight dialect to their speech in region E, thus explaining the misclassification as B. The confusion with F on the other hand, can be explained by a similarity in vowel quality, at least for Swabian (part of F) and Saxon (part of E). The two dialects are usually distinguished by human listeners in terms of dialect specific grammar and vocabulary, but they do show similarities in pronounciation. Lastly, there is dialect D with a reasonable number of speakers, that still only scored 13.3% accuracy in RVG-S and 8.9% in RVG-R. In both modes of speech, F was primarily confused for A and B indicating a siginificant number of test speakers who speak standard German.

In sum, while there are slight differences to be observed between the results of the read and spontaneous speech classifier, both matrices show mainly similar results as to the broad classification of the three regions.

5. Conclusions

Our experiments focused on the comparison of ADC models for two different modes of spoken German – read and spontaneous speech from the speech corpora RVG and DH. We have found the classification accuracies of a basic GMM-UBM classification system to be relatively similar on both modes of speech, which leads to the conclusion that the distinguishable features of a speaker's dialect are based on the same mechanisms and less influenced by the speaking mode. Our indicative results require further clarification and also more sophisticated classification methods in a next step. To have more robust results, more train/test sets or k-fold Cross Validation need to be tried. Different classifiers like Support Vector Machines, logistic regression or Artificial Neural Networks could be employed. We will also test the different corpora and speaking modes against their corresponding models to further determine the influence of the mode of speech on the performance of dialect classification.

6. Acknowledgments

We would like to thank the IDS Mannheim for providing their corpus "Deutsch Heute" in our experiments on German dialects.

7. References

[1] A. Hanani, M. J. Russell, and M. J. Carey, "Human and computer recognition of regional accents and ethnic groups from British English speech," *Computer Speech & Language*, vol. 27, pp. 59–74, 2013. [Online]. Available: https://doi.org/10.1016/j.csl.2012.01.003

[2] M. Najafian, S. Khurana, S. Shon, A. Ali, and J. R. Glass, "Exploiting convolutional neural networks for phonotactic based dialect identification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, 2018, pp. 5174–5178. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461486

[3] H. Wang and V. J. van Heuven, "Relative contribution of vowel quality and duration to native language identification in foreign-accented English," in *Proceedings of the 2nd International Conference on Cryptography, Security and Privacy, ICCSP 2018, Guiyang, China, March 16-19, 2018*, 2018, pp. 16–20. [Online]. Available: https://doi.org/10.1145/3199478.3199507

[4] G. Brown, "Automatic accent recognition systems and the effects of data on performance," in *Odyssey 2016: The Speaker and Language Recognition Workshop, Bilbao, Spain, June 21-24, 2016*, 2016, pp. 94–100. [Online]. Available: https://doi.org/10.21437/Odyssey.2016-14

[5] S. Bougrine, H. Cherroun, and D. Ziadi, "Hierarchical classification for spoken Arabic dialect identification using prosody: Case of Algerian dialects," *CoRR*, vol. abs/1703.10065, 2017. [Online]. Available: http://arxiv.org/abs/1703.10065

[6] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic dialect identification using phonotactic modeling," in *Proceedings of the Workshop on Computational Approaches to Semitic Languages, SEMITIC@EACL 2009, Athens, Greece, March 31, 2009*, 2009, pp. 53–61. [Online]. Available: https://aclanthology.info/papers/W09-0807/w09-0807

[7] M. Akbacak, D. Vergyri, A. Stolcke, N. Scheffer, and A. Mandal, "Effective Arabic dialect classification using diverse phonotactic models," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011, pp. 737–740. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2011/i11_0737.html

[8] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S.-Y. Yoon, "Accent detection and speech recognition for Shanghai-accented Mandarin," in *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, 2005, pp. 217–220. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2005/i05_0217.html

[9] J. Hou, Y. Liu, T. F. Zheng, J. Ø. Olsen, and J. Tian, "Multi-layered features with SVM for Chinese accent identification," in *2010 International Conference on Audio, Language and Image Processing*, 2010, pp. 25–30. [Online]. Available: https://doi.org/10.1109/ICALIP.2010.5685023

[10] Y. Lei and J. H. L. Hansen, "Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, pp. 85–96, 2011. [Online]. Available: https://doi.org/10.1109/TASL.2010.2045184

[11] P. A. Torres-Carrasquillo, D. E. Sturim, D. A. Reynolds, and A. McCree, "Eigen-channel compensation and discriminatively trained Gaussian mixture models for dialect and accent recognition," in *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, 2008, pp. 723–726. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2008/i08_0723.html

[12] F. Biadsy, J. Hirschberg, and M. Collins, "Dialect recognition using a phone-GMM-supervector-based SVM kernel," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 753–756. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2010/i10_0753.html

[13] F. Biadsy, "Automatic dialect and accent recognition and its application to speech recognition," Ph.D. dissertation, Columbia University, 2011. [Online]. Available: https://doi.org/10.7916/D8M61S68

[14] M. A. Zissman, T. P. Gleason, D. Rekart, and B. L. Losiewicz, "Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, ICASSP '96, Atlanta, Georgia, USA, May 7-10, 1996*, 1996, pp. 777–780. [Online]. Available: https://doi.org/10.1109/ICASSP.1996.543236

[15] N. B Chittaragi, A. Prakash, and S. Koolagudi, "Dialect identification using spectral and prosodic features on single and ensemble classifiers," *Arabian Journal for Science and Engineering*, vol. 43, 2017. [Online]. Available: https://doi.org/10.1007/s13369-017-2941-0

[16] M. Najafian, S. Safavi, P. Weber, and M. J. Russell, "Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems," in *Odyssey 2016: The Speaker and Language Recognition Workshop, Bilbao, Spain, June 21-24, 2016*, 2016, pp. 132–139. [Online]. Available: https://doi.org/10.21437/Odyssey.2016-19

[17] Q. Zhang, H. Boril, and J. H. L. Hansen, "Supervector pre-processing for PRSVM-based Chinese and Arabic dialect identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 2013, pp. 7363–7367. [Online]. Available: https://doi.org/10.1109/ICASSP.2013.6639093

[18] G. Liu and J. H. L. Hansen, "A systematic strategy for robust automatic dialect identification," in *Proceedings of the 19th European Signal Processing Conference, EUSIPCO 2011, Barcelona, Spain, August 29 - Sept. 2, 2011*, 2011, pp. 2138–2141. [Online]. Available: http://ieeexplore.ieee.org/document/7074191/

[19] A. Lazaridis, E. el Khoury, J. Goldman, M. Avanzi, S. Marcel, and P. N. Garner, "Swiss french regional accent identification," in *Odyssey 2014: The Speaker and Language Recognition Workshop, Joensuu, Finland, June 16-19, 2014*, 2014. [Online]. Available: https://isca-speech.org/archive/odyssey_2014/abstracts.html#abs29

[20] M. Stadtschnitzer, "Robust Speech Recognition for German and Dialectal Broadcast Programmes," Ph.D. dissertation, University of Bonn, Germany, 2018. [Online]. Available: http://hss.ulb.uni-bonn.de/2018/5236/5236.htm

[21] J. Dobbriner and O. Jokisch, "Towards a dialect classification in German speech samples," in *Proc. 21th International Conference Speech and Computer (SPECOM), August 20-25, 2019*. Istanbul, Turkey: Springer LNAI, 2019.

[22] S. Burger and F. Schiel, "RVG 1 - a database for regional variants of contemporary German," in *Proc. of the 1st Int. Conf. on Language Resources and Evaluation*, Granada, Spain, 1998, pp. 1083–1087. [Online]. Available: https://www.phonetik.uni-muenchen.de/forschung/publikationen/Burger-98-RVG1.ps

[23] S. Kleiner, "'Deutsch heute' und der Atlas zur Aussprache des deutschen Gebrauchsstandards," in *Regionale Variation des Deutschen*. Berlin/Boston: de Gruyter, 2015, pp. 489–518.

[24] H. Mettke, *Mittelhochdeutsche Grammatik*. Leipzig, Germany: Bibliographisches Institut, 1989.

[25] A. Larcher, K. A. Lee, and S. Meignier, "An extensible speaker identification sidekit in Python," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, 2016, pp. 5095–5099. [Online]. Available: https://doi.org/10.1109/ICASSP.2016.7472648