

Text-Independent Dialect Classification in Read and Spontaneous Speech

Oliver Jokisch and Johanna Dobbriner

Leipzig University of Telecommunications (HFTL), Institute of Communications Engineering, Germany
 Technological University Dublin, School of Computing, Ireland
 jokisch@hft-leipzig.de, johanna.dobbriner@gmail.com

Abstract

Linguistic diversity and the fundamental freedom of users of language technology (LT) to access information and knowledge in their own language(s) or dialect(s) lead to certain requirements with regard to truly multilingual language technologies, in particular for under-resourced languages and application domains. One key issue is the low-threshold creation of high-quality speech corpora and the corresponding annotation data to train powerful analysis or classification algorithms as a base for state-of-the-art language technology. Dialects constitute an important part of the mentioned linguistic diversity. In this contribution, we shortly discuss basic concepts of automatic dialect classification (ADC) with a focus on methods that do not require expensive prior annotation or labeling. Starting with text-independent ADC methods for well-studied major languages and summarizing our results of a case study on read and spontaneous German, we convey necessary development steps for under-resourced language data and a possible processing chain.

Keywords: dialect classification, read and spontaneous speech, under-resourced language, corpus design, feature selection

Résumé

Sprachliche Vielfalt und ein Grundrecht von Sprachtechnologie-Anwendern, auf Informationen und Wissen in ihrer eigenen Sprache oder sogar ihrem eigenen Dialekt zuzugreifen, führen zu dezidierten Anforderungen an tatsächlich mehrsprachige Technologien, insbesondere für Sprachen und Anwendungsdomänen mit geringen Ressourcen. Das zentrale Thema ist eine niedrighschwellige Erstellung hochwertiger Sprachkorpora und zugehöriger Annotationsdaten, um leistungsstarke Analyse- oder Klassifizierungsalgorithmen zu trainieren, die eine Grundlage aktueller Sprachtechnologie darstellen. Dialekte sind ein wesentlicher Bestandteil der genannten sprachlichen Vielfalt. In diesem Beitrag diskutieren wir Konzepte der automatischen Dialektklassifizierung (ADC) mit dem Fokus auf Methoden, die keine aufwendige, vorherige Annotation erfordern. Ausgehend von textunabhängigen ADC-Methoden für ausgiebig untersuchte Hauptsprachen und einer Zusammenfassung der Ergebnisse einer Fallstudie zu gelesenen und spontanem Deutsch leiten wir Entwicklungs- und Verarbeitungsschritte für unterrepräsentierte Sprachdaten ab.

1. Dialect Classification in Major Languages

Particularly from a perspective of language and speech technologies, the differences between dialects are commonly less distinctive than the ones between single languages or language groups. Typical intra-language variations are small, with less-defined borders between dialect realizations. Both native and migrant speakers often exhibit a mixture of different dialects with regard to their vita. Therefore, Automatic Dialect Classification (ADC) combines known principles from language and speaker identification to automatically recognize a regional dialect of a given language from speech samples or corresponding transcripts. ADC methods can constitute a complementary technology in the area of under-resourced languages, e.g. to increase the performance of Automatic-Speech-Recognition (ASR) modules or to enable Intelligent-Language-Tutoring (ILT) systems with the goal to reduce or even to improve a regional accent.

Methods of accent reduction or improvement require robust ADC algorithms to identify the dialect and to evaluate the training progress, preferably avoiding transcribed speech. The so-called text-independent dialect classification has been researched by several authors, e.g. for English (Hanani et al., 2013; Najafian et al., 2018; Wang and van Heuven, 2018; Brown, 2016), Arabic (Bougrine et al., 2017; Biadisy et al., 2009; Akbacak et al., 2011) and Chinese (Zheng et al., 2005; Hou et al., 2010; Lei and Hansen, 2011). The ADC approaches can be roughly categorized as either acoustic/phonetic (Torres-Carrasquillo et al., 2008;

Biadisy et al., 2010; Biadisy, 2011), phonotactic (Biadisy et al., 2009; Akbacak et al., 2011; Zissman et al., 1996) or prosodic (Bougrine et al., 2017; Chittaragi et al., 2017) including variations and combinations in features, modeling and classification methods, cf. (Najafian et al., 2016; Zhang et al., 2013). The most common ADC approaches rely on Mel-Frequency Cepstral Coefficients (MFCCs) for feature analysis and Gaussian Mixture Models (GMMs) with a Universal Background Model (UBM) for the classification task, followed by UBM adaptation to each of the target dialects (Hanani et al., 2013; Brown, 2016; Liu and Hansen, 2011; Lazaridis et al., 2014).

Apart from the above-mentioned works on major languages, text-independent dialect classification is still an under-researched and under-resourced topic – even with regard to German as a major European language. However, German dialect classification, based on phonotactic and acoustic approaches, was previously studied as part of an ASR system for broadcast speech (Stadtschnitzer, 2018).

Following the ADC approaches in other major languages we summarize our results of previous case studies on German ADC (Dobbriner and Jokisch, 2019a; Dobbriner and Jokisch, 2019b) with restricted training data, in which we tested various feature combinations for read and spontaneous speech from two corpora with 500 and 830 speakers respectively. Both modes of speech differ in multiple ways, so they are not pooled in the same model. Afterwards, we discuss the lessons learned from the viewpoint of under-resourced language data.

2. Text-Independent Dialect Classification in Read and Spontaneous German Speech

2.1. Speech Corpora

There are various German speech databases for multiple tasks within LT research and development, including selected databases with regionally accented speech, such as “Regional Variants of German 1” (RVG) (Burger and Schiel, 1998) and “Deutsch Heute” (DH) (Kleiner, 2015), that we used in our study (Dobbriner and Jokisch, 2019a). Both corpora are well-annotated and appropriate for many linguistic studies. In terms of training and test material, in particular for state-of-the-art methods in (deep) learning, a few hundred speakers with a few ten phrases per speaker has to be treated as low-resourced data.

RVG is a corpus within the BAS CLARIN Repository (Burger and Schiel, 1998), which comprises recordings of 500 speakers from nine different dialect regions in Germany. There are samples of about 1 min. of spontaneous speech as well as single numbers, commands and 30 phrases per speaker, recorded by four microphones simultaneously. The corpus is divided into nine dialect regions, illustrated by the sample speakers in Figure 1 with regard to their current home when the corpus was recorded, 1996 – 1997.

The DH corpus, recorded 2006 – 2009 in Germany, Switzerland and Austria by the Institute of the German Language (IDS) in Mannheim, contains variations in contemporary spoken German. In total, DH includes 830 speakers, mainly from high schools and further education centers. There are different parts of read and task-prompted speech. Contrary to RVG, the speakers in DH were not assigned a dialect in the database, i. e., we processed our own assignment with regard to (Mettke, 1989) and (Burger and Schiel, 1998). Table 1 summarizes the number of speakers in the different regions and dialects. The RVG speakers are not distributed uniformly over Germany, and the dialect re-

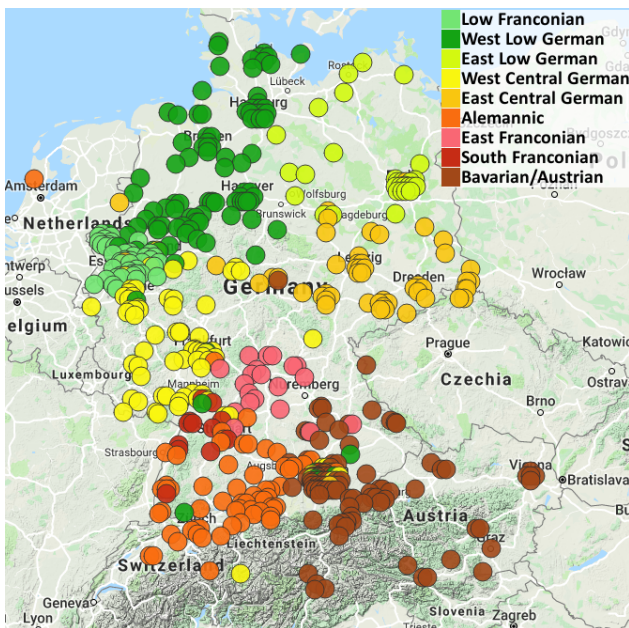


Figure 1: Dialects of RVG speakers by current home

Table 1: No. of speakers per dialect and spontaneous/read subcorpora, extension S/R (Dobbriner and Jokisch, 2019a)

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

gions themselves vary in size, which leads to imbalanced classes. The DH recording sites, on the other hand, are uniformly distributed, but the varying size of the dialect regions leads to imbalanced classes as well.

2.2. Parameter Extraction and Classification

To compare ADC for spoken versus read German, we developed a tool chain (Dobbriner and Jokisch, 2019b) according to the 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 test.

The first step consists of extracting Mel-Frequency Cepstral Coefficients (MFCC) with a sampling rate of 8kHz, frame length of 25ms, Hamming-windowing and 10ms frame shift. The resulting feature vectors consist of 12 MFCC and the spectral energy per frame. The feature vectors are processed in Step 2 by using Voice Activity Detection (VAD) through an energy threshold, RASTA-filtering and Cepstral Mean and Variance Normalization (CMVN). Additionally, delta and double delta, and Shifted Delta Cepstra (SDC) are computed from the MFCC to incorporate temporal context for each frame. In later experiments a sampling rate of 16kHz with similar overall results (Dobbriner, 2019) was tested, which led to higher calculation complexity. Afterwards, the speech data is randomly divided into a speaker-disjunct training set and a test set. Step 3 is accumulating the feature vectors of all training speakers and training the UBM by Expectation-Maximization (EM) for 256 or 512 gaussians, which had proven to be successful in prior ADC research. In step 4, the maximum-a-posteriori (MAP) algorithm is used, to adapt the means of the UBM to each dialect by using all speakers of this dialect category 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 is maximized in the distributions of the

adapted model. All test samples are scored in step 5 for every adapted model using log-likelihood, and the highest score per sample is determined as the corresponding dialect. Lastly, the weighted accuracy of the model is calculated by dividing all correctly classified test samples per class by the total number of test samples per class, aiming at the average accuracy over all classes. We always refer to a weighted accuracy measure, since our classes are imbalanced in their number of speakers, cf. Table 1.

Our ADC processing chain, including feature extraction, classification and evaluation is based on the Python toolkit “Sidekit” (Larcher et al., 2016), which was originally designed for speaker identification with a certain similarity to the task of dialect classification.

2.3. Experiments and Results

To realistically experiment with our restricted German dialect data, we switched between a coarse-grained dialect classification, which divided the speakers into just three main regions (low/North, central and upper/South German), that are widely agreed among linguists, and a fine-grained partition of both corpora, RVG and DH, into nine dialect regions with the disadvantage of sparse training data in some classes. Longer speech samples and monologues up to one minute proved to be suitable for training and testing of our classifiers on spontaneous speech in “RVG-S” (Dobbriner and Jokisch, 2019b). While maintaining an appropriate sample duration for the classifiers and to increase the number of samples per dialect, we therefore concatenated approximately 30 read phrases per speaker in “RVG-R” to three files of ten phrases each. For subcorpus “DH-R”, we selected one minute of read speech per speaker for training and testing the ADC system.

All speakers were randomly partitioned into speaker-disjunct sets for training (80%) and testing (20%). Applying the processing chain described in section 2.2., we systematically analyzed different feature combinations in the three- and nine-dialect classification, based on RASTA, Delta/DoubleDelta, SDC and CMVN. Beside the best method from our previous study (GMM based on 512 gaussians), we tested the same feature combinations with 256 gaussians only to determine, whether similar accuracies can be achieved by smaller models with lower calculation complexity. Furthermore, to visualize effects of the random speaker selection, we repeated the classification. The different train/test-set constellations are marked by “0” or “1” in the following. As an example for spontaneous speech (RVG-S corpus), Figure 2 summarizes the test results of the more challenging nine-dialect classification: The weighted accuracies are sorted by feature combination and model type. Colored bars represent the spontaneous/read subcorpora (RVG-S, RVG-R, DH-R) and concrete test sets (0/1). Considering a chance level of 11.1 %, the nine-dialect classification in the test set stretches over a huge range from a weighted overall accuracy of 14.0% to 35.3%, and the two test sets per corpus behave diverse too. In a three-dialect classification, the overall accuracies span from 37.6%, which is barely above chance level (33.3%), up to 56.0%. In both classification scenarios, the RVG-R subcorpus reached the highest accuracies, while the test results on

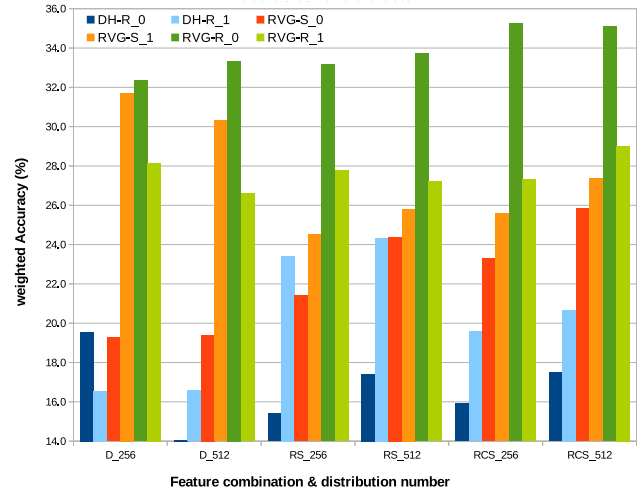


Figure 2: Nine-dialect classification for varying feature sets and 256/512 GMMs: R – RASTA, D – Delta/DoubleDelta, S – SDC, C – CMVN (Dobbriner and Jokisch, 2019a)

DH-R were significantly worse. For the RVG-S subcorpus, the RCS feature combination performed best, whereas the highest accuracies for DH-R were reached with the RS feature set. More gaussians in the GMM lead to higher classification accuracies – in (Dobbriner and Jokisch, 2019b) we surveyed constellations of 64 ... 512 gaussians. Overall, the resulting accuracies are far from optimal but certainly above chance level. As a baseline system, the GMM-UBM approach seems to be effective for distinguishing dialect samples, and aside from the model quality, there are some potential reasons for the low accuracies observed, e.g. some dialects and dialect groups are similar and therefore hard to distinguish, even for human listeners. Besides, a few speakers in the test corpora articulate close to standard German, which may contradict to the forced-assigned dialects in the training. A general shortcoming is the unequal distribution of speakers per dialect as well as the unbalanced size of dialect regions. For a three-dialect categorization, the variation of speech within the regions seems too high, so that our models are not specific enough to allow for a robust classification. As shown in the confusion matrix of the best three-dialect model in Table 2, both northern and southern German are distinguished relatively well, but the central region is frequently confused with the northern region.

The results of the nine-dialect classification on RVG-S are presented in the corresponding confusion matrix in Table 3. Based on the described acoustic approach, our results for RVG-S – a weighted overall accuracy of 31.7% for nine dialects (and 53.2% for three dialects) – outperform the re-

Table 2: Three-dialect accuracies (RVG-S)

	North	Central	South
North	23	10	6
Central	9	10	10
South	5	6	23
Σ	37	26	39
Accuracy(%)	62.2	38.5	59.0

Table 3: Nine-dialect accuracies (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	1	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

sults of a phonotactic approach on RVG-S (Stadtschnitzer, 2018), which led to a nine-dialect accuracy of 19.2% only. In contrast, a second approach in (Stadtschnitzer, 2018) with acoustic-spectral features and a convolutional neural network (CNN) classifier on a small, well-annotated corpus achieved 56.7% accuracy on four dialect classes and 77.1% on two classes, but due to the different corpus and other constellations of classes and speakers, a direct comparison with our results is not possible.

Some peculiarities of the German dialects and spontaneous vs. read speech as well as potential explanations for the observed confusions in RVG-S, RVG-R and DH-R are addressed in (Dobbriner, 2019). After some modifications in feature processing, and in particular by a manual correction and re-assignment of dialect speakers from the original RVG and DH corpora into other/partly new classes, our classification results could be improved. The nine-dialect accuracy could be thereby increased up to 36.3%. Another trial, only differentiating between standard and dialect speech, achieved at a maximum accuracy of 76.0%.

In general, the ADC accuracies of our basic GMM-UBM classification system are similar for spontaneous and read speech, which indicates that the distinguishable features of a speakers’ dialect are based on the same mechanisms and less influenced by the speaking mode. Of course, the surveyed approach requires a sophisticated back-end classification method – in a further step we tested different classifiers like support vector machines, logistic regression or artificial neural networks but the accuracies diversified about 2% only for our example of a nine-dialect classification.

3. ADC Conception for Under-resourced Languages and some Conclusions

As introduced for major languages in section 1., our modeling via GMM-UBM has proved a good baseline performance for German dialect classification too, in particular in the context of the rather low amount of training data. The ADC task does not require transcribed speech. The proposed tool chain in section 2.2. – feature extraction and processing, UBM computing, maximum-a-posteriori (MAP) adaptation to different dialects and scoring of test samples for each dialect model – based on the Python toolkit “Sidekit”, is appropriate for all relevant tasks in analysis and classification. Toolboxes like WEKA (Frank

et al., 2016), can support the classification by alternative methods but the potential accuracy improvements seem to be limited. With regard to a few hundred speakers and phrases, deep learning techniques seem superfluous.

Acoustic, namely spectral, features like MFCC are suitable for the ADC task, and modifications in the feature analysis and processing offer potential for optimization. Our results suggest that ADC may also work on shorter audio samples below a length of 1 min. Phonotactic and prosodic measures can be applied as well, but they have been barely discriminative in our dialect classification tests. A combination of approaches such as the Phone-Supervector method in (Biadsky, 2011), combining conventional phone recognition and GMM-mean super vectors on Arabic ADC, is an interesting option, if adequate components like a phone recognizer are available for the language in question.

For under-resourced languages and applications, the design and annotation quality of the training corpus are the most significant factor of influence. Naturally, ADC training requires well-annotated classes that reflect current regional varieties of the language and contain a sufficiently large number of speakers as well as somewhat balanced classes. Regional varieties may even change within a decade due to our dynamic life environment including migration and media influence, although that may be less of a concern for languages spoken only in an isolated location. To cover more than two (standard vs. dialect speech) or three regional dialects, the minimum requirement is a few hundred speakers with a few tens of longer phrases. A corpus size of about 500+ speakers as in our experiments demands a precise annotation and effort in manual corrections.

To construct mid-size corpora up to a few thousand speakers, existing speech data from different sources with similar recording conditions can be merged, which usually calls for a new, consistent annotation of the samples according to dialect and strength of dialect, preferably by semi-automatic means. Local broadcast programs as in (Stadtschnitzer, 2018) can be an appropriate source of information and should be a good option for cooperation. Another, potentially low-cost method, is based on crowd-sourcing approaches to reach volunteers more easily and widespread, as demonstrated with the “Voice Äpp” for Swiss German (Leemann et al., 2015) and the “English Dialects App” for British varieties (Leemann et al., 2018).

4. Acknowledgment

We would like to thank IDS Mannheim for providing the corpus “Deutsch Heute” (partly used in our experiments). The open-source corpus “Regional Variants of Contemporary German” from the Bavarian Archive for Speech Signals/CLARIN Repository was also quite helpful.

5. Bibliographical References

- Akbaçak, M., Vergyri, D., Stolcke, A., Scheffer, N., and Mandal, A. (2011). Effective Arabic dialect classification using diverse phonotactic models. In *INTER-SPEECH, Florence, Italy, August 2011*, pages 737–740.
- Biadsky, F., Hirschberg, J., and Habash, N. (2009). Spoken Arabic dialect identification using phonotactic modeling. In *Proc. Workshop on Computational Approaches*

- to Semitic Languages, *SEMITIC@EACL 2009, Athens, March 2009*, pages 53–61.
- Biadisy, F., Hirschberg, J., and Collins, M. (2010). Dialect recognition using a phone-GMM-supervector-based SVM kernel. In *INTERSPEECH Makuhari, Japan, September 2010*, pages 753–756.
- Biadisy, F. (2011). *Automatic Dialect and Accent Recognition and its Application to Speech Recognition*. Ph.D. thesis, Columbia University.
- Bougrine, S., Cherroun, H., and Ziadi, D. (2017). Hierarchical classification for spoken Arabic dialect identification using prosody: Case of Algerian dialects. *CoRR*, abs/1703.10065.
- Brown, G. (2016). Automatic accent recognition systems and the effects of data on performance. In *Odyssey: The Speaker and Language Recognition Workshop, Bilbao, Spain, June 2016*, pages 94–100.
- Burger, S. and Schiel, F. (1998). RVG 1 - a database for regional variants of contemporary German. In *Proc. of the 1st Int. Conf. on Language Resources and Evaluation*, pages 1083–1087, Granada, Spain.
- Chittaragi, N. B., Prakash, A., and Koolagudi, S. G. (2017). Dialect identification using spectral and prosodic features on single and ensemble classifiers. *Arabian Journal for Science and Engineering*, 43.
- Dobbriner, J. and Jokisch, O. (2019a). Implementing and evaluating methods of dialect classification on read and spontaneous German speech. In *Proc. 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE), September 2019*, pages 53–58, Graz, Austria.
- Dobbriner, J. and Jokisch, O. (2019b). Towards a dialect classification in German speech samples. In *Proc. 21th Intern. Conf. Speech and Computer (SPECOM), August 2019*, pages 64–74, Istanbul, Turkey. Springer LNAI.
- Dobbriner, J. (2019). Automatic Dialect Classification in Spoken German. Master’s thesis, Univ. Leipzig/HfTL.
- Frank, E., Hall, M. A., and Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", 4th Ed.* Morgan Kaufmann, Amsterdam.
- Hanani, A., Russell, M. J., and Carey, M. J. (2013). Human and computer recognition of regional accents and ethnic groups from British English speech. *Computer Speech & Language*, 27:59–74.
- Hou, J., Liu, Y., Zheng, T. F., Olsen, J. Ø., and Tian, J. (2010). Multi-layered features with SVM for Chinese accent identification. In *Intern. Conf. on Audio, Language and Image Processing*, pages 25–30.
- Kleiner, S. (2015). ‘Deutsch heute’ und der Atlas zur Aussprache des deutschen Gebrauchsstandards. In *Regionale Variation des Deutschen*, pages 489–518. de Gruyter, Berlin/Boston.
- Larcher, A., Lee, K. A., and Meignier, S. (2016). An extensible speaker identification sidekit in Python. In *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP, Shanghai, March 2016*, pages 5095–5099.
- Lazaridis, A., el Khoury, E., Goldman, J., Avanzi, M., Marcel, S., and Garner, P. N. (2014). Swiss french regional accent identification. In *Odyssey 2014: The Speaker and Language Recognition Workshop, Joensuu, Finland, June 16-19, 2014*.
- Leemann, A., Kolly, M.-J., Goldman, J.-P., Dellwo, V., Hove, I., Almajai, I., Grimm, S., Robert, S., and Wanitsch, D. (2015). Voice app: a mobile app for crowdsourcing Swiss German dialect data. In *INTERSPEECH, Dresden, Germany, September 2015*, pages 2804–2808.
- Leemann, A., Kolly, M.-J., and Britain, D. (2018). The English dialects app: The creation of a crowdsourced dialect corpus. *Ampersand*, 5:1-17.
- Lei, Y. and Hansen, J. H. L. (2011). Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese. *IEEE Trans. Audio, Speech & Language Processing*, 19:85–96.
- Liu, G. and Hansen, J. H. L. (2011). A systematic strategy for robust automatic dialect identification. In *Proc. 19th European Signal Processing Conference, EU-SIPCO, Barcelona, August 2011*, pages 2138–2141.
- Mettke, H. (1989). *Mittelhochdeutsche Grammatik*. Bibliographisches Institut, Leipzig, Germany.
- Najafian, M., Safavi, S., Weber, P., and Russell, M. J. (2016). Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems. In *Odyssey: The Speaker and Language Recognition Workshop, Bilbao, June 2016*, pages 132–139.
- Najafian, M., Khurana, S., Shon, S., Ali, A., and Glass, J. R. (2018). Exploiting convolutional neural networks for phonotactic based dialect identification. In *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP, Calgary, April 2018*, pages 5174–5178.
- Stadtschnitzer, M. (2018). *Robust Speech Recognition for German and Dialectal Broadcast Programmes*. Ph.D. thesis, University of Bonn, Germany.
- Torres-Carrasquillo, P. A., Sturim, D. E., Reynolds, D. A., and McCree, A. (2008). Eigen-channel compensation and discriminatively trained Gaussian mixture models for dialect and accent recognition. In *INTERSPEECH 2008, Brisbane, September 2008*, pages 723–726.
- Wang, H. and van Heuven, V. J. (2018). Relative contribution of vowel quality and duration to native language identification in foreign-accented English. In *Proc. 2nd Intern. Conf. on Cryptography, Security and Privacy, ICCSP 2018, Guiyang, March 2018*, pages 16–20.
- Zhang, Q., Boril, H., and Hansen, J. H. L. (2013). Supervector pre-processing for PRSVM-based Chinese and Arabic dialect identification. In *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP, Vancouver, May 2013*, pages 7363–7367.
- Zheng, Y., Sproat, R., Gu, L., Shafran, I., Zhou, H., Su, Y., Jurafsky, D., Starr, R., and Yoon, S.-Y. (2005). Accent detection and speech recognition for Shanghai-accented Mandarin. In *INTERSPEECH, Lisbon, Portugal, September 2005*, pages 217–220.
- Zissman, M. A., Gleason, T. P., Rekart, D., and Losiewicz, B. L. (1996). Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech. In *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing Conference Proceedings, ICASSP, Atlanta, USA, May 1996*, pages 777–780.