Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6423–6428
Marseille, 11–16 May 2020

Using Automatic Speech Recognition in Spoken Corpus Curation

Jan Gorisch,[1] Michael Gref,[2] Thomas Schmidt[1]
[1]Leibniz-Institute for the German Language (IDS), Germany
[2]Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Germany
gorisch@ids-mannheim.de, michael.gref@iais.fraunhofer.de, thomas.schmidt@ids-mannheim.de

Abstract

The newest generation of speech technology caused a huge increase of audio-visual data nowadays being enhanced with orthographic transcripts such as in automatic subtitling in online platforms. Research data centers and archives contain a range of new and historical data, which are currently only partially transcribed and therefore only partially accessible for systematic querying. Automatic Speech Recognition (ASR) is one option of making that data accessible. This paper tests the usability of a state-of-the-art ASR-System on a historical (from the 1960s), but regionally balanced corpus of spoken German, and a relatively new corpus (from 2012) recorded in a narrow area. We observed a regional bias of the ASR-System with higher recognition scores for the north of Germany vs. lower scores for the south. A detailed analysis of the narrow region data revealed – despite relatively high ASR-confidence – some specific word errors due to a lack of regional adaptation. These findings need to be considered in decisions on further data processing and the curation of corpora, e.g. correcting transcripts or transcribing from scratch. Such geography-dependent analyses can also have the potential for ASR-development to make targeted data selection for training/adaptation and to increase the sensitivity towards varieties of pluricentric languages.

Keywords: oral corpora, automatic transcription, ASR, corpus curation, pluricentric, spoken German, Ripuarian

1. Introduction

The Archive for Spoken German (AGD, Stift and Schmidt (2014), http://agd.ids-mannheim.de) is part of the CLARIN-D Centre IDS and specializes in data of spoken German, mostly corpora of natural interaction and data on varieties of German. The AGD develops and curates such corpora and makes them available to the scientific community. Besides detailed metadata on recordings and speakers, the key to accessibility are transcripts of the recordings allowing systematic queries and the application of other automated methods. So far, the high cost of manual transcription results in large parts of the corpora remaining untranscribed and therefore accessible only to a limited degree. Automatic Speech Recognition (ASR) is often claimed to be a way through that transcription bottleneck. The present paper describes some first steps in exploring whether or not that claim can be fulfilled and if so, for which type of data ASR is suitable and what factors influence the quality of ASR results.

The latest developments in the field of ASR – artificial neural networks, deep learning, and the intorduction of LF-MMI models (Povey et al., 2016) – have heaved speech technology from a level that was merely useful for limited vocabulary and clean audio tasks to a level where it could be applied to various recording conditions and large vocabulary challenges as can be found in Oral History interviews (Gref et al., 2018; Leh et al., 2018). Most of the developments have been pushed by the Kaldi-ASR-toolkit (Povey et al., 2011). Although the Word Error Rate (WER) is still below a level where a reader of an automatically derived transcript could trust that every word is correct, most of the extracted content words can already be considered as an initial starting point for historical or language researchers for accessing the content of interview data.

The specific challenge is therefore to create transcripts with a relatively low investment of time, a factor that becomes almost irrelevant considering automatic processing. Therefore we want to explore how far we can get with a state-of-the-art ASR system.

With the ultimate aim of providing correct transcripts to the users of the data center, the aim of this paper is to evaluate the performance of a state-of-the-art ASR system with respect to the usability of the resutling transcripts in the sense of (i) which ASR-transcripts are good enough for certain research communities, (ii) which ASR-transcripts are worth sending to manual correction, i.e. can we save time by correcting transcripts rather than by transcribing the audio from scratch? As ASR systems tend to perform better with data that are similar to the data that they were trained on, an additional aim (iii) of this paper is to reveal the weak points of an ASR system, i.e. those points where the system lacks training data in order to improve systems in a targeted manner. These lacks can be age related, gender related, or regional. The latter is especially relevant for pluricentric languages, such as German, i.e. the data in the AGD.

We assume that the quality of ASR results depends on several factors: proximity to the standard (or similarity to training data) / recording quality (background noise etc.) / degree of interactivity (overlap, speaker change). This would require to compose a systematic test-set and rigorous testing. The data in the AGD would allow for such an enterprise of creating such a test-set with a wide range of properties. With the pilot study presented in this paper we intend to demonstrate this potential.

2. Material and Method

2.1. Corpora

The AGD[1] contains data of various origins and various types. It contains conversational corpora, variation corpora

[1]Most of the corpora are disseminated through the Database of Spoken German (http://dgd.ids-mannheim.de). Corpora that have not been sufficiently curated, yet, are accessible through the personal service of the AGD (http://agd.ids-mannheim.de).

from within the continuous language area of German and variation corpora from outside that area (extraterritorial varieties, also called speech islands). The user community is as varied as the corpora, ranging from phoneticians, over dialectologists, conversation analysts, ethnomethodologists to historical linguists.

For this study we took data from three corpora: the Pfeffer-Corpus (PF), FOLK and the BETV-Corpus[2], which we describe in more detail below. For the experiment on regionality, we chose PF. For analysing typical ASR error types qualitatively, we chose data from BETV.

The Pfeffer-Corpus (AGD-PF, 1961) is suitable for testing ASR-regionality as it was recorded with high-quality technical equipment, the speech is colloquial, from city-like regions (including major cities in Germany, Switzerland and Austria, like Hamburg, Berlin, Frankfurt, Bremen, Bern, Zürich, Innsbruck, Vienna, etc., cf. Figure 2). The speakers are interviewed, however the interviewer was instructed to let mainly the target person speak. The varietal span is not as extreme as e.g. in the Zwirner-Corpus (AGD-ZW, 1956), where speakers were especially recruited from small towns, who hadn't moved about much and who had a relatively low socio-economic status.

From the Pfeffer-Corpus, we chose from each recording-place one female and one male target speaker randomly, making up a collection of altogether 112 recordings of durations between 7 and 16 minutes summing up to 21 hours and 22 minutes. From the metadata we extracted the geocodes (latitude and longitude coordinates).

2.2. Automatic Speech Recognition

We had access through a REST-API to the Audio-Mining System from the Fraunhofer IAIS (Schmidt et al., 2016)[3]. The model that was employed at the stage of processing is described in (Gref et al., 2019): The acoustic model is trained on 1000h broadcast data (Stadtschnitzer et al., 2014) that is 3-fold noise & reverberation augmented and 3-fold speed perturbed (corresponding altogether to 9000h). The speed perturbation is based on work by Ko et al. (2015). The language model was trained on 1.6 billion sentences/tokens from which a pronunciation dictionary containing 2 Million entries was automatically generated using grapheme-2-phoneme conversion following (Bisani and Ney, 2008) and using the German pronunciation dictionary *Phonolex*[4] from the Bavarian Archive for Speech Signals (BAS).

The ASR system provides the metric "asrQuality" that corresponds to a self-estimated confidence of how good the resulting transcription might be, based on the average recognition confidences across all word tokens (the arithmetic mean). The confidence per word is based on the recalculation of the probabilities in the search lattice, a graph that is unfolded by the acoustic model and the language model, i.e. the path through the lattice with the highest probability. The values of asrQuality may vary between 0 and 100. Additionally, we took the orthographic transcripts as references to the ASR-hypotheses and calculated the Word-Error-Rate (WER).

We tested the correspondence between the two metrics 'asrQuality' and 'WER' based on the PF-data in order to check if we can assume such a correspondence – in future, for ASR-outputs for which we do not have any reference transcripts available. The asrQuality and WER were negatively correlated, $r(108) = -.89, p < .001$ as shown in Figure 1. We also checked manually the outliers with WER=83.5%, Q=88 (male speaker in BERN, CH); WER=51.45%, Q=90; WER=58.58%, Q=84 (both male and female speakers from Cottbus, DE). It turned out that the alignment of text and audio of the underlying reference transcripts were bad. This had an effect on the calculation of the WER, but not on the asrQuality.

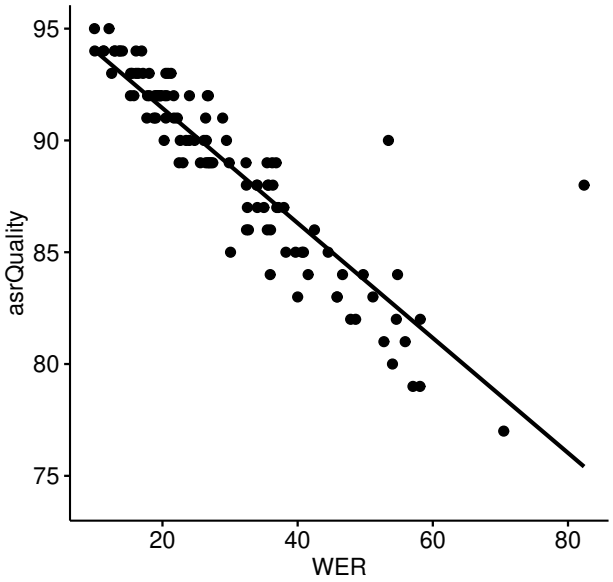

Figure 1: Scatterplot of WER and asrQuality with regression line.

It is also quite reasonable to assume that there must be a lower score in word recognition if the acoustic + language model produce a lower probability and vice versa. We cannot however assume that there would be a direct correspondence between WER and asrQuality, i.e. that an asrQuality score of e.g. 85 would correspond to a WER of 15%. The tendency is that the WER is even higher than the inverse asrQuality, as the acoustics (+ language model) might still be relatively high, even if the most probable word is not correct (false positive FP), which is sometimes the case for homonymes or similarly sounding words. This does not seem to be balanced out by false negatives (FN), where a correct word could be found even when a relatively low confidence is observed. This can be the case when the acoustic probability of the current word deviates from the acoustics of the word in the training. A last constellation is true negatives (TN), where incorrectly recognized words are accompanied by a relatively low confidence, which might be the case for e.g. out-of-vocabulary errors (OOV), where none of the words in the vocabulary

[2]The two recordings of BETV that we analysed here are also part of FOLK.

[3]cf. https://www.iais.fraunhofer.de/en/business-areas/content-technologies-and-services.html

[4]Phonolex: https://www.phonetik.uni-muenchen.de/Bas/BasPHONOLEXeng.html

can be confidently matched with the acoustics (+ language model) and the system has no other option than taking the best match, which is necessarily wrong. All these constellations are summarized in Table 1 and show that a simple measure such as WER can only represent a part of the complexity of recognition errors and their causes.

TP (correct + high conf.)	FP (in-corr + high conf.)
FN (correct + low conf.)	TN (in-corr + low conf.)

Table 1: Possible constellations of correctly vs. incorrectly recognized words (corresponding to high vs. low WER) and high vs. low confidence (corresponding to high/low "asrQuality").

The results of the recognition task, i.e. the relationship between asrQuality and regionality, are presented in Section 3.1.

2.3. Benchmarking and Qualitative Analysis

For the aforementioned transcribed test data, we had the BenchmarkViewer available, an additional tool provided by the Fraunhofer IAIS to calculate the WER (among other measures) based on the hypothesis and reference transcripts. We observed the asrQuality–WER relationship (the higher the asrQuality, the lower the WER) as shown in Table 2. The events "BUND_01" and "BUND_02" are political debates in the German parliament (Bundestag). The events "PODI_01, 02, 03" are podium discussions. Both types of events are currently built up for "FOLK" (the research and teaching corpus of spoken German (FOLK, 2019)). The events from the "BETV" corpus are political talk-shows from the regional TV-channel in the German-speaking area in Belgium (AGD-BETV, 2019). The events from the "PF" corpus are interviews from the 1960s. (AGD-PF, 1961).

Corpus	Event(s)	dur	Q	WER	ci-WER
FOLK	BUND_01	1h58	94	14.30	13.33
FOLK	BUND_02	1h32	94	18.71	17.41
FOLK	PODI_01	1h21	93	17.35	16.26
FOLK	PODI_02	0h59	91	24.98	23.88
FOLK	PODI_03	1h03	92	22.81	21.83
BETV	E_00001	0h58	92	19.87	18.70
BETV	E_00002	0h56	92	22.82	21.85
PF	112 events	21h22	89	31.24	30.02

Table 2: relationship between asrQuality (Q) and Word-Error-Rate (WER) in [%] or case-insensitive WER (ci-WER) for recordings from the corpora FOLK and BETV (averaged values for the events from corpus PF, including the outliers mentioned regarding Figure 1).

Table 2 also shows that the overall WER of the recognizer between 14.3% and 24.98% for the contemporary data is relatively good. For the historical data from the Pfeffer-Corpus it is slightly worse with 31.24%.
With the help of the BenchmarkViewer we extracted miss-recognized words and analysed the error-sources by comparing phonetic realisations with the phonetics of the word in the reference transcript and the phonetics of the word that was hypothesized by the system, cf. Section 3.2.

3. Results

3.1. Regionality

In order to approximate regionality, we refer to the latitude and longitude of the place where the individual speaker grew up and developed his/her way of speaking. Figure 2 illustrates this regional distribution[5]. The ASR quality for each recording is illustrated on a blue-yellow scale with blue indicating high ASR-Quality and yellow indicating low ASR-Quality.

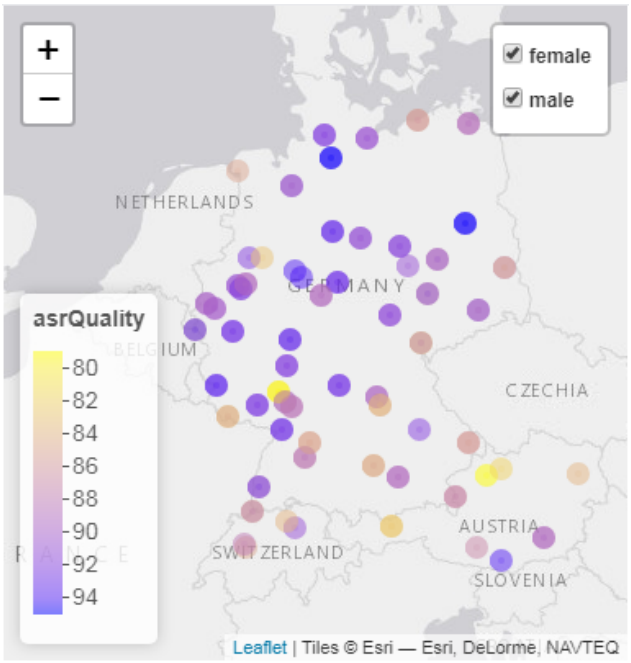

Figure 2: Distribution of speakers' variety and ASR-Quality.

There is no noticeable difference in ASR-Quality for female vs. male speakers as shown in Figure 3.

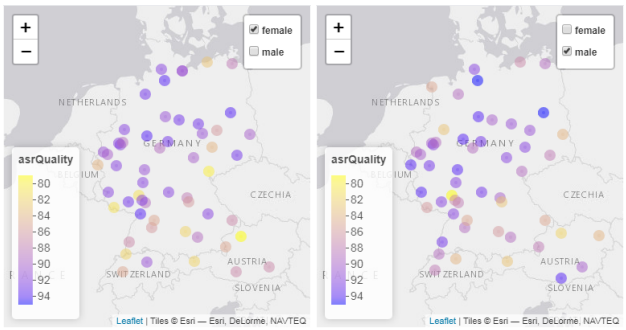

Figure 3: Regional Distribution of ASR-Quality for female (left) and male speakers (right).

Regarding our hypothesis that there is a regional influence

[5]The map was plotted with the function shiny (Chang et al., 2019) in R (R Core Team, 2019) and the open-source JavaScript library Leaflet https://leafletjs.com/.

on the ASR-System, the ASR-Quality seems to decline from the north to the south.
Having the geographic directions available in the metadata in form of latitude and longitude coordinates, we fed them into a correlation analysis. The North-South dimension is indicated by the latitude. A Spearman correlation[6] shows a positive coefficient ($rho = 0.371$) indicating that an increase in latitude corresponds with an increase in asrQuality. The correlation is significant ($p < 0.0001$), indicating that the ASR-System performs better with data from the north than data from the south (cf. scatter plot in Figure 4). The longitude did not show a relationship with asrQuality ($rho = -0.162$, $p = 0.09$).

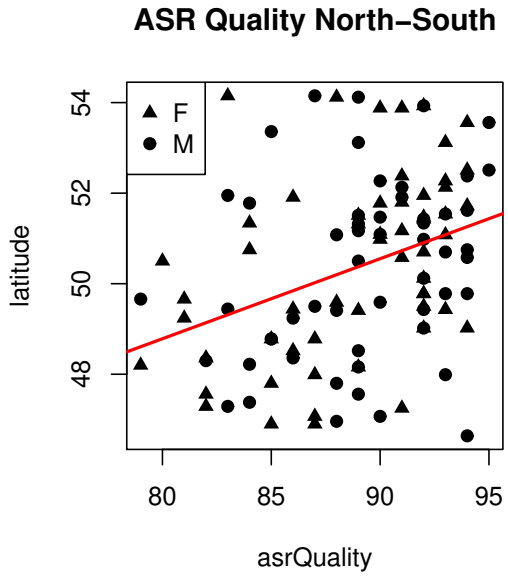

Figure 4: Scatterplot of latitude vs. asrQuality ("F" = female, "M" = male) with the regression line from a linear model.

3.2. Typical ASR Error Types

As mentioned above, the underlying data for this section stem from the BETV-Corpus (cf. the two recordings from Table 2), that are recordings from the two towns Amel and Burg Reuland in the German speaking area in east Belgium. The regional variety is at the intersection of the varieties Ripuarian and Mosel-Franconian, cf. the dialectal map by Wiesinger (1983, p.836). The participants of the BETV-Corpus are local politicians and TV-moderators who speak the "regional standard" of that area.
Apart from errors due to overlapping talk, homonymes, capitalisation or the complex German orthography (compounding), some of the error types we observed are also attributable to the specific regional variety.
Two types of recognition errors seem to be systematically frequent. Of course, there are common OOV errors concerning e.g. local names (Büllingen, Bütgenbach, Elsenborn, Kaiserbaracke) or current topics of the discussion (Entschlackung, makabere Unterstellung, Asphaltwerk). A second type of error is caused by consistent linguistic variation that is typical for that specific region as described by Münch (1904). For example, the word "nur" (merely) was recognized as "noch" (yet). This is due to the regional pronunciation [nuːχ] vs. the standard [nuːʁ]. The [χ] was therefore attributed to another word that contained a [x] following a vowel in the standard variety of German, in this case the "noch", cf. Münch (1904, §40 p.35f.). Another example is [waχtən] "warten" (wait) being recognized as "wachten" (guard – simple past) or [saːɣən] "sagen" (say) being recognized as "sahen" (saw – simple past of "see"), cf. Münch (1904, §111 p.92f.). A selection of these and other examples are shown in Table 3. The table reads as in the caption.
Table 3 shows that the recognizer has problems with specific phenomena in Ripuarian, where the orthographic "er", "ör", "or", "ur" are mostly diphthongized or vocalized to a-schwa in "standard German" while in Ripuarian, it seems that these phenomena are produced with a uvular – or at least velar – voiceless fricative. The orthographic "g" as in "sagen" is in standard German a voiced velar plosive, while in Ripuarian it is almost disappearing or reduced to an intervocalic approximant.

4. Discussion

It is common sense that an ASR System can only recognize data (speech) that it was either trained on or adapted to. Therefore, the training data should contain all types of data, covering quiet and noisy environments, young and old speakers, male and female, in different interactional settings, etc. Here we looked at one aspect: regionality by either keeping the other parameters constant or by covering them throughout.
The relationship between latitude and asrQuailty shows that there is either a bias in the training-data, e.g. more training data from the north than the south of the German speaking area. Or there is a tendency to speak more northern when being on broadcast – however this might also have been a similar setting for the speakers of the Pfeffer-Corpus being interviewed, who don't seem to have that tendency.
The qualitative analysis of one of the German varieties (Ripuarian) shows that despite the relatively high asrQuality (92), there are still some region-related errors that could be tackled by either employing training-data from that specific area, or by introducing other pronunciation variants into the pronunciation dictionary – or another mechanism – of the ASR framework or architecture.

5. Conclusions

Testing an Automatic Speech Recognition System with data that is enhanced with metadata containing information on the regional background of the speakers allowed us to reveal geographic gaps, where a system performs significantly less well in one region than in another.
From the perspective of a data center being interested in creating transcripts for archived audio recordings, it is at least to be expected that automatically created transcripts can be less trusted for some areas and it would be necessary to invest more time/money for correcting them.
A direct outcome of this pilot study is that we made the recognition results, i.e. transcripts, for the entire BETV-

[6] We chose a Spearman correlation as we cannot expect a linear relationship between latitude and asrQuality.

Reference	R-canonical	Hypothesis	H-canonical	observed	R-corr.	H-incorr.	oth. incorr.	total
dort	dɔɐt	doch	dɔx	dɔχt	38	18	5	61
nur	nuːɐ	noch/nun	nɔx/nuːn	nuːχ	30	4/5	4	43
noch	nɔx	nur	nuːɐ	nɔ	73	1	10	84
sagen	saːgən	sahen	saːhən	saːɣən	58	3	5	66
Dörfer	dœɐfə	doch vor	dɔx foːɒ	dœχfɔ	2	1	1	4
Dorf	dɔɐf	doch	dɔx	dɔχf	4	1	0	5
Vorteile	foːɐtaɪlə	Bruchteile	brʊxtaɪlə	vɔχtaɪlə	1	1	0	2
gebucht	gəbuːxt	Geburt	gəbuːɐt	gəbuːt	0	1	0	1
Geburten	gəbuːɐtən	gebucht	gəbuːχt	gəbuːχtən	0	1	0	1
aber	aːbɐ	aber auch	aːbɐ aʊx	aːbəχ	101	5	19	125

Table 3: Examples of ASR-confusions due to variety-specific pronunciations. 'Reference' is the manually transcribed words, 'Hypothesis' is the automatically recognized words, 'observed' is the phonetic pronunciation, that was then either correctly recognized as the Reference ('R-correct') or incorrectly as the Hypothesis ('H-incorrect'). Sometimes it was confused with other incorrect words ('oth. incorr.').

Corpus (10 x 1h recordings) available to the research community with the latest release 2.13 of the Database for Spoken German (DGD, Schmidt (2014), http://dgd.ids-mannheim.de). This is now the first corpus of the AGD whose transcripts are created entirely automatically.

6. Limitations and Future Work

Sofar, the BenchmarkViewer calculates error rates, precision and recall, etc. on different parameters – words (based on insertions, deletions, substitutions), speaker-diarization, punctuation, etc.). An additional feature that we have in mind are the classification of different word error types, e.g. missing hesitation markers, compounds ("Asphalt Werk" vs. "Alphaltwerk"), overlaps, etc. Once they are classified, the remaining errors are most likely due to regional or speaker-related characteristics, which are useful to analyse for both developers of speech technology and for linguists. Extending such regionality testing to other available corpora, e.g. the Zwirner-Corpus as mentioned above, should make it possible to make finer grained evaluations. This can also be enhanced with more sophisticated analysis methods such as geographical clustering. We also haven't used all the available metadata yet (age, place of birth, language background of parents, etc.) that could be included in such analyses.

The data could also be used to train/adapt the ASR-System to these regional variants and for evaluating the new System. ASR developers would also need to think about the trade-off: up to what point does it make sense to train a system with all varieties of a language, and at what point does it make sense to split a system and create recognition for individual varieties of a pluricentric language.

From the data center perspective, we also need to consider which way we want to be going once correcting automatically derived transcripts is getting faster than transcribing from scratch, as this development needs to find a couterpart in software development from transcription-software to correction-software.

7. Acknowledgements

We would like to thank the Fraunhofer IAIS for providing API access to the Audio Mining system and the BenchmarkViewer, Sascha Wolfer and Sandra Hansen for advice on the statistical analysis, and Ralf Knöbl for informative exchanges on the Ripuarian variety.

8. Bibliographical References

Bisani, M. and Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J., (2019). *shiny: Web Application Framework for R*. R package version 1.4.0.

Gref, M., Köhler, J., and Leh, A. (2018). Improved transcription and indexing of oral history interviews for digital humanities research. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Gref, M., Schmidt, C., Behnke, S., and Köhler, J. (2019). Two-staged acoustic modeling adaption for robust speech recognition by the example of german oral history interviews. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 796–801. IEEE.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *16th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3586–3589.

Leh, A., Köhler, J., Gref, M., and Himmelmann, N. (2018). Speech analytics in research based on qualitative interviews. experiences from ka3. *VIEW Journal of European Television History and Culture*, 7(14).

Münch, F. (1904). *Grammatik der ripuarisch-fränkischen Mundart*. Friedrich Cohen, Bonn, Germany.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE Signal Processing Society, December.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *17th Annual Con-*

ference of the International Speech Communication Association (Interspeech), pages 2751–2755.

R Core Team, (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schmidt, C., Stadtschnitzer, M., and Köhler, J. (2016). The Fraunhofer IAIS audio mining system: Current state and future directions. In *12. ITG Symposium on Speech Communication*, pages 1–5.

Schmidt, T. (2014). The Database for Spoken German–DGD2. In *Ninth international conference on Language Resources and Evaluation (LREC)*, pages 1451–1457, Reykjavik, Iceland. European Language Resources Association (ELRA).

Stift, U.-M. and Schmidt, T. (2014). Mündliche korpora am IDS: vom deutschen spracharchiv zur datenbank für gesprochenes deutsch. In Melanie Steinle et al., editors, *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, pages 360–375. Institut für Deutsche Sprache, Mannheim, Germany.

Wiesinger, P. (1983). Die Einteilung der deutschen Dialekte. In Werner Besch, et al., editors, *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung*, pages 807–900. de Gruyter, Berlin, New York.

9. Language Resource References

AGD-BETV. (2019). *Belgische TV-Debatten*. Archive for Spoken German, distributed via the DGD, Database for Spoken German.

AGD-PF. (1961). *Deutsche Umgangssprachen: Pfeffer-Korpus*. Archive for Spoken German, distributed via the DGD, Database for Spoken German.

AGD-ZW. (1956). *Deutsche Mundarten: Zwirner-Korpus*. Archive for Spoken German, distributed via the DGD, Database for Spoken German.

FOLK. (2019). *Research and Teaching Corpus of Spoken German*. Archive for Spoken German, distributed via the DGD, Database for Spoken German.

Stadtschnitzer, M., Schwenninger, J., Stein, D., and Köhler, J. (2014). Exploiting the large-scale German broadcast corpus to boost the Fraunhofer IAIS speech recognition system. In *International Conference on Language Resources and Evaluation (LREC)*, pages 3887–3890.