INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Longitudinal Speaker Clustering and Verification Corpus with Code-Switching Frisian-Dutch Speech

Emre Yılmaz[1], *Jelske Dijkstra*[2], *Hans Van de Velde*[2], *Frederik Kampstra*[3], *Jouke Algra*[3], *Henk van den Heuvel*[1] *and David Van Leeuwen*[1]

[1]CLS/CLST, Radboud University, Nijmegen, Netherlands
[2]Fryske Akademy, Leeuwarden, Netherlands
[3]Omrop Fryslân, Leeuwarden, Netherlands

{e.yilmaz, h.vandenheuvel, d.vanleeuwen}@let.ru.nl, {jdijkstra, hvandevelde}@fryske-akademy.nl, {frederik.kampstra, jouke.algra}@omropfryslan.nl

Abstract

In this paper, we present a new longitudinal and bilingual broadcast database designed for speaker clustering and text-independent verification research. The broadcast data is extracted from the archives of Omrop Fryslân which is the regional broadcaster in the province of Fryslân, located in the north of the Netherlands. Two speaker verification tasks are provided in a standard enrollment-test setting with language consistent trials. The first task contains target trials from all speakers available appearing in at least two different programs, while the second task contains target trials from a subgroup of speakers appearing in programs recorded in multiple years. The second task is designed to investigate the effects of ageing on the accuracy of speaker verification systems. This database also contains unlabeled spoken segments from different radio programs for speaker clustering research. We provide the output of an existing speaker diarization system for baseline verification experiments. Finally, we present the baseline speaker verification results using the Kaldi GMM- and DNN-UBM speaker verification system. This database will be an extension to the recently presented open source Frisian data collection and it is publicly available for research purposes.

Index Terms: Speaker clustering, speaker diarization, speaker verification, ageing effects, bilingual data

1. Introduction

Speaker clustering and verification tasks have been relevant for various biometric and forensic applications [1–4]. One desired application is to use speech utterances for authentication of secure actions performed via automated systems. Moreover, automatically identifying speakers with their speech in different recordings such as telephone conversation, radio programs, meetings facilitates the manual labeling work.

Several databases have been prepared by NIST for the global evaluation of text-independent speaker verification system between 1996–2016 [1]. The goal of these evaluations is to propose a unified framework for all researchers to test their techniques on conversational telephony speech. These conversations are in North American English, except the last evaluation organized in 2016 which contains recordings in multiple languages, namely Tagalog, Cantonese, Mandarin and Cebuano. Recently, some other databases which aim to offer different challenges than typical verification setting of NIST SREs have also been proposed [5–7].

Multilingual speaker verification systems which can operate on more than one language have been researched in the last two decades [8–11]. These systems have been tested on multilingual databases aimed for language identification experiments (e.g., NIST language identification development (LID) data [12], OGI multi-language telephone speech data [13]) or on small-sized bilingual databases designed for speaker identification experiments.

In this work, we present a new bilingual and longitudinal broadcast database designed for speaker clustering and verification experiments. This database is based on the FAME! speech corpus [14, 15]. This corpus contains manually annotated radio broadcasts with Frisian-Dutch code-switching (CS) speech. Frisian is mostly spoken in the province of Fryslân which is located in the north of the Netherlands. The native speakers of Frisian are Frisian-Dutch bilingual and often code-switch in daily conversations. To the best of our knowledge, this database is the first bilingual broadcast database designed for speaker clustering and verification research.

The presented database has been collected in the scope of the FAME! (Frisian Audio Mining Enterprise) Project. This project aims to build a spoken document retrieval system for the disclosure of the archives of Omrop Fryslân[2] (Frisian Broadcast) covering a large time span from 1950s to present and a wide variety of topics. For accurate document retrieval, one milestone in this project is the integration of speaker diarization and speaker verification system that can be applied to a large longitudinal data set. For this purpose, we have prepared the proposed database by reorganizing the annotated parts of the FAME! speech corpus in a standard enrollment-test setting with language- and gender-consistent trials. Language consistency in a code-switching scenario is only applied to monolingual segments, i.e., segments with code-switching are included in trials. Due to the longitudinal nature of the data, a separate set of trials is created with enrollment segments chosen with recordings dates far from the recording dates of the test utterances. The initial research investigating the effects of ageing on speaker verification systems are presented in [16–18]. With these attributes, the database offers researchers the possibility to test their speaker clustering and verification systems on bilingual and longitudinal speech data. Finally, a large amount of unlabeled data is added for training purposes. However, unlike the unlabeled data provided at NIST SRE 2016[3], each unlabeled

[1] https://www.nist.gov/itl/iad/mig/speaker-recognition

[2] Omrop Fryslân is the regional public broadcaster of the province of Fryslân. (http://www.omropfryslan.nl)

[3] https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016

http://dx.doi.org/10.21437/Interspeech.2017-301

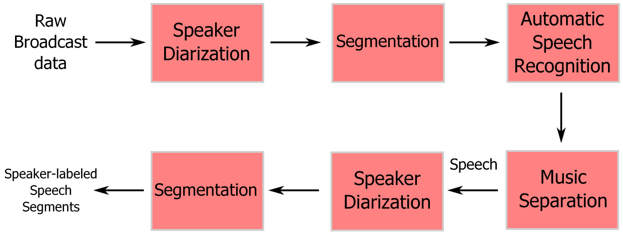

Figure 1: *Preprocessing the unlabeled broadcast data to extract speaker-labeled speech segments*

recording contains speech segments from a radio program with multiple speakers. Using this unlabeled in-domain data, more accurate speaker models can be obtained for target speaker verification tasks.

This paper is organized as follows. Section 2 describes the properties of the labeled Frisian-Dutch data used for the verification experiments. The organization of the database is described in Section 3 and the baseline verification results are presented in Section 4. Section 5 concludes the paper.

2. Frisian-Dutch Radio Broadcast Data

The bilingual FAME! speech database contains radio broadcasts in Frisian and Dutch. This bilingual data contains Frisian-only and Dutch-only utterances as well as mixed utterances with inter-sentential, intra-sentential and intra-word CS [19]. These recordings include language switching cases and speaker diversity, and have a large time span (1966–2015). The content of the recordings is very diverse, including radio programs about culture, history, literature, sports, nature, agriculture, politics, society and languages.

The radio broadcast recordings have been manually annotated and cross-checked by two bilingual native Frisian speakers. The annotation protocol designed for this CS data includes three kinds of information: the orthographic transcription containing the uttered words, speaker details such as the gender, dialect, name (if known) and spoken language information. The language switches are marked with the label of the switched language. There are 10 speakers who are labeled to speak in both languages. For further details, we refer the reader to [14].

The total duration of the manually annotated radio broadcasts sums up to 18.5 hours. The stereo audio data has a sampling frequency of 48 kHz and 16-bit resolution per sample. All data is subsampled to 16 kHz and reduced to single channel data. The available meta-information helped the annotators to identify these speakers and mark them either using their names or the same label (if the name is not known). Later, a manual check has been performed by the second author, who is also a bilingual native Frisian-Dutch speaker, to improve the quality of the speaker annotations with the help of Omrop Fryslân employees. There are 334 identified and 120 unidentified speakers in the FAME! speech database. 51 of the identified speaker appear at least in 2 different years in the database. These speakers are mostly program presenters and celebrities appearing multiple times in different recordings over years.

3. FAME! Speaker Verification Database

The annotations of the Frisian-Dutch radio broadcasts are used to extract the segments containing speech and each speech segment is classified based on its speaker, program name, recording year, and language. This database also contains unlabeled spoken segments from 363 different radio programs. These speech segments resemble the target utterances in the verification tasks and training the speaker verification system on this data is expected to provide the most accurate speaker models. Since each of the radio programs contain speech from more than one speaker, speaker clustering (diarization) is required before training the verification system. The output of an existing speaker diarization system is also included in the database for baseline verification experiments.

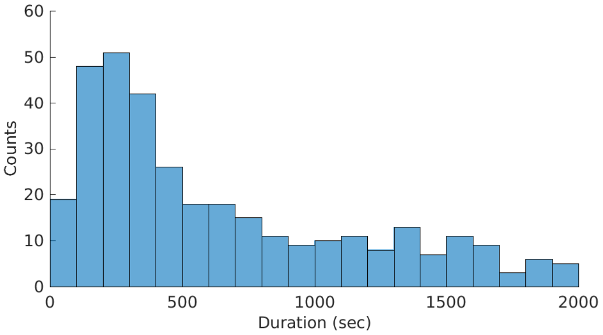

Figure 2: *Duration distribution of the unlabeled speech segments*

Two speaker verification tasks are provided in a standard enrollment-test setting with language consistent trials. The first task contains target trials from all speakers available appearing in at least two different programs, while the second task contains target trials from a subgroup of speakers appearing in programs recorded in multiple years. The second task is designed to investigate the effects of ageing on the accuracy of speaker verification systems. For each task, 3 trial lists are created with durations of 3, 10 and 30 seconds. The details of each component are presented in the following subsections.

3.1. Speaker Clustering Task

The first challenge in the proposed database is to label speakers in radio programs that are extracted from the same radio broadcast archive. Frisian is a low-resourced language with no available in-domain data to train a speaker verification system operating on broadcast data. Having a small subset of the radio archives annotated, there is a large amount of raw broadcast data that can be used to train a bilingual speaker verification system in an unsupervised setting. Therefore, we selected 150 hours of raw radio broadcast data and this data has been preprocessed using a speaker diarization system and an automatic speech recognition (ASR) system to extract the speech segments with no or mild background music. The block diagram of the preprocessing the raw broadcast data for automatic speaker labeling is given in Figure 1. Based on the speaker diarization output, long radio programs are segmented with a reasonable separation of music segments from speech segments. To identify the content of each segment, they are fed to an ASR system and a subset of these segments are chosen based on the number of words and average word length of the text hypothesized by the ASR. After removing the segments that are suspected to be music based on the ASR output, the speech segments in each program are automatically labeled with a speaker id by applying the same speaker diarization system. The total duration distribution of these segments for each program are given in Figure 2. These segments are used for learning a universal background model, a T matrix for i-vector extraction and a PLDA model in the experiments presented in Section 4.

Table 1: *Speaker verification task statistics*

	3 s			10 s			30 s		
	all	female	male	all	female	male	all	female	male
	Complete Database (described in Sec. 3.2.1)								
# of enroll. segments	7188	1952	5236	2094	625	1469	581	163	418
# of test segments	7842	2378	5464	2039	566	1473	594	169	425
# of enroll. speakers	245	66	179	222	62	160	162	44	118
# of test speakers	236	75	161	218	65	153	165	41	124
# of trials	19,763,834	2,902,328	16,861,506	1,481,416	217,755	1,263,661	120,180	16,395	103,785
# of target trials	198,315	104,605	93710	14,413	7675	6738	1138	594	544
# of target trials (in %)	1.0%	3.6%	0.6%	1.0%	3.5%	0.5%	0.9%	3.6%	0.5%
	Ageing Database (described in Sec. 3.2.2)								
# of enroll. segments	15,460	7802	7658	3760	1772	1988	847	409	438
# of test segments	14,506	4593	9913	4367	1298	3069	1243	336	907
# of enroll. speakers	46	15	31	42	13	29	32	10	22
# of test speakers	319	89	230	331	80	251	253	49	204
# of trials	2,054,354	793,413	1,260,941	248,575	76,712	171,863	28,651	8952	19,699
# of target trials	299,239	167,339	131,900	21,908	12,005	9903	1728	971	757
# of target trials (in %)	14.6%	21.1%	10.5%	8.8%	15.6%	5.8%	6.0%	10.8%	3.8%

3.2. Speaker Verification Tasks

3.2.1. Complete Database

This component of the database contains speaker verification trials from all speakers appearing in the FAME! Speech Corpus. The target trials contain speech from speakers appearing at least in two different programs implying that there are no target trials in which two segments from the same program are compared. Cross-gender and cross-language trials are also excluded. It is important to note that cross-linguality applies only to monolingual segments. Code-switching (bilingual) segments are actually included in enrollment and test data and they may be compared with mono- or bilingual segments.

During the preparation phase, we have firstly created a list of the speech segments from the FAME! Speech Corpus with the speaker name, program name, start and end time of the segment, spoken language and the year in which the program is broadcasted. These non-overlapping segments are later merged and trimmed to produce fixed length segments of 3, 10 and 30 seconds. The merging is performed in a way that the final segment will contain speech from the same speaker spoken in the same program and the same language.

The trials are created by grouping the segments that are extracted from different programs and randomly assigning them as enrollment and test data. The segments uttered by the speakers that appear only in one program are added to either enrollment or test data to be used in nontarget trials.

3.2.2. Data Controlled for Ageing Effects

The second verification setup contains enrollment data only from 51 speakers who appear in multiple years. The goal of this setup is to choose trials in a controlled manner to allow the measurement of ageing effects on speaker verification systems. The target trials are extracted as 2-combination of all available years for a speaker to maximize the number of possible trials. For each target segment, all segments from the same year spoken by another speaker are used as test segments of multiple nontarget trials. To increase the amount of trials, we allow a single year mismatch with 10-second segments and two-year mismatch with 30-second segments.

To analyse the effect of the ageing, three subgroups of trials are created based on the time difference in recording date between the enrollment and test segments. The first subgroup (1-3) contains the trials with difference of 1 to 3 years. The second and third subgroups include the trials with difference of 4 to 10 (4-10) and more than 10 years (>10) respectively. The boundaries are chosen to yield comparable amounts of trials for each subgroup.

4. Baseline Verification Experiments and Results

4.1. Experimental setup

The unlabeled radio programs is preprocessed to obtain automatically created speaker labels as illustrated in Figure 1. For this purpose, the IDIAP speaker diarization system has been used [20]. The speech-music classification is achieved based on the ASR output which has been developed last year in the scope of the FAME! Project [21, 22]. Some statistics about the verification tasks are given in Table 1. For the verification tasks, no development data is provided due to the limited amount of available manually annotated speech data.

We perform speaker verification experiments using the KALDI speaker verification systems described in [23]. Being familiar with the ASR toolkit of KALDI [24], we opt for the KALDI speaker verification system due to the high compatibility with the resources generated for the ASR system. Two types of speaker verification systems are available in this toolkit, namely a Gaussian mixture model-universal background model (GMM-UBM) and a deep neural network-UBM (DNN-UBM). We detailed these recognizers in the following paragraphs.

The GMM-UBM speaker recognizer extracts 20 MFCCs with a frame shift of 10 ms and a frame length of 25 ms with deltas and delta-deltas. Mean normalization is applied over a 3 second window. For the GMM-UBM, a diagonal covariance matrix is trained initially by applying 4 EM iterations followed by another 4 iterations with a full-covariance matrix. The i-vector extractor is obtained after 5 EM iteration on the training data and it generates 600-dimensional i-vectors. Finally, the i-vector mean subtraction and length normalization is applied

Table 2: *Speaker verification results in EER (%) and weighted EER (%)*

	3 s			10 s			30 s		
	pooled	female	male	pooled	female	male	pooled	female	male
	Complete Database (EER (%))								
GMM-UBM	21.4	25.3	16.6	14.0	17.7	9.7	10.5	13.6	6.8
DNN-UBM (matched DNN)	16.6	20.1	13.0	10.1	13.0	7.4	7.4	10.1	4.8
DNN-UBM (mismatched DNN)	16.6	19.8	13.1	10.0	13.1	6.9	7.2	9.6	4.6
	Ageing Database (EER (%))								
GMM-UBM	22.4	24.4	18.1	14.1	15.4	10.4	10.1	10.5	6.9
DNN-UBM (matched DNN)	17.3	19.3	13.9	10.3	12.0	7.8	7.4	8.2	4.8
DNN-UBM (mismatched DNN)	17.2	19.0	14.0	10.3	11.9	7.4	7.1	7.9	4.0
	Ageing Database (weighted EER (%))								
GMM-UBM	20.4	22.3	21.3	14.9	16.4	14.4	10.0	11.1	9.5
DNN-UBM (matched DNN)	18.6	20.4	18.8	12.9	13.0	13.0	8.3	8.9	8.1
DNN-UBM (mismatched DNN)	18.4	19.5	18.9	13.2	13.0	13.3	8.1	8.4	7.8

before calculating the PLDA scores.

For the DNN-UBM, we have trained a time-delay deep neural network [23, 25, 26] to extract sufficient statistics required for i-vector extractor training. For this purpose, we have used both the FAME! Speech Corpus which is the only manually annotated Frisian broadcast database and the CGN corpus [27] which only contains Dutch speech data. This FAME! Speech Corpus designed for ASR evaluation on code-switching speech data is also publicly available[4] and the Frisian KALDI recognition scripts can be found at the main KALDI repository[5]. The broadcast components of the CGN database containing 107 hours of speech have been used for training the DNN. The former DNN is trained on the same speech segments included in the verification task. Due to this overlap, we refer to this DNN as 'matched'. The latter DNN is referred to as 'mismatched' as it is trained on an unseen database with only Dutch speech. The DNN models are trained on high-resolution MFCC features with 40-dimensional MFCC features with a frame length of 25 ms. Cepstral mean normalization is performed over a window of 6 seconds. The DNN has 6 hidden layers with p-norm input dimension of 2500 and output dimension of 250. The splicing details are given in [23].

The equal error rate (EER) and the weighted EER [28] are used to quantify the quality the of the speaker verification system. The former metric is used for the complete database, while the latter is used for the ageing database to balance the influence of the speakers on the final performance figures (cf. [17]).

4.2. Results

The speaker verification results are presented in Table 2. The results of the complete and ageing database are presented in the upper and lower panel respectively. The EERs presented on 3-second segments are around 20%, while using 30-second segments approximately halves the EERs. There is a large performance gap between genders and female voices appear to be significantly more challenging than male voices. The DNN-UBM system with the matched DNN provides the best results. This is expected as DNN targets referring to tristate phones are trained on the same speech with the verification task. The mismatch DNN performs similar to the DNN-UBM system using a DNN trained the FAME! speech corpus.

In Figure 3, the EERs obtained on different ageing categories have been presented. The verification accuracy drops as the difference between the recording year increases for 3- and 10-second segments which is consistent with the previous research [18]. For 30-second segments, the weighted EER obtained on age categories 1-3 is 4.1% lower than 4-10, while the results obtained on more than 10 years and 4-10 are similar with a weighted EER of 10.5%.

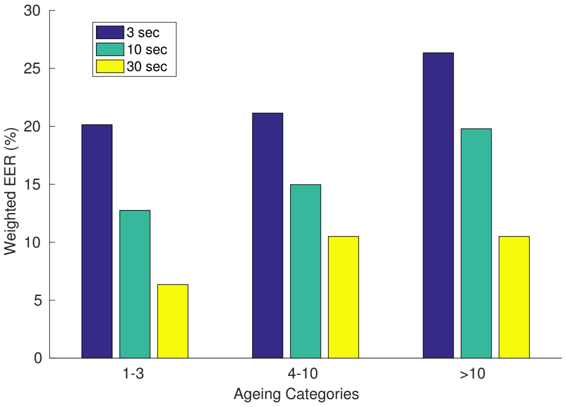

Figure 3: *GMM-UBM results on different ageing groups*

5. Conclusion

In this work, we have presented a new longitudinal and bilingual database for speaker clustering and verification research. The Frisian-Dutch radio broadcast data is extracted from the archives of the local broadcaster and a small subset of this data is manually annotated with orthographic transcription and speaker information. This component is designed to perform two speaker verification experiments, one using all available data and the other controlled for ageing effects on speaker verification. Moreover, a large amount of unlabeled data is also provided. This part of the data is expected to be used for training a speaker verification system after applying speaker clustering/diarization. Considering the longitudinal character, this database is going to enable the research of speaker tracking and diarization over a large time period and speaker aging effects on speaker recognition systems. The database and recognition scripts will be publicly available for research purposes.

6. Acknowledgements

This research is funded by the NWO Project 314-99-119 (Frisian Audio Mining Enterprise).

[4] http://www.ru.nl/clst/datasets/

[5] https://github.com/kaldi-asr/kaldi/tree/master/egs/fame

7. References

[1] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. ICASSP*, vol. 4, May 2002, pp. 4072–4075.

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.*, vol. 2004, pp. 430–451, Jan. 2004.

[3] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J. F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 95–103, March 2009.

[4] N. Singh, R. A. Khan, and R. Shree, "Applications of speaker recognition," *Procedia Engineering*, vol. 38, pp. 3122 – 3126, 2012.

[5] A. Lawson, A. Stauffer, E. Cupples, S. Wenndt, W. Bray, and J. Grieco, "The multi-session audio research project (MARP) corpus: goals, design and initial findings," in *Proc. INTERSPEECH*, 2009, pp. 1811–1814.

[6] F. Kelly, N. Brümmer, and N. Harte, "Eigenageing compensation for speaker verification," in *Proc. INTERSPEECH*, 2013, pp. 1624–1628.

[7] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The 2016 speakers in the wild speaker recognition evaluation," in *Proc. INTERSPEECH*, Sept. 2016, pp. 823–827.

[8] R. Auckenthaler, M. J. Carey, and J. S. D. Mason, "Language dependency in text-independent speaker verification," in *Proc. ICASSP*, vol. 1, 2001, pp. 441–444 vol.1.

[9] B. Ma and H. Meng, "English-Chinese bilingual text-independent speaker verification," in *Proc. ICASSP*, vol. 5, May 2004, pp. 293–296.

[10] N. T. Kleynhans and E. Barnard, "Language dependence in multilingual speaker verification," in *Sixteenth Annual Symposium of the Pattern Recognition Association of South Africa PRASA*, Nov. 2005.

[11] I. Luengo, E. Navas, I. Sainz, I. Saratxaga, J. Sanchez, I. Odriozola, and I. Hernaez, "Text independent speaker identification in multilingual environments," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008.

[12] "CALLFRIEND American English-Non-Southern Dialect LDC96S46," 1996, Philadelphia: Linguistic Data Consortium.

[13] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proceedings of the International Conference on Spoken Language Processing*, Oct. 1992.

[14] E. Yılmaz, M. Andringa, S. Kingma, F. Van der Kuip, H. Van de Velde, F. Kampstra, J. Algra, H. Van den Heuvel, and D. Van Leeuwen, "A longitudinal bilingual Frisian-Dutch radio broadcast database designed for code-switching research," in *Proc. LREC*, 2016, pp. 4666–4669.

[15] E. Yılmaz, H. Van den Heuvel, J. Dijkstra, H. Van de Velde, F. Kampstra, J. Algra, and D. Van Leeuwen, "Open source speech and language resources for Frisian," in *Proc. INTERSPEECH*, San Francisco, CA, USA, Sept. 2016, pp. 1536–1540.

[16] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification with long-term ageing data," in *2012 5th IAPR International Conference on Biometrics (ICB)*, 2012, pp. 478–483.

[17] G. Doddington, "The effect of target/non-target age difference on speaker recognition performance," in *Proc. Odyssey 2012: The Speaker and Language Recognition Workshop*, Singapore, June 2012.

[18] F. Kelly and J. H. L. Hansen, "Score-aging calibration for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2414–2424, Dec 2016.

[19] C. Myers-Scotton, "Codeswitching with English: types of switching, types of communities," *World Englishes*, vol. 8, no. 3, pp. 333–346, 1989.

[20] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1382–1393, Sept 2009.

[21] E. Yılmaz, H. Van den Heuvel, and D. A. Van Leeuwen, "Investigating bilingual deep neural networks for automatic speech recognition of code-switching Frisian speech," in *Proc. Workshop on Spoken Language Technology for Under-resourced Languages (SLTU)*, May 2016, pp. 159–166.

[22] E. Yılmaz, H. van den Heuvel, and D. van Leeuwen, "Code-switching detection using multilingual DNNS," in *IEEE Spoken Language Technology Workshop (SLT)*, Dec 2016, pp. 610–616.

[23] D. Snyder, D. G. Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 92–97.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, Dec. 2011.

[25] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. ICASSP*, May 2014, pp. 1695–1699.

[26] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. INTERSPEECH*, 2015, pp. 3214–3218.

[27] N. Oostdijk, "The spoken Dutch corpus: Overview and first evaluation," in *Proc. LREC*, 2000, pp. 886–894.

[28] D. A. van Leeuwen, "Overall performance metrics for multi-condition speaker recognition evaluations," in *Proc. INTERSPEECH*, Brighton, UK, Sept. 2009, pp. 908–911.