LANGUAGE DIARIZATION FOR SEMI-SUPERVISED BILINGUAL ACOUSTIC MODEL TRAINING

Emre Yılmaz[1,2]*, Mitchell McLaren*[2]*, Henk van den Heuvel*[1] *and David A. van Leeuwen*[1]

[1] CLS/CLST, Radboud University, Nijmegen, Netherlands
[2] Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA

ABSTRACT

In this paper, we investigate several automatic transcription schemes for using raw bilingual broadcast news data in semi-supervised bilingual acoustic model training. Specifically, we compare the transcription quality provided by a bilingual ASR system with another system performing language diarization at the front-end followed by two monolingual ASR systems chosen based on the assigned language label. Our research focuses on the Frisian-Dutch code-switching (CS) speech that is extracted from the archives of a local radio broadcaster. Using 11 hours of manually transcribed Frisian speech as a reference, we aim to increase the amount of available training data by using these automatic transcription techniques. By merging the manually and automatically transcribed data, we learn bilingual acoustic models and run ASR experiments on the development and test data of the FAME! speech corpus to quantify the quality of the automatic transcriptions. Using these acoustic models, we present speech recognition and CS detection accuracies. The results demonstrate that applying language diarization to the raw speech data to enable using the monolingual resources improves the automatic transcription quality compared to a baseline system using a bilingual ASR system.

Index Terms— Language diarization, code-switching, bilingual acoustic modeling, semi-supervised training, Frisian language

1. INTRODUCTION

Spontaneous language switches in a single conversation, also known as code-switching (CS), is prominent in multilingual societies in which minority languages are influenced by the majority language or majority languages are influenced by *lingua francas*, such as English and French. West Frisian (Frisian henceforth) is a regional language spoken in the northern provinces of the Netherlands with approximately half a million bilingual speakers. These speakers switch between the Frisian and Dutch languages in daily conversations. In the FAME! Project, the influence of this language alteration on modern ASR systems is explored with the objective of building a robust recognizer that can handle this phenomenon. The main focus has been developing robust acoustic models operating on bilingual speech delving into the automatic speech recognition and CS detection aspects [1].

The impact of CS and other kinds of language switches on speech-to-text systems has recently received research interest, resulting in several robust acoustic modeling [2–8] and language modeling [9–11] approaches for CS speech. Given that CS involves more than one language, it is foreseeable that automatic language identification (LID) could assist in the automatic speech recognition (ASR) of CS speech [12–15]. One fundamental approach is assigning language labels in advance with the help of a LID system and performing recognition of each language separately using a monolingual ASR system at the back-end. These systems may suffer from error propagation between the language identification front-end and ASR back-end, since language identification is still a challenging problem especially in the case of intra-sentence CS and closely related languages.

To cope with this problem, all-in-one ASR approaches, which do not directly incorporate a language identification system, have also been proposed [3, 6, 8]. In our previous work, we compared such an all-in-one system with another system using ground-truth language labels. Ground truth labels resulted in minor improvements in bilingual recognition performance (with a WER of 33.9% compared to 34.7% for the best performing bilingual ASR) at a cost of additional computation [1].

To date, our research has focused on the under-resourced Frisian language using semi-supervised acoustic model training techniques that can effectively leverage the little amount of manually transcribed Frisian data for training a bilingual ASR system. In our recent work [16], we used a multilingually trained ASR system to automatically transcribe raw broadcast data. Later, a subset of this data with *reliable* automatic transcriptions was combined with the manually annotated data and used for retraining (fine-tuning) the multilingually trained models aiming to improve recognition performance due to the considerable increase in the amount of training data. For generating the automatic transcriptions, we also applied lattice rescoring using bilingual language models to examine their impact on the quality of the automatic transcription. This type of semi-supervised acoustic model training has been researched intensively and various training strategies and data selection criteria have been proposed [17–22].

In this paper, we investigate the impact of using a language diarization (LD) system during this automatic transcription procedure. To the best of our knowledge, using LD on raw speech data in a semi-supervised bilingual acoustic modeling training setting has never been investigated. Using LD is expected to provide improvements in the acoustic modeling quality in this scenario for the following reasons. First, the LID errors on the raw broadcast recordings should be reduced due to the significant increase in recording durations which provides better context for the diarization process. Second, the marginal gains reported in [1] can indirectly improve the acoustic modeling quality, i.e., due to more accurate automatic transcription. Finally, the contribution of the lattice rescoring for improved automatic transcriptions is expected to be higher using monolingual LMs compared to using a bilingual LM as done in [16]. We delve into each of these aspects in the rest of the paper.

This work has been performed during a research visit to SRI International. This visit has been funded by the NWO Project 314-99-119 (Frisian Audio Mining Enterprise) and SRI International.

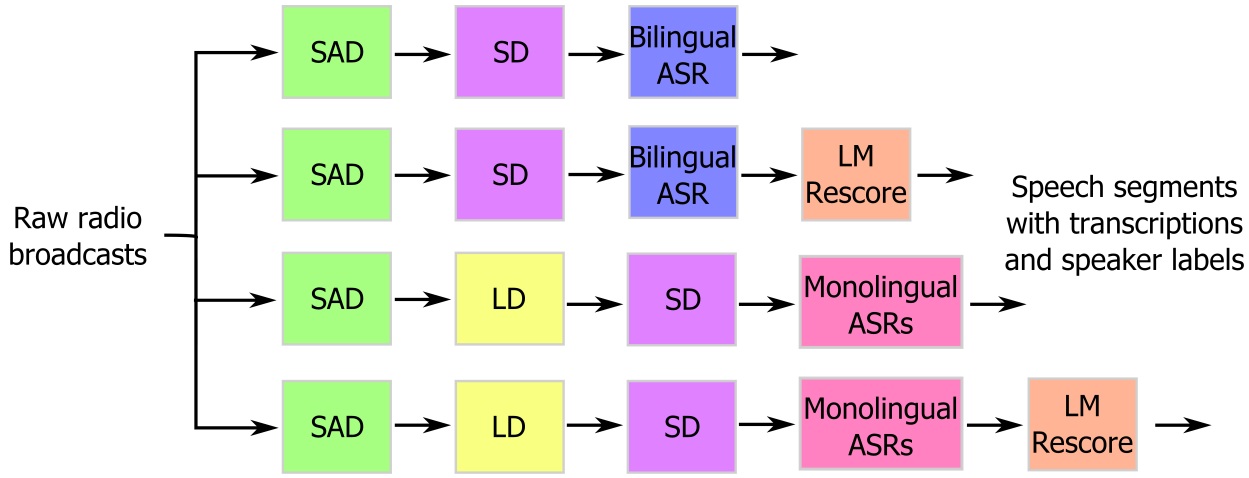

Fig. 1. Overview of the automatic transcription systems

This paper is organized as follows. Section 2 introduces the demographics and the linguistic properties of the Frisian language and summarizes the Frisian-Dutch radio broadcast database that was collected for CS and longitudinal speech research. Section 3 details the proposed semi-supervised bilingual acoustic model training approach with a language diarization system. The experimental setup is described in Section 4 and the recognition results are presented in Section 5.

2. THE FRISIAN LANGUAGE AND FRISIAN-DUTCH RADIO BROADCAST DATABASE

Frisian belongs to the North Sea Germanic language group, which is a subdivision of the West Germanic languages. Linguistically, three Frisian languages exist: (1) West Frisian, spoken in the province of Fryslân in the Netherlands; (2) East Frisian, spoken in Saterland in Lower Saxony in Germany; and (3) North Frisian, spoken in the northwest of Germany, near the Danish border. Historically, Frisian shows many parallels with Old English. However, today the Frisian language is under growing influence of the Dutch language due to long-lasting and intense language contact. Frisian has about half a million speakers. A recent study shows that about 55% of all inhabitants of Fryslân speak Frisian as their first language, which is about 330 000 people [23]. All speakers of Frisian are at least bilingual, since Dutch is the main language used in education in Fryslân.

The bilingual FAME! speech database, which has been collected in the context of the *Frisian Audio Mining Enterprise* Project, contains radio broadcasts in Frisian and Dutch. The FAME! project aims to build a spoken document retrieval system operating on the bilingual archive of the regional public broadcaster Omrop Fryslân (Frisian Broadcast Organization). This bilingual data contains Frisian-only and Dutch-only utterances as well as mixed utterances with inter-sentential, intra-sentential and intra-word CS [24]. To design an ASR system that can handle the language switches, a representative subset of recordings has been extracted from this radio broadcast archive. 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: (1) the orthographic transcription containing the uttered words; (2) speaker details such as the gender, dialect and name (if known); and (3) spoken language information. The language switches are marked with the label of the switched language. For further details, we refer the reader to [25].

Two kinds of language switches are observed in the broadcast data in the absence of segmentation information. First, a speaker may switch language in a conversation (*within-speaker switches*). Secondly, a speaker may be followed by another one speaking in the other language. For example, the presenter may narrate an interview in Frisian, while several excerpts of a Dutch-speaking interviewee are presented (*between-speaker switches*). Both type of switches pose a challenge to ASR systems and must be handled carefully.

3. AUTOMATIC TRANSCRIPTION OF RAW BROADCAST DATA

The automatic transcription schemes compared in this work are illustrated in Figure 1. For all systems, speech segments are extracted from a large amount of raw broadcast data using a robust speech activity detection system [26]. The following steps differ for each system and each component is detailed in the following sections.

3.1. Speaker Diarization

The speech segments are further labeled with speaker ids by using a speaker diarization (SD) system. The speaker diarization system used in these experiments attempts to cluster speaker voices in a recording such that each unique voice is assigned to a single cluster and each cluster only has the voice of one speaker. The intent of diarization in the current context is to aid in speaker-adaptive training. Errors from diarization (i.e, the allocation of speech to the wrong cluster) are expected to have limited impact on ASR since the errors will likely be due to similar sounding speakers.

3.2. Bilingual ASR

Baseline automatic transcription is achieved by using a multilingually trained system that is also used in [1]. Multilingual data from

closely related high-resourced languages, i.e., Dutch and English, is used for training deep neural network (DNN)-based acoustic models to obtain more robust acoustic models against the language switches between the under-resourced Frisian language and Dutch. The multilingual DNN training scheme resembles prior work [27–29] and is achieved in two steps. Firstly, the English and Dutch data are used together with the Frisian data in the initial multilingual training step to obtain more accurate shared hidden layers. After training the shared hidden layers, the softmax layer obtained during the initial training phase is replaced with one that is specific to the target recognition task. In the second step, the complete DNN is retrained bilingually (on Frisian and Dutch) to fine-tune the DNNs for the target CS Frisian and Dutch speech.

Each speech segment is labeled with a speaker id hypothesized by the speaker diarization system. These speaker labels are useful for applying speaker adaptive training using speaker-adapted speech features. After removing very short segments, these segments are automatically transcribed by the bilingual ASR. The most likely hypothesis output by the recognizer is used as the reference transcription. After obtaining the transcriptions, the manually and automatically transcribed data is merged to obtain the combined Frisian-Dutch broadcast data which is used for the training of the final acoustic model.

3.3. Language Diarization and Monolingual ASRs

The proposed automatic transcription incorporates a language diarization system between the speech activity detection and speaker diarization. The speech segments extracted from each recording are merged and language scores are assigned to the overlapping speech segments of N seconds with a frame shift with K seconds for $N > K$. The language detection scores are assigned using a LD system based on bottleneck features modeled using an i-vector framework followed by a Gaussian backend [30] (Additional details provided in Section 4.2). For each segment of K seconds, we apply majority voting among all language detection scores to decide on the assigned language label. The arguable choice of majority voting is due to the limited number of hypothesized language switches avoiding false alarms.

The language-labeled segments are automatically transcribed using monolingual resources at the back-end. Having one low-resourced and one high-resourced language as mixed languages, the monolingual resources of the highly resourced language, Dutch in this case, is expected to provide better ASR quality compared to a bilingual system which is trained to recognize both languages. Moreover, the multilingual training approach described in Section 3.2 can be used for getting acoustic models that can recognize Frisian speech only. Such a system also provides better recognition than a bilingual system and, hence, improved automatic transcriptions.

3.4. Language Model Rescoring

Lattice rescoring using a bilingual and two monolingual recurrent neural network (RNN) LMs has also been applied to extract alternative transcriptions. In [16], the transcriptions extracted with and without the rescoring stage have given similar results. However, in that case, the rescoring was performed with a bilingual LM with a higher perplexity than the monolingual LMs used in this paper. Therefore, we include the rescoring stage in the experiments, expecting to see improved transcription quality due to the lower perplexity of the monolingual LMs.

4. EXPERIMENTAL SETUP

4.1. Databases

The training data of the FAME! speech database comprises 8.5 hours and 3 hours of speech from Frisian and Dutch speakers respectively. The development and test sets consist of 1 hour of speech from Frisian speakers and 20 minutes of speech from Dutch speakers each. All speech data has a sampling frequency of 16 kHz.

The untranscribed radio broadcast data extracted from the same archive with the FAME! speech database consists of 256 hours 50 minutes of audio, including 159 hours 27 minutes of speech in total. The speech-only segments are fed either to the speaker diarization system described in Section 3.1 or to the language diarization system described in Section 3.3.

4.2. Implementation Details

The recognition experiments are performed using the Kaldi ASR toolkit [31]. We train a conventional context dependent Gaussian mixture model-hidden Markov model (GMM-HMM) system with 40k Gaussians using 39 dimensional mel-frequency cepstral coefficient (MFCC) features including the deltas and delta-deltas to obtain the alignments for DNN training. A standard feature extraction scheme is used by applying Hamming windowing with a frame length of 25 ms and frame shift of 10 ms. DNNs with six hidden layers and 2048 sigmoid hidden units at each hidden layer are trained on the 40-dimensional feature-level maximum likelihood linear transformations (fMLLR) [32] features with the deltas and delta-deltas. The DNN training is done by mini-batch stochastic gradient descent with an initial learning rate of 0.008 and a minibatch size of 256. The time context size is 11 frames achieved by concatenating ± 5 frames. We further apply sequence training using a state-level minimum Bayes risk (sMBR) criterion [33].

The bilingual lexicon contains 110k Frisian and Dutch words. The number of entries in the lexicon is approximately 160k due to the words with multiple phonetic transcriptions. The phonetic transcriptions of the words that do not appear in the initial lexicons are learned by applying grapheme-to-phoneme (G2P) bootstrapping [34, 35]. The lexicon learning is done only for the words that appear in the training data using the G2P model learned on the corresponding language. We use the Phonetisaurus G2P system [36] for creating phonetic transcriptions.

The details of the multilingually trained bilingual system that we used for the automatic transcription are found in [1]. The monolingual Frisian system used for automatically annotating the Frisian-labeled speech segments is also multilingually trained on the same data as the bilingual system, but retrained only on the Frisian training data. This system provides a WER of 32.7% and 30.9% on the Frisian component of development and test sets respectively. The monolingual Dutch system is trained on 110 hours of Dutch speech from the CGN corpus [37]. Further details of the Dutch ASR system are available in [1]. This system provides a WER of 27.9% and 23.8% on the Dutch component of development and test sets.

Standard bilingual and monolingual 3-gram with interpolated Kneser-Ney smoothing and RNN LMs are trained with 300 hidden units for recognition and lattice rescoring respectively. All language models are trained on a bilingual text corpus containing 37M Frisian and 8.8M Dutch words. The Frisian text is extracted from Frisian novels, news articles, Wikipedia articles and orthographic transcriptions of the FAME! training data.

SRI International's OLIVE software package is used for speaker diarization and language identification [38]. Speaker diarization is

Table 1. Word error rates in % obtained on the Frisian-only (fy), Dutch-only (nl) and code-switching (fy-nl) segments in the FAME! development and test sets

		Development				Test			
		fy	nl	fy-nl	all	fy	nl	fy-nl	all
# of Frisian words		9190	0	2381	11,571	10,753	0	1798	12,551
# of Dutch words		0	4569	533	5102	0	3475	306	3781
Approach	Training Data								
baseline	Man. Trans.	33.6	41.7	44.8	37.8	32.0	40.2	47.5	35.7
sad_sd	Man.+Auto. Trans.	31.2	35.7	44.2	34.7	29.3	34.8	46.7	32.7
sad_sd_rescore	Man.+Auto. Trans.	31.1	34.7	44.2	34.4	28.9	33.2	46.1	32.1
sad_ld_sd	Man.+Auto. Trans.	33.5	31.4	46.3	35.0	30.0	28.5	47.4	31.8
sad_ld_sd_rescore	Man.+Auto. Trans.	32.7	31.0	44.9	**34.2**	29.0	28.7	47.4	**31.2**

largely based on the process defined in [39]. Specifically, i-vectors [40] are first extracted from two-second segments of audio with 50% overlap prior to an exhaustive comparison of segments with probabilistic linear discriminant analysis (PLDA) to provide a matrix of scores representing speaker similarity. These scores are then transformed into a distance matrix, with distances being computed as the opposite of the log-likelihood ratios (LLR) obtained with PLDA shifted by the maximum LLR obtained for any pair of samples. Consequently, the minimum distance is 0. Finally, hierarchical clustering with average linkage method is used to generate a clustering tree which is pruned to ensure that each cluster has a cophenetic distance no greater than a pre-defined threshold tuned on a held-out corpus of telephone conversations.

The language diarization system is based on bottleneck features extracted from a DNN trained on English telephone data from the Fisher and Switchboard datasets to predict more than 2000 tied triphone states (senones). Bottleneck features of 80 dimensions are used to extract 400 dimensional i-vectors [40] from a framework that used a 2048-component universal background model (UBM). The Dutch and Frisian language classes are modeled by Gaussians with shared covariance. The LLR of the detected languages is calculated using calibration with a multi-class objective function [41] trained on the same data used to estimate the Gaussian parameters. The language diarization is performed with a frame length of N=30 sec. and a frame shift K=10 sec. As an initial step, we mainly aim for detecting *between-speaker switches* (cf. Section 2) in the raw broadcast data in this work. Therefore, a coarse LD with a frame shift on the order of 10 seconds is viable for this purpose. Investigating a finer LD that enables detecting intra-sentential and intra-word CS in raw data remains as future work.

4.3. Recognition and CS Detection Experiments

The baseline system is trained only on the manually annotated data. The other ASR systems incorporate acoustic models trained on the combined data which is automatically transcribed in various ways. These systems are tested on the development and test data of the FAME! speech database and the recognition results are reported separately for Frisian only (fy), Dutch only (nl) and mixed (fy-nl) segments. The overall performance (all) is also provided as a performance indicator. The recognition performance of the ASR system is quantified using the Word Error Rate (WER). The word language tags are removed while evaluating the ASR performance.

After the ASR experiments, we compare the CS detection performance of these recognizers. For this purpose, we use a different LM strategy. We train separate monolingual LMs, and interpolated between them with varying weights, effectively varying the prior for the detected language. For each LM, we generate the ASR output for each utterance. Then, we extract word-level segmentation files in `.ctm` format for each LM weight. By comparing these alignments with the ground truth word-level alignments (obtained by applying forced alignment using the baseline recognizer), a time-based CS detection accuracy metric is calculated. Specifically, we label each frame with a language tag for the ground truth and hypothesized alignments and calculate the total duration of the frames in the reference alignments with a mismatch with the hypothesized language tag. The missed Frisian (Dutch) time is calculated as the ratio of total duration of frames with a Frisian (Dutch) tag in the reference alignment which is aligned to frames without a Frisian (Dutch) tag to the total number of frames with a Frisian (Dutch) tag in the reference alignment. CS detection accuracy is evaluated by reporting the equal error rates (EER) calculated based on the detection error tradeoff (DET) graph [42] plotted for visualizing the CS detection performance. The presented code-switching detection results indicate how well the recognizer detects the switches and hypothesizes words in the switched language.

5. RESULTS

5.1. ASR Results

The WER obtained on each component of the FAME! development and test data is presented in Table 1. The upper panel of this table presents the number of Frisian and Dutch words in order to clarify the language priors in each subset. The baseline ASR system trained on the manually transcribed data provides a WER of 37.8% on the development and 35.7% on the test data. Only by adding the automatically transcribed data after applying SAD and SD improves the bilingual acoustic models, providing around 3% absolute reduction in WER on both datasets.

Bilingual rescoring brings minor improvements reducing the WER from 34.7% to 34.4% on the development data and 32.7% to 32.1% on the test data. Using language diarization for automatic transcription considerably improves the recognition performance on the test data reducing the WER from 32.7% to 31.8% without LM rescoring and 32.1% to 31.2% with LM rescoring. Comparable recognition performance is provided on the development data. The best performing ASR system on both datasets (sad_ld_sd_rescore) is trained using the automatic transcriptions provided by monolingual ASRs and monolingual rescoring.

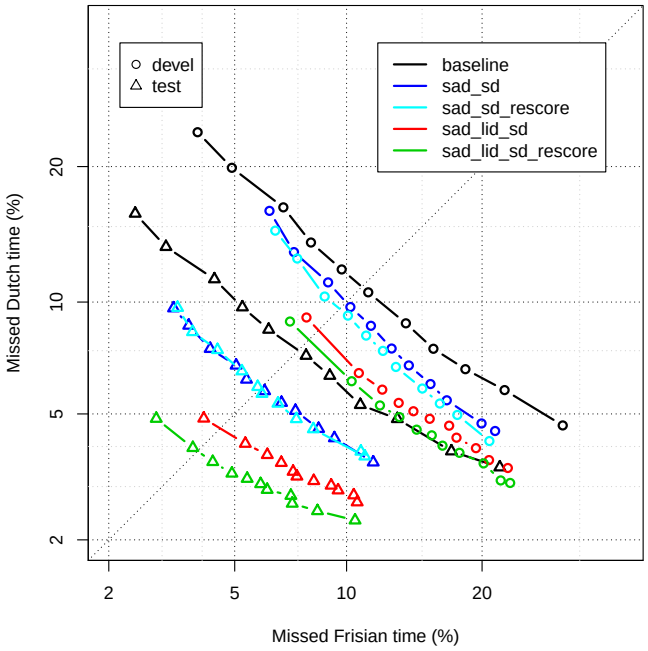

Fig. 2. Code-switching detection performance obtained on the FAME! development and test sets

5.2. CS Detection Results

The CS detection performance obtained on each component of the FAME! development and test data is presented in Figure 2. The baseline ASR system provides an EER of 10.9% on the development data and 7.5% on the test data. Increasing the amount of training data by automatic transcription also helps with CS detection accuracy and the sad_sd system reduces the EER to 9.9% and 5.9% on the development and test data respectively. Similar to the ASR results, bilingual rescoring also provides minor improvements by providing an EER of 9.6% on the development and 5.8% on the test data.

The sad_ld_sd system performs considerably better CS detection with an EER of 8.5% on the development and 4.5% on the test data. Applying monolingual LM rescoring during automatic transcription further improves the CS detection reducing the EERs to 8.1% on the development and 3.9% on the test data. From these results, we can conclude that using language diarization for semi-supervised acoustic model training boosts the quality of the language tags assigned by the final acoustic model and helps with detecting possible language switches. Moreover, monolingual LM rescoring provides larger improvements in CS detection compared to bilingual LM rescoring.

6. CONCLUSIONS

This paper describes a semi-supervised bilingual acoustic model training approach that uses a language diarization system to assign language labels to the speech segments extracted from the untranscribed data. Later, these language-labeled segments are automatically transcribed using monolingual resources and merged with the manually transcribed data for training bilingual acoustic models. We compare these acoustic models to others that are obtained using bilingual ASR for the automatic transcription. The acoustic models trained in the proposed manner provide minor improvement in ASR performance and considerable improvement in CS detection performance. Further, monolingual LM rescoring is found to provide larger improvements in automatic transcription quality compared to bilingual LM rescoring.

7. ACKNOWLEDGEMENTS

We would like to thank Aaron Lawson for his support with the arrangement of this research visit and Diego Castan and Mahesh Nandwana for their support with the SRI software used for the language diarization.

8. REFERENCES

[1] 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.

[2] G. Stemmer, E. Nöth, and H. Niemann, "Acoustic modeling of foreign words in a German speech recognition system," in *Proc. EUROSPEECH*, 2001, pp. 2745–2748.

[3] Dau-Cheng Lyu, Ren-Yuan Lyu, Yuang-Chin Chiang, and Chun-Nan Hsu, "Speech recognition on code-switching among the Chinese dialects," in *Proc. ICASSP*, May 2006, vol. 1, pp. 1105–1108.

[4] Ngoc Thang Vu, Dau-Cheng Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, Eng-Siong Chng, T. Schultz, and Haizhou Li, "A first speech recognition system for Mandarin-English code-switch conversational speech," in *Proc. ICASSP*, March 2012, pp. 4889–4892.

[5] T. I. Modipa, M. H. Davel, and F. De Wet, "Implications of Sepedi/English code switching for ASR systems," in *Pattern Recognition Association of South Africa*, 2015, pp. 112–117.

[6] T. Lyudovyk and V. Pylypenko, "Code-switching speech recognition for closely related languages," in *Proc. SLTU*, 2014, pp. 188–193.

[7] C. H. Wu, H. P. Shen, and C. S. Hsu, "Code-switching event detection by using a latent language space model and the delta-Bayesian information criterion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1892–1903, Nov 2015.

[8] 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.

[9] Ying Li and Pascale Fung, "Code switching language model with translation constraint for mixed language speech recognition," in *Proc. COLING*, Dec. 2012, pp. 1671–1680.

[10] H. Adel, N.T. Vu, F. Kraus, T. Schlippe, Haizhou Li, and T. Schultz, "Recurrent neural network language modeling for code switching conversational speech," in *Proc. ICASSP*, 2013, pp. 8411–8415.

[11] H. Adel, K. Kirchhoff, D. Telaar, Ngoc Thang Vu, T. Schlippe, and T. Schultz, "Features for factored language models for code-switching speech," in *Proc. SLTU*, May 2014, pp. 32–38.

[12] J. Weiner, Ngoc Thang Vu, D. Telaar, F. Metze, T. Schultz, Dau-Cheng Lyu, Eng-Siong Chng, and Haizhou Li, "Integration of language identification into a recognition system

for spoken conversations containing code-switches," in *Proc. SLTU*, May 2012.

[13] Dau-Cheng Lyu, Eng-Siong Chng, and Haizhou Li, "Language diarization for code-switch conversational speech," in *Proc. ICASSP*, May 2013, pp. 7314–7318.

[14] Yin-Lai Yeong and Tien-Ping Tan, "Language identification of code switching sentences and multilingual sentences of under-resourced languages by using multi structural word information," in *Proc. INTERSPEECH*, Sept. 2014, pp. 3052–3055.

[15] K. R. Mabokela, M. J. Manamela, and M. Manaileng, "Modeling code-switching speech on under-resourced languages for language identification," in *Proc. SLTU*, 2014, pp. 225–230.

[16] E. Yılmaz, H. Van den Heuvel, and D. A. Van Leeuwen, "Exploiting untranscribed broadcast data for improved code-switching detection," in *Proc. Interspeech*, Aug. 2017, p. Accepted for publication.

[17] J.-L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *ICSLP'98*, 1998, pp. 1335–1338.

[18] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–129, 2002.

[19] G. Riccardi and D. Hakkani-Tur, "Active learning: theory and applications to automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 504–511, July 2005.

[20] K. Yu, M. Gales, L. Wang, and P. C. Woodland, "Unsupervised training and directed manual transcription for LVCSR," *Speech Communication*, vol. 52, no. 78, pp. 652 – 663, 2010.

[21] T. Tsutaoka and K. Shinoda, "Acoustic model training using committee-based active and semi-supervised learning for speech recognition," in *Proc. of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, Dec 2012, pp. 1–4.

[22] R. Lileikyte, A. Gorin, L. Lamel, J.-L. Gauvain, and T. Fraga-Silva, "Lithuanian broadcast speech transcription using semi-supervised acoustic model training," *Procedia Computer Science*, vol. 81, pp. 107 – 113, 2016.

[23] Provinsje Fryslân, "De Fryske taalatlas 2015. De Fryske taal yn byld," 2015, Available at http://www.fryslan.frl/taalatlas.

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

[25] 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.

[26] M. Graciarena, L. Ferrer, and V. Mitra, "The SRI System for the NIST OpenSAD 2015 speech activity detection evaluation," in *Proc. Interspeech*, 2016, pp. 3673–3677.

[27] S. Thomas, S. Ganapathy, and H. Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in *Proc. ICASSP*, March 2012, pp. 4269–4272.

[28] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc. SLT*, Dec 2012, pp. 246–251.

[29] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*, 2013, pp. 7304–7308.

[30] Y. Song, B. Jiang, Y. Bao, S. Wei, and L.R. Dai, "I-vector representation based on bottleneck features for language identification," *Electronics Letters*, vol. 49, no. 24, pp. 1569–1570, 2012.

[31] 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.

[32] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75 – 98, 1998.

[33] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. INTERSPEECH*, 2013, pp. 2345–2349.

[34] M. Davel and E. Barnard, "Bootstrapping for language resource generation," in *Pattern Recognition Association of South Africa*, 2003, pp. 97–100.

[35] S. R. Maskey, A. B. Black, and L. M. Tomokiyo, "Bootstrapping phonetic lexicons for new languages," in *Proc. ICLSP*, 2004, pp. 69–72.

[36] J. R. Novak, N. Minematsu, and K. Hirose, "Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework," *Natural Language Engineering*, pp. 1–32, 9 2015.

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

[38] A. Lawson, M. McLaren, H. Bratt, M. Graciarena, H. Franco, C. George, A. Stauffer, C. Bartels, and J. Van Hout, "Open language interface for voice exploitation (OLIVE)," in *Proc. INTERSPEECH*, 2016, pp. 377–378.

[39] G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *Proc. IEEE Spoken Language Technology Workshop SLT*, 2014, pp. 413–417.

[40] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[41] N. Brummer and D. A. Van Leeuwen, "On calibration of language recognition scores," in *Proc. IEEE Odyssey - The Speaker and Language Recognition Workshop*, June 2006, pp. 1–8.

[42] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, Sep. 1997, pp. 1895–1898.