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

Exploiting Untranscribed Broadcast Data for Improved Code-Switching Detection

Emre Yılmaz, Henk van den Heuvel and David Van Leeuwen

CLS/CLST, Radboud University, Nijmegen, Netherlands
{e.yilmaz, h.vandenheuvel, d.vanleeuwen}@let.ru.nl

Abstract

We have recently presented an automatic speech recognition (ASR) system operating on Frisian-Dutch code-switched speech. This type of speech requires careful handling of unexpected language switches that may occur in a single utterance. In this paper, we extend this work by using some raw broadcast data to improve multilingually trained deep neural networks (DNN) that have been trained on 11.5 hours of manually annotated bilingual speech. For this purpose, we apply the initial ASR to the untranscribed broadcast data and automatically create transcriptions based on the recognizer output using different language models for rescoring. Then, we train new acoustic models on the combined data, i.e., the manually and automatically transcribed bilingual broadcast data, and investigate the automatic transcription quality based on the recognition accuracies on a separate set of development and test data. Finally, we report code-switching detection performance elaborating on the correlation between the ASR and the code-switching detection performance.

Index Terms: code-switching, bilingual ASR, under-resourced languages, Frisian language

1. Introduction

Spontaneous change between two languages in a single conversation, also known as code-switching (CS), is mostly noticeable in minority languages influenced by the majority language or majority languages influenced by *lingua francas* such as English and French. West Frisian (Frisian henceforth) has approximately half a million bilingual speakers and these speakers often code-switch between the Frisian and Dutch languages in daily conversations. In the scope of 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 the development of robust acoustic models operating on bilingual speech delving into the automatic speech recognition and CS detection aspects [1].

Impact of CS and other kinds of language switches on the speech-to-text systems have recently received research interest, resulting in several robust acoustic modeling [2–8] and language modeling [9–11] approaches for CS speech. Language identification (LID) is a relevant task for the automatic speech recognition (ASR) of CS speech [12–15]. One fundamental approach is to label speech frames with the spoken language and perform recognition of each language separately using a monolingual ASR system at the back-end. These systems have the tendency to suffer from error propagation between the language identification front-end and ASR back-end, since language identification is still a challenging problem especially in case of intra-sentence CS. To alleviate this problem, all-in-one ASR approaches, which do not directly incorporate a language identification system, have also been proposed [3, 6, 8].

Multilingual training of deep neural network (DNN)-based ASR systems has provided some improvements in the automatic recognition of both low- and high-resourced languages [16–25]. Some of these techniques incorporate multilingual DNNs for feature extraction [16, 21, 26, 27]. Training DNN-based acoustic models on multilingual data to obtain more reliable posteriors for the target language has also been investigated, e.g., [19, 20, 24]. In previous work, we have explored the recognition and code-switching detection performance of multilingual DNN models applied to the code-switching Frisian speech [1]. Multilingual data from closely related high-resourced languages, i.e., Dutch and English, is used for training 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 the prior work, e.g., in [19] 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 which 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, unlike the previous applications using multilingual DNN training for the recognition of a single target language.

In this work, we use the baseline bilingual ASR system trained on 11.5 hours of manually annotated data to automatically transcribe raw broadcast data. Later, a subset of this data with *reliable* transcriptions are combined with the manually annotated data aiming to obtain better acoustic modeling due to the considerable increase in the amount of training data. This type of semi-supevised acoustic model training has been researched intensively and various training strategies and data selection criteria have been proposed, e.g., in [28–33]. While getting the automatic transcriptions, we apply lattice rescoring using several language models (LM) such as a larger N-gram and a recurrent neural network (RNN) LM in order to investigate their impact on the quality of the automatic transcription. For assessing the quality of the automatic transcription, ASR experiments have been performed on a separate development and test data using the new acoustic models trained on the combined data. Finally, CS detection experiments have been conducted using the new models to reveal the relation between the ASR performance and CS detection which has never been investigated before to the best of our knowledge.

This paper is organized as follows. Section 2 introduces the demographics and the linguistic properties of the Frisian language. Section 3 summarizes the Frisian-Dutch radio broadcast database that has recently been collected for CS and lon-

http://dx.doi.org/10.21437/Interspeech.2017-391

gitudinal speech research. Section 4 summarizes how the untranscribed data is processed to extract speaker-labeled speech segments and automatic transcription is performed. The experimental setup is described in Section 5 and the recognition results are presented in Section 6. Section 7 concludes the paper.

2. Frisian Language

Frisian belongs to the North Sea Germanic language group, which is a subdivision of the West Germanic languages. Linguistically, there are three Frisian languages: West Frisian, spoken in the province of Fryslân in the Netherlands, East Frisian, spoken in Saterland in Lower Saxony in Germany, and North Frisian, spoken in the northwest of Germany, near the Danish border. These three varieties of Frisian are mutually barely intelligible [34]. The current paper focuses on the West Frisian language only and, following common practice, we will use the term Frisian for it.

Historically, Frisian shows many parallels with Old English. However, nowadays the Frisian language is under growing influence of Dutch 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 [35]. All speakers of Frisian are at least bilingual, since Dutch is the main language used in education in Fryslân.

The Frisian alphabet consists of 32 characters including all letters used in English and six others with diacritics, i.e., â, ê, é, ô, û, ú. The Frisian phonetic alphabet consists of 20 consonants, 20 monophthongs, 24 diphthongs, and 6 triphthongs. Frisian has more vowels compared to Dutch which has 13 monophthongs and 3 diphthongs [36]. Dutch consonants are similar to the Frisian ones. There are three main dialect groups in Frisian, i.e., Klaaifrysk (Clay Frisian), Wâldfrysk (Wood Frisian) and Súdwesthoeksk (Southwestern). Although these dialects differ mostly on phonological and lexical levels, they are mutually intelligible [37].

3. Frisian-Dutch Radio Broadcast Database

The bilingual FAME! speech database, which has been collected in the scope 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 [38]. To be able 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: 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. For further details, we refer the reader to [39].

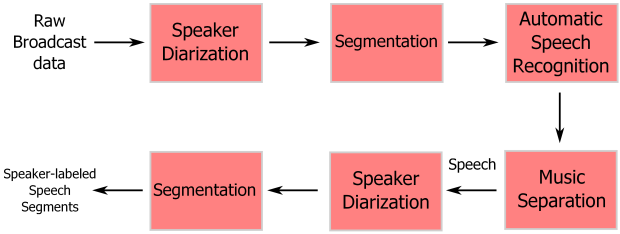

Figure 1: *Preprocessing the unlabeled broadcast data to extract speaker-labeled speech segments*

4. Processing Untranscribed Broadcast Data

The first task is to automatically annotate the raw radio programs with transcriptions and speaker information. These radio programs are extracted from the same radio broadcast archive with the FAME! Speech Corpus. Frisian is a low-resourced language with no available in-domain data, hence, the bilingual ASR system is expected to benefit from this untranscribed broadcast data. 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 of the raw broadcast data for automatic speech detection and 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 spoken content of each segment, they are fed to an ASR system and a subset of these segments are chosen based on the total number of words and average word length of the text hypothesized by the ASR. The segments that are suspected to be music based on the ASR output, e.g., segments with very small average word length and/or very few number of hypothesized words, are removed. The remaining speech segments in each program are automatically labeled with a speaker id by applying the same speaker diarization system.

The most likely hypothesis output by the recognizer is used as the reference transcription. Lattice rescoring using a 5-gram and an RNN LM have also been applied to extract alternative transcriptions. After obtaining the transcriptions, the manually and automatically transcribed data is merged to obtain the combined Frisian-Dutch broadcast data. The final acoustic models are trained on this combined database and the recognition and CS detection performance of these models are compared with the recognizer only trained on the manually transcribed data.

5. Experimental Setup

5.1. Recognition and CS Detection Experiments

The baseline recognizer uses the multilingually trained DNNs described in [1]. The same recognizer is used to automatically transcribe the raw broadcast data. A 3-gram with interpolated Kneser-Ney smoothing is trained using the SRILM toolkit [40] and this LM is used during the recognition phase. Similarly, a 5-gram with interpolated Kneser-Ney smoothing is trained for lattice rescoring purposes. Finally, a standard RNN LM is trained with 300 hidden units also for lattice rescoring. All these models are trained on a bilingual text corpus containing 37M Frisian and 8.8M Dutch words. The Frisian text is extracted

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*

		Devel				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.	32.7	38.4	44.0	36.2	29.7	34.9	46.6	33.0
norescoring	Man.+Auto. Trans.	32.3	**37.0**	43.9	35.6	**28.8**	**32.9**	**45.2**	**31.8**
rnnrescoring	Man.+Auto. Trans.	33.0	39.4	45.6	37.0	29.6	34.6	47.3	33.0
5Grnnrescoring	Man.+Auto. Trans.	**31.4**	37.2	**43.6**	**35.1**	28.9	33.0	46.2	32.0

from Frisian novels, news articles, wikipedia articles and orthographic transcriptions of the FAME! training data.

The final acoustic models trained on the combined data 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 used a different LM strategy. We trained separate monolingual LMs, and interpolated between them with varying weights, effectively varying the prior for the detected language. For each LM, we have generated 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 has been calculated. Specifically, we label each frame with a language tag for the ground truth and hypothesized alignments and calculate the total duration of frames in the reference alignments with a mismatch with hypothesized language tag. The missed Frisian (Dutch) time is calculated as the ratio of total duration of frames with Frisian (Dutch) tag in the reference alignment which is aligned to frames without Frisian (Dutch) tag to the total number of frames with Frisian (Dutch) tag in the reference alignment. The CS detection accuracy is evaluated by reporting the equal error rates (EER) calculated based on the detection error tradeoff (DET) graph [41] plotted for visualizing the CS detection performance. The presented code-switching detection results indicate how well the recognizer can detect the switches and hypothesize words in the switched language.

5.2. Implementation Details

Due to the superior performance of multilingual DNN training detailed in [1], the ASR systems used in the recognition experiments incorporate a multilingual lexicon with Frisian, Dutch and English words. The entries in the multilingual lexicon are extracted from the initial Fluency[1] Frisian (340k entries), ELEX[2] Dutch (600k entries) and CMU[3] English (134k entries) lexicons based on their presence in the transcriptions of all available training data and the text corpus used for LM training. In pilot experiments, modeling all Frisian vowels at the monophthong level has provided the best recognition performance. Therefore, all diphthongs and triphthongs are modeled as a sequence of their monophthong constituents.

The multilingual lexicon contains 144k Frisian, Dutch and English words. The number of entries in the lexicon is around 200k due to the words with multiple phonetic transcriptions. The phonetic transcriptions of the words which do not appear in the initial lexicons are learned by applying grapheme-to-phoneme (G2P) bootstrapping [42, 43]. 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 [44] for creating phonetic transcriptions. The OOV rates in the complete development and test set are 2.7% and 2.3%.

The IDIAP speaker diarization system [45] has been used for the preprocessing of the raw broadcast data illustrated in Figure 1. The total duration of the raw broadcast data is 150 hours 22 minutes and the duration of the remaining data after the preprocessing is 76 hours 40 minutes. The manually transcribed bilingual training data contains 8.5 hours of Frisian and 3 hours of Dutch speech. The recognition experiments are performed using the Kaldi ASR toolkit [46]. We train a conventional context dependent GMM-HMM system with 40k Gaussians using 39 dimensional 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 6 hidden layers and 2048 sigmoid hidden units at each hidden layer are trained on the 40-dimensional log-mel filterbank 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 [47].

6. Results

6.1. ASR Results

The recognition results obtained on the development and test sets of the FAME! speech database are given in Table 1. The WERs provided by the baseline recognizer and the recognizers trained on combined data (Man.+Aut. Trans) are presented in rows and the lowest WER of each column is marked in bold. 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 recognizer trained on the manually transcribed

[1] http://www.fluency.nl/

[2] http://tst-centrale.org/en/tst-materialen/lexica/e-lex-detail

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

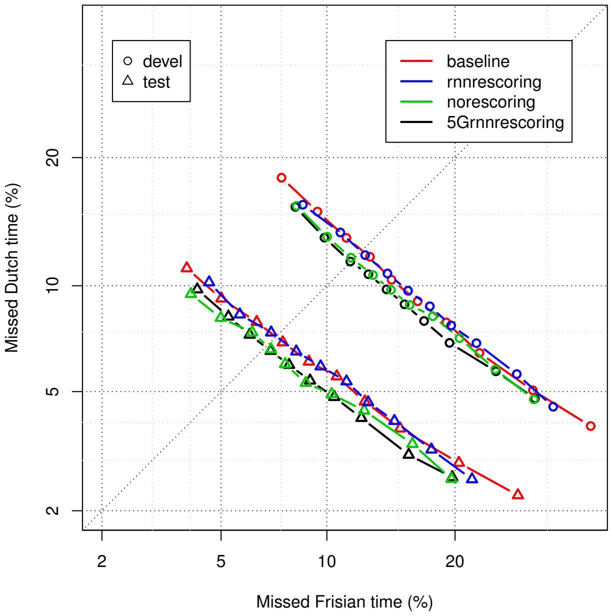

Figure 2: *Time-based code-switching detection evaluation obtained on the FAME! development and test sets*

data has a WER of 36.2% on the development set and 33.0% on the test set. The recognizers trained on combined data with automatic transcriptions provided by the baseline recognizer (norescoring) and two-stage lattice rescoring (5Grnnrescoring), i.e., 5-gram followed by an RNN LM, provide better recognition performance with a WER of 35.6% and 35.1% on the development set and 31.8% and 32.0% on the test set respectively. The performance of the system trained using the automatic transcriptions obtained after RNN rescoring (rnnrescoring) is comparable to the baseline recognizer. In the following subsection, we will investigate the relation between the ASR performance with the CS detection performance.

6.2. CS Detection Results

The DET curves provided by different approaches are plotted in Figure 2. Each point on these curves is obtained for a different language model weight. The time-based CS detection accuracy is lower on the development data with an EER of 12.2% compared to the test data with an EER of 7.3% using the baseline recognizer. Similar to the ASR performance, the CS detection accuracy provided by the systems trained using the automatic transcriptions created by norescoring and 5Grnnrescoring are better with EER values of 11.7% and 11.5% on the development set and 6.9% and 6.7% on the test set respectively. The rnnrescoring approach performs similar to the baseline recognizer yielding an EER of 12.2% on the development set and 7.3% on the test set which is also in parallel with the ASR performance.

6.3. Discussion

The recognition and CS detection experiments have shown that there is a strong correlation between the recognition accuracy and the detection of the CS words even using a primitive bilingual language model trained on text with very limited CS examples. Moreover, we can conclude that the ASR and CS detection performance of the bilingual recognizer can be improved by adding automatically transcribed speech data. However, the impact of lattice rescoring is not prominent given the similar performance of norescoring and 5Grnnrescoring approaches.

7. Conclusions

In this work, we make use of raw bilingual broadcast data to improve a multilingually trained ASR system designed for Frisian-Dutch code-switched speech. The initial system is trained on a small manually transcribed broadcast data extracted from the same archive. This recognizer is applied to the untranscribed broadcast data to create transcriptions automatically based on the recognizer output. Then, the manually and automatically transcribed bilingual broadcast data is combined and used for training new acoustic models. We first investigate the automatic transcription quality of different automatic transcription approaches based on the recognition accuracies on a separate set of development and test data. Furthermore, the CS detection performance of these recognizers are also presented to explore on how correlated the ASR accuracy is with the CS detection accuracy. From these results, it can be concluded that the recognition and CS detection performance of different acoustic models exhibit similar behaviour.

8. Acknowledgements

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

9. 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] D.-C. Lyu, R.-Y. Lyu, Y.-C. Chiang, and C.-N. Hsu, "Speech recognition on code-switching among the Chinese dialects," in *Proc. ICASSP*, vol. 1, May 2006, pp. 1105–1108.

[4] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, and H. 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] Y. Li and P. 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. Vu, F. Kraus, T. Schlippe, H. 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, N. T. 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, N. T. Vu, D. Telaar, F. Metze, T. Schultz, D.-C. Lyu, E.-S. Chng, and H. Li, "Integration of language identification into a recognition system for spoken conversations containing code-switches," in *Proc. SLTU*, May 2012.

[13] D.-C. Lyu, E.-S. Chng, and H. Li, "Language diarization for code-switch conversational speech," in *Proc. ICASSP*, May 2013, pp. 7314–7318.

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

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

[18] G. Heigold, V. Vanhoucke, A. W. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proc. ICASSP*, 2013, pp. 8619–8623.

[19] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*, 2013, pp. 7304–7308.

[20] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Proc. ICASSP*, 2013, pp. 7319–7323.

[21] Z. Tuske, J. Pinto, D. Willett, and R. Schluter, "Investigation on cross- and multilingual MLP features under matched and mismatched acoustical conditions," in *Proc. ICASSP*, May 2013, pp. 7349–7353.

[22] K. M. Knill, M. Gales, S. Rath, P. Woodland, C. Zhang, and S.-X. Zhang, "Investigation of multilingual deep neural networks for spoken term detection," in *Proc. ASRU*, Dec 2013, pp. 138–143.

[23] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Proc. ICASSP*, May 2014, pp. 7639–7643.

[24] A. Das and M. Hasegawa-Johnson, "Cross-lingual transfer learning during supervised training in low resource scenarios," in *Proc. INTERSPEECH*, 2015, pp. 3531–3535.

[25] A. Mohan and R. Rose, "Multi-lingual speech recognition with low-rank multi-task deep neural networks," in *Proc. ICASSP*, April 2015, pp. 4994–4998.

[26] N. T. Vu, F. Metze, and T. Schultz, "Multilingual bottle-neck features and its application for under-resourced languages," in *Proc. SLTU*, May 2012.

[27] K. Vesely, M. Karafiat, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proc. SLT*, Dec 2012, pp. 336–341.

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

[29] 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.

[30] 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.

[31] 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.

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

[33] R. Lileikyt, 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.

[34] A. P. Versloot, "Mechanisms of language change. vowel reduction in 15th century West Frisian," Ph.D. dissertation, University of Groningen, 2008.

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

[36] G. Booij, *The phonology of Dutch*. Oxford University Press, 1995.

[37] J. Popkema, *Frisian Grammar: The Basics*. Afûk, Leeuwarden, 2013.

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

[39] 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.

[40] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Proc. ICSLP*, 2002, pp. 901–904.

[41] 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.

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

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

[44] 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.

[45] 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.

[46] 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.

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