6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)
29-31 August 2018, Gurugram, India

Code-Switching Detection with Data-Augmented Acoustic and Language Models

Emre Yılmaz[1,2], *Henk van den Heuvel*[1] *and David A. van Leeuwen*[3]

[1] CLS/CLST, Radboud University, Nijmegen, Netherlands
[2] Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore
[3] ICIS, Radboud University, Nijmegen, Netherlands

eleey@nus.edu.sg, {h.vandenheuvel, d.vanleeuwen}@let.ru.nl

Abstract

In this paper, we investigate the code-switching detection performance of a code-switching (CS) automatic speech recognition (ASR) system with data-augmented acoustic and language models. We focus on the recognition of Frisian-Dutch radio broadcasts where one of the mixed languages, namely Frisian, is under-resourced. Recently, we have explored how the acoustic modeling (AM) can benefit from monolingual speech data belonging to the high-resourced mixed language. For this purpose, we have trained state-of-the-art AMs on a significantly increased amount of CS speech by applying automatic transcription and monolingual Dutch speech. Moreover, we have improved the language model (LM) by creating CS text in various ways including text generation using recurrent LMs trained on existing CS text. Motivated by the significantly improved CS ASR performance, we delve into the CS detection performance of the same ASR system in this work by reporting CS detection accuracies together with a detailed detection error analysis.

Index Terms: code-switching detection, language switches, language recognition, under-resourced languages, Frisian language

1. Introduction

Detecting the language switches in speech with code-switching (CS) is a relevant application while designing automatic speech recognition (ASR) systems for multilingual societies. 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 [1–11] and language modeling [12–17] approaches for CS speech. Our research in the FAME! project focuses on developing an all-in-one CS ASR system using a Frisian-Dutch bilingual acoustic and language model that allows language switches [7, 18].

One major challenge when building a CS ASR system is the lack of training speech and text data to train reliable acoustic and language models that can accurately recognize the uttered word and its language. The latter is relevant in language pairs such as Frisian and Dutch with orthographic similarity and shared vocabulary. For this purpose, we have proposed automatic transcription strategies developed for CS speech to increase the amount of training speech data in previous work [19,20]. Thanks to the increased CS training data [20], we reported improvements in the ASR and CS detection accuracy using fully connected deep neural networks (DNN). Later in [21], we have investigated how to combine the out-of-domain speech data from the high-resourced mixed language with this increased amount of CS speech data for acoustic training of a more recently proposed neural network architecture consisting of time-delay and recurrent layers [22].

For the language modeling, we have thus far incorporated a standard bilingual language model which is mostly trained on a combination of monolingual text from Frisian and Dutch. The only component in the training text with CS is the transcriptions of the FAME! Corpus [23] training speech data. To enrich this baseline LM, in [21], we generate CS text either by training long short-term memory (LSTM) language models on the very small amount of CS text extracted from the transcriptions of the training speech data and synthesize much larger amounts of CS text using these models [24] or by translating Dutch text extracted from the transcriptions of a large Dutch speech corpora. The latter provides CS text as some Dutch words, such as proper nouns (most commonly person, place and institution names) which are a prominent source of code-switching, are not translated to Frisian while the neighboring words are. We further include the transcriptions provided by the best performing automatic transcription strategies described in [20].

In this work, we investigate the code-switching detection performance of different ASR systems using these augmented acoustic and language models by presenting the detection performance. Moreover, we compare the frequency of the hypothesized language switches, duration distribution of the monolingual segments, accuracy of the assigned language tags and the most common erroneously recognized word pairs with different language tags to provide further insight into the quality of the hypothesized language switches.

2. 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 Frisian language shows many parallels with Old English. However, nowadays it is under growing influence of Dutch due to long lasting and intense language contact. Almost all speakers of Frisian are at least bilingual, since Dutch is the main language used in education in Fryslân.

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 [25]. 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). For

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

10.21437/SLTU.2018-27

Table 1: *Acoustic data composition of different training setups used in the recognition experiments (in hours)*

Training data	Annot.	Frisian	Dutch	Flemish	Total
(1) FAME! [23]	Manual	8.5	3.0	-	11.5
(2) Frisian Broad. [20]	Auto.	125.5		-	125.5
(3) CGN-NL [27]	Manual	-	442.5	-	442.5
(4) CGN-VL [27]	Manual	-	-	307.5	307.5

further details, we refer the reader to [23].

3. Data Augmentation

3.1. Acoustic Modeling

In previous work, we described several automatic annotation approaches to enable using of a large amount of raw bilingual broadcast data for acoustic model training in a semi-supervised setting. For this purpose, we performed various tasks such as speaker diarization, language and speaker recognition and LM rescoring on raw broadcast data for automatic speaker and language tagging [19, 20] and later used this data for acoustic training together with the manually annotated (reference) data. These approaches improved the recognition performance on the CS due to the significant increase in the available CS training data.

Our later efforts focus on the possible improvements in the acoustic modeling that can be obtained using other datasets with a much larger amount of monolingual speech data from the high-resourced mixed language, which is Dutch in our scenario. Previously, adding even a small portion of this Dutch data resulted in severe recognition accuracy loss in the low-resourced mixed language. After the significant increase in the CS training speech data, we explore to what extent one can benefit from the greater availability of resources for Dutch. Given that the Dutch spoken in the target CS context is characterized by the West Frisian accent [26], we further include speech data from a language variety of Dutch, namely Flemish, to investigate its contribution towards the accent-robustness of the final acoustic model.

3.2. Language Modeling

Language modeling in a CS scenario mainly suffers from lack of adequate training material, as CS rarely occurs (mostly in an informal context such personal messages and tweets) in written resources. Therefore, finding enough training material for a language model that can model word sequences with CS is very challenging. Our main source for CS text was the transcriptions of the training speech data which comprises a very small proportion of the bilingual training text compared to the monolingual Frisian and monolingual Dutch text.

The CS language model (LM) used in this work is enriched by generating code-switching text in three ways: (1) generating text using recurrent LMs trained on the transcriptions of the training CS speech data, (2) adding the transcriptions of the automatically transcribed CS speech data and (3) translating Dutch text extracted from the transcriptions of a large Dutch speech corpora. The perplexity improvements provided by each component is presented in [21].

4. Experimental Setup

4.1. Databases

4.1.1. Spoken Material

Details of all training data used for the experiments are presented in Table 1. The training data of the FAME! speech corpus 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 raw radio broadcast data extracted from the same archive as the FAME! speech corpus consists of 256.8 hours of audio, including 159.5 hours of speech based on the SAD [28] output. The amount of total raw speech data extracted from the target broadcast archive after removing very short segments is 125.5 hours. We refer to this automatically annotated data as the 'Frisian Broadcast' data.

Monolingual Dutch speech data comprises the complete Dutch and Flemish components of the Spoken Dutch Corpus (CGN) [27] that contains diverse speech material including conversations, interviews, lectures, debates, read speech and broadcast news. This corpus contains 442.5 and 307.5 hours of Dutch and Flemish data respectively.

4.1.2. Written Material

The baseline language models are trained on a bilingual text corpus containing 37M Frisian and 8.8M Dutch words. Almost all Frisian text is extracted from monolingual resources such as Frisian novels, news articles, Wikipedia articles. The Dutch text is extracted from the transcriptions of the CGN speech corpus which has been found to be very effective for language model training compared to other text extracted from written sources. The transcriptions of the FAME! training data is the only source of CS text and contains 140k words.

Using this small amount of CS text, we train LSTM-LM and generate text with 10M, 25M, 50M and 75M words. The translated CS text contains 8.5M words. Finally, we use the automatic transcriptions provided by the best-performing monolingual and bilingual automatic transcription strategy which contains 3M words in total. The details of these strategies are described in [20]. The final training text corpus after including the generated text has 107M words.

4.2. Implementation Details

The recognition experiments are performed using the Kaldi ASR toolkit [29]. 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 training a lattice-free maximum mutual information (LF-MMI) [30] TDNN-LSTM [22] AM (1 standard, 6 time-delay and 3 LSTM layers). We use 40-dimensional MFCC as features combined with i-vectors for speaker adaptation [31]. Further details are provided in [21].

The baseline language models are standard bilingual 3-gram with interpolated Kneser-Ney smoothing and an RNN-LM [32] with 400 hidden units used for recognition and lattice rescoring respectively. The RNN-LMs with gated recurrent units (GRU) [33] and noise contrastive estimation [34] are trained. A tied LSTM-LM [35] with 650 hidden units per layer and 650-dimensional word embeddings is used for CS text generation which is trained for 40 epochs using the example Pytorch implementation available at `https://github.com/pytorch/examples`.

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-

Table 2: *WER (%) obtained on the development and test set of the FAME! Corpus - Different AM training data is identified with the numbers which are defined in Table 1.*

				Dev.				Test				Total
				fy	nl	fy-nl	all	fy	nl	fy-nl	all	
# of Frisian words				9190	0	2381	11,571	10,753	0	1798	12,551	24,122
# of Dutch words				0	4569	533	5102	0	3475	306	3781	8883
ASR System	AM train data	LM train data	Lang. Tag									
Baseline ASR	(1)	Orig.	No	36.4	43.7	48.2	40.3	31.5	39.5	47.9	35.2	37.8
Baseline ASR	(1)	Orig.	Yes	37.9	48.3	53.3	43.3	32.8	42.5	51.9	37.2	40.3
ASR_AA	(1+2)	Orig.	No	31.1	35.2	42.4	34.1	28.6	31.8	44.0	31.2	32.7
ASR_AA	(1+2)	Orig.	Yes	32.9	37.8	48.2	36.8	29.5	34.5	49.4	33.1	35.0
ASR_AA_NL	(1+2+3)	Orig.	No	25.8	27.4	36.9	28.1	24.8	24.6	37.6	26.4	27.2
ASR_AA_NL	(1+2+3)	Orig.	Yes	27.4	30.7	43.1	30.9	25.8	27.8	44.1	28.5	29.7
ASR_AA_NL-VL	(1+2+3+4)	Orig.	No	26.4	26.9	36.2	28.1	24.5	23.2	38.5	26.0	27.1
ASR_AA_NL-VL	(1+2+3+4)	Orig.	Yes	27.7	30.5	42.7	31.0	25.5	26.1	43.5	27.8	29.4
ASR_AA_NL-VL_CS-LM	(1+2+3+4)	Orig.+Gen.	No	24.6	26.7	33.3	26.5	22.5	22.4	32.9	23.8	25.2
ASR_AA_NL-VL_CS-LM	(1+2+3+4)	Orig.+Gen.	Yes	25.9	30.2	38.0	29.1	23.5	26.7	38.6	26.0	27.6

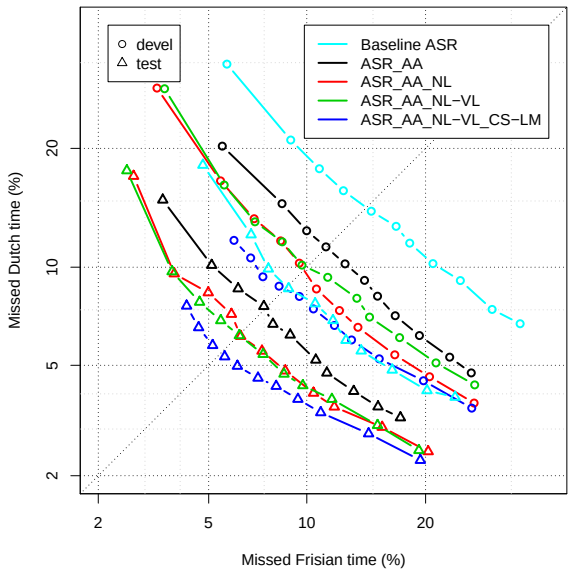

Figure 1: *Code-switching detection performance obtained on the FAME! development and test sets*

phoneme (G2P) bootstrapping [36, 37]. The lexicon learning is carried out 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 [38] for creating phonetic transcriptions.

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 with and without language tags. The results with language tags identify the quality of the assigned language tags for each component. 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).

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 for each LM weight. By comparing these alignments with the ground truth word-level alignments, a time-based CS detection accuracy metric is calculated [19]. CS detection accuracy is evaluated by reporting the equal error rates (EER) calculated based on the detection error tradeoff (DET) graph [39] 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 and Discussion

5.1. ASR Results

The ASR results with and without language tags provided by different ASR systems are presented in Table 2. The baseline ASR trained only on the FAME speech corpus has a total WER of 37.8% without language tags and 40.3% with language tags. Applying data augmentation to the AM and LM consistently improves the ASR performance [21], while the performance gap between the results with and without language tag stays relatively constant between 2%-2.5% absolute indicating a relatively constant source of language recognition errors that cannot be eliminated using the augmented AM and LM.

Furthermore, for all systems, the recognition accuracy on the Dutch only and mixed segments reduces more dramatically due to the language recognition errors compared to the Frisian segments. This suggests the CS LM (yielding the lowest perplexity on the development text) used during the ASR experiments has the tendency to assign Frisian language tags more frequently than Dutch to the phonetically and orthographically similar words. To address these observations, an error analysis on language tag confusions is given in Section 5.3.

5.2. CS Detection Results

The CS detection performance provided by the ASR systems on each component of the FAME! development and test data is presented in Figure 1. The baseline ASR system provides an EER of 14.4% (8.7%) on the development (test) data. Adding automatically annotated in-domain speech and monolingual Dutch speech data (ASR_AA_NL) reduces the CS detection accuracy considerably with an EER of 9.8% (6.3%). The improved AM yielding a 12.2% (8.8%) absolute WER improvement compared

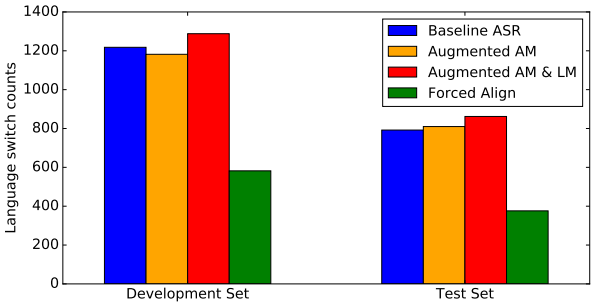

Figure 2: *Hypothesized language switch counts*

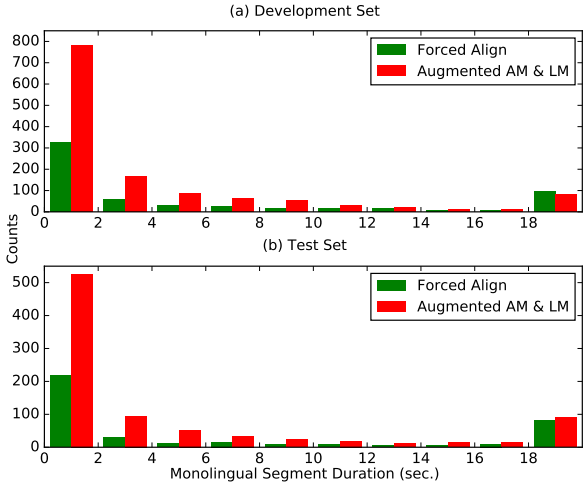

Figure 3: *Duration distribution of monolingual segments*

to the baseline ASR (cf. Table 2) also improves the quality of the assigned language tags with an absolute EER reduction of 4.6% (2.4%).

Comparable to the ASR results, adding monolingual Flemish speech data does not bring any improvement in the CS detection accuracy. Finally, enriching the CS LM using generated textual training data improves the CS detection accuracy by reducing the EER to 8.6% (5.5%) which is the best CS detection accuracy among the presented systems.

5.3. Discussion

After reporting the ASR and CS detection accuracies, we analyze the CS cases hypothesized by different ASR systems. We first present the language switch counts for the baseline ASR, the ASR with the augmented AM (ASR_AA_NL-VL) and the ASR with augmented AM and LM (ASR_AA_NL-VL_CS-LM). Later, the histogram plot of the monolingual segment durations provided by the best-performing ASR system in Table 2 is compared with the forced alignment output (used as the ground truth). Finally, the most frequent (word-language tag) pair confusions are listed to demonstrate the characteristics of the words that are resulting in CS detection errors.

Figure 2 illustrates the language switch counts belonging to different ASR systems. In general, all ASR systems have the tendency to overestimate the number of actual language switches annotated by bilingual Frisian-Dutch speakers. This overestimation in the number of language switches is not noticeable in our duration-based CS detection metric as hypothesizing very short erroneous language switches is less penalized compared to incorrect language tags assigned over longer segments which gives a better indication of the general language recognition capability of the corresponding ASR system. Another observation is the consistent increase in the amount of hypothesized language switches after enriching the CS LM with generated CS textual data containing new CS examples. As expected, the ASR system with augmented AM and LM allows more language switches compared to the other ASR systems on both datasets.

Table 3: *Most frequent (word-language tag) confusions of the best-performing ASR system on the development and test data*

Dev.			Test		
Ref. word	Hyp. word	Count	Ref. word	Hyp. word	Count
en-nl	en-fy	26	en-nl	en-fy	42
de-fy	de-nl	24	dat-nl	dat-fy	33
dat-nl	dat-fy	18	de-fy	de-nl	29
wat-nl	wat-fy	16	is-nl	is-fy	25
het-nl	it-fy	15	ja-nl	ja-fy	24
dat-fy	dat-nl	15	it-fy	't-nl	24
it-fy	't-nl	14	de-nl	de-fy	23
is-nl	is-fy	14	het-nl	it-fy	21
de-nl	de-fy	13	ja-fy	ja-nl	20
in-fy	een-nl	12	dat-fy	dat-nl	20

As a next step, we further compare the distribution of monolingual segments at the output of the best-performing ASR and forced alignment of the orthographic transcriptions to get a better idea about the overestimation of language switch counts. As it can be seen from Figure 3, the output of the ASR system contains more monolingual segments that are shorter then 10 seconds compared to the forced alignment output. Especially, there is a significant difference in the number of segments that are shorter than 2 seconds. From Figure 2 and 3, it can be concluded that the best-performing ASR system suffers from false alarms (by hypothesizing very short erroneous language switches) despite providing the best detection performance using a duration-based metric. Investigating smoothing techniques for language tag switches to reduce these false alarm remains as a future work.

In Table 3, the most frequently substituted (word-language tag) pairs are listed as the main source of the CS detection errors. From these lists, it can be seen that most common language tag errors are orthographically identical (or similar) short filler words with the same meaning both in Frisian and Dutch. The orthographically identical words, some of which are appearing in this list, induces the 2%-2.5% absolute WER difference between the ASR results with and without language tags. However, these errors are less of a concern for the ultimate goal of building a spoken document retrieval system.

6. Conclusions

This paper explores the CS detection performance of a state-of-the-art CS ASR with data-augmented acoustic and language model. After reporting significant improvements in the ASR accuracy, we explore how well the same ASR system assigns language tags compared to a baseline ASR system using only manually annotated data in the scenario of under-resourced Frisian-Dutch CS speech. The CS detection experiments demonstrate the superior CS detection performance of the data-augmented CS ASR system. Further CS analysis of the ASR output gives insight about the quality of the hypothesized CS cases: (1) the best-performing ASR system hypothesizes more language switches compared to the manual annotations, (2) this overestimation is mostly due to numerous very short erroneous language switches, and (3) the main source of CS detection errors are orthographically identical short filler words which are invisible in the absence of the language tags.

7. References

[1] 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.

[2] 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.

[3] 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 sys-tem for Mandarin-English code-switch conversational speech," in *Proc. ICASSP*, March 2012, pp. 4889–4892.

[4] T. I. Modipa, M. H. Davel, and F. De Wet, "Implications of Se-pedi/English code switching for ASR systems," in *Pattern Recog-nition Association of South Africa*, 2015, pp. 112–117.

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

[6] 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.

[7] 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. SLTU*, May 2016, pp. 159–166.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Z. Zeng, H. Xu, T. Y. Chong, E. S. Chng, and H. Li, "Improving N-gram language modeling for code-switching speech recognition," in *Proc. APSIPA ASC*, 2017, pp. 1–6.

[16] I. Hamed, M. Elmahdy, and S. Abdennadher, "Building a first language model for code-switch Arabic-English," *Procedia Computer Science*, vol. 117, pp. 208 – 216, 2017.

[17] E. van der Westhuizen and T. Niesler, "Synthesising isiZulu-English code-switch bigrams using word embeddings," in *Proc. INTERSPEECH*, 2017, pp. 72–76.

[18] 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.

[19] E. Yılmaz, M. McLaren, H. Van den Heuvel, and D. A. Van Leeuwen, "Language diarization for semi-supervised bilin-gual acoustic model training," in *Proc. ASRU*, Dec. 2017, pp. 91–96.

[20] "Semi-supervised acoustic model training for speech with code-switching," *Submitted to Speech Communication*, 2018. [Online]. Available: goo.gl/NKiYAF

[21] E. Yılmaz, H. Van den Heuvel, and D. A. Van Leeuwen, "Acoustic and textual data augmentation for improved ASR of code-switching speech," in *Accepted for publication, INTERSPEECH*, Sept. 2018.

[22] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, March 2018.

[23] 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.

[24] I. Sutskever, J. Martens, and G. Hinton, "Generating text with recurrent neural networks," in *Proc. ICML-11*, June 2011, pp. 1017–1024.

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

[26] R. v. Bezooijen and J. Ytsma, "Accents of Dutch: personality impression, divergence, and identifiability," *Belgian Journal of Lin-guistics*, pp. 105–129, 1999.

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

[28] 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.

[29] 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.

[30] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. INTERSPEECH*, 2016, pp. 2751–2755.

[31] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. ASRU*, Dec 2013, pp. 55–59.

[32] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. INTERSPEECH*, 2010, pp. 1045–1048.

[33] J. Chung, C. Gulcehre, and Y. Kyung Hyun Cho, Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv:1412.3555*, 2014.

[34] X. Chen, X. Liu, M. J. F. Gales, and P. C. Woodland, "Recurrent neural network language model training with noise contrastive estimation for speech recognition," in *Proc. ICASSP*, April 2015, pp. 5411–5415.

[35] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Proc. INTERSPEECH*, 2012, pp. 194–197.

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

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

[38] 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.

[39] 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.