CODE-SWITCHING DETECTION USING MULTILINGUAL DNNS

Emre Yılmaz, Henk van den Heuvel and David van Leeuwen

CLS/CLST, Radboud University, Nijmegen, Netherlands

ABSTRACT

Automatic speech recognition (ASR) of code-switching speech requires careful handling of unexpected language switches that may occur in a single utterance. In this paper, we investigate the feasibility of using multilingually trained deep neural networks (DNN) for the ASR of Frisian speech containing code-switches to Dutch with the aim of building a robust recognizer that can handle this phenomenon. For this purpose, we train several multilingual DNN models on Frisian and two closely related languages, namely English and Dutch, to compare the impact of single-step and two-step multilingual DNN training on the recognition and code-switching detection performance. We apply bilingual DNN retraining on both target languages by varying the amount of training data belonging to the higher-resourced target language (Dutch). The recognition results show that the multilingual DNN training scheme with an initial multilingual training step followed by bilingual retraining provides recognition performance comparable to an oracle baseline recognizer that can employ language-specific acoustic models. We further show that we can detect code-switches at the word level with an equal error rate of around 17% excluding the deletions due to ASR errors.

Index Terms— Language contact, multilingual DNN, code-switching, Frisian, under-resourced languages

1. INTRODUCTION

Code-switching (CS) is defined as the continuous alternation between two languages in a single conversation. 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 speakers who are mostly bilingual and it is common practice to code-switch between the Frisian and Dutch languages in daily conversations. In the scope of the FAME! Project, the influence of this spontaneous language switching on modern ASR systems is explored with the objective of building a robust recognizer that can handle this phenomenon.

In addition to the well-established research line in linguistics, implications of CS and other kinds of language switches for speech-to-text systems have recently received some research interest, resulting in some robust acoustic modeling [1–5] and language modeling [6–8] approaches for CS speech. Language identification (LID) is a relevant task for the automatic speech recognition (ASR) of CS speech [9–12]. 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 [2, 5].

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 [13–22]. Some of these techniques incorporate multilingual DNNs for feature extraction [13, 18, 23, 24]. Training DNN-based acoustic models on multilingual data to obtain more reliable posteriors for the target language has also been investigated, e.g., [16, 17, 21].

In this work, we explore the recognition and code-switching detection performance of multilingual DNN models applied to the code-switching Frisian speech. 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 [16] 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 approaches using multilingual DNN training for the recognition of a single target language.

The performance of multilingual DNN models is compared to a baseline ASR system using the oracle LID information at the front-end and three different recognizers at the back-end. In this way, we compare the recognition performance of multilingual DNNs and a conventional CS recognition system with an ideal LID at the front-end which has never been explored to the best of our knowledge. Moreover, we vary the amount of the high-resourced target language, i.e., Dutch, to quantify the feasible amount of Dutch training data for the multilingual DNN to perform reasonably well on both languages. Finally, we discuss the word-level CS detection performance of the recognizer described above to provide some insight into how well this recognizer can cope with the language switches.

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 longitudinal speech research. Section 4 summarizes the fundamentals of the DNN-HMM ASR system and describes the two-step multilingual training of DNNs applied to the CS Frisian speech. The experimental setup is described in Section 5 and the recognition results are presented in Section 6. Section 7 concludes the paper.

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

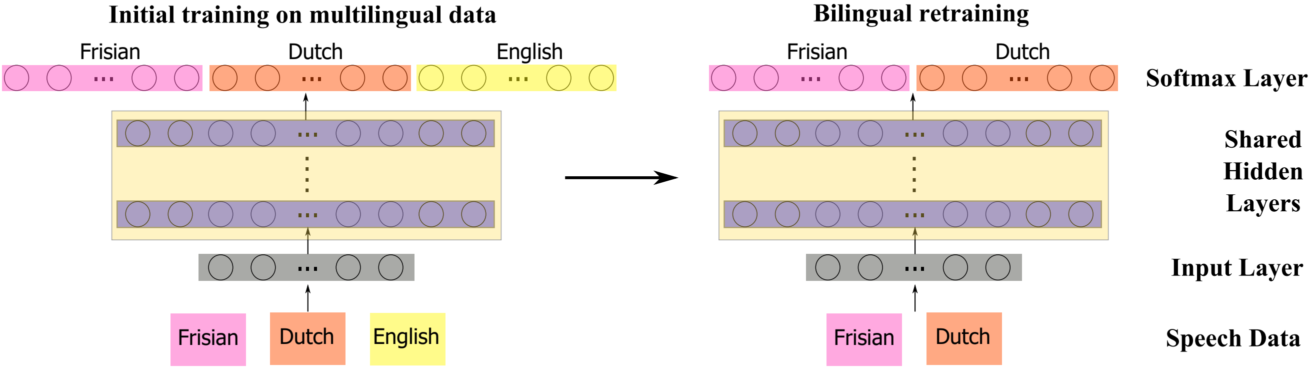

Fig. 1. Multilingual DNN training scheme designed for code-switching Frisian speech

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 [25]. 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 [26]. 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 [27]. 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 [28].

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 [29]. 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 longitudinal and bilingual nature of the material enables research into several fields such as language variation in Frisian over years, formal versus informal speech, CS trends, speaker tracking and diarization over a large time period.

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 [30].

It is important to note that two kinds of language switches are observed in broadcast data in the absence of segmentation information. Firstly, 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 instance, 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 the ASR systems and have to be handled carefully.

4. MULTILINGUAL DNN TRAINING

4.1. Fundamentals

Our DNN consists of L layers of M artificial neurons and the output of the $(l-1)^{\text{th}}$ layer with M_{l-1} neurons is the input of the l^{th} layer with M_l neurons, which is formulated as $\mathbf{v}_l = f(\mathbf{z}_l) = f(\mathbf{W}_l\mathbf{v}_{l-1} + \mathbf{b}_\mathbf{l})$ where the dimensions of $\mathbf{v}_l$, $\mathbf{W}_l$, $\mathbf{v}_{l-1}$ and $\mathbf{b}_\mathbf{l}$ are M_l, $(M_l \times M_{l-1})$, M_{l-1} and M_l respectively. M_0 is the number of neurons in the input layer which is equal to the dimension of the speech features. The non-linear activation function f maps an M_{l-1} vector to an M_{l-1} vector. The activation function applied at the output layer is the softmax function in order to get output values in the range $[0,1]$ for the hidden Markov model (HMM) state posterior probabilities.

The DNN-HMM training is achieved in three main stages [31, 32]. Firstly, a GMM-HMM setup is trained to obtain the structure of the DNN-HMM model, initial HMM transition probabilities and training labels of the DNNs. Then, the pretraining algorithm described in [33] is applied to obtain a robust initialization for the DNN model. Finally, the back-propagation algorithm [34] is applied to train the DNN that will be used as the emission distribution of the HMM states.

Table 1. Data composition of different training setups used in the recognition experiments (in hours)

Training data	Frisian	Dutch	English	Total
fy	8.5	-	-	8.5
fy-nl	8.5	3.0	-	11.5
fy-nl+	8.5	20.5	-	29.0
fy-nl++	8.5	110.0	-	118.5
fy-en	8.5	-	141.5	150.0
fy-nl++-en	8.5	110.0	141.5	260.0

4.2. Multilingual training

Using language resources from high-resourced languages for the recognition of an under-resourced language is common practice [35–37]. Being an under-resourced language, Frisian also lacks adequate speech data to train acoustic models that can provide accurate enough recognition. Therefore, an ASR system working on Frisian benefits from bootstrapping data from other closely related languages such as Dutch and English. Moreover, the code-switching nature of Frisian requires to incorporate bilingual resources for the ASR system to handle unexpected switches to Dutch.

The multilingual training scheme applied in this paper is illustrated in Figure 1. In this multilingual training scheme, the phones of each language are modeled separately, e.g., by appending a language identifier to every phone of a word based on the language of its lexicon. We refer the reader to [38] in which the impact of phone merging in the context of CS ASR is explored. The words in each lexicon are also tagged with language identifier to be able to evaluate CS detection accuracy.

The multilingual DNN training is performed on spectral features that allow the cross-lingual knowledge transfer. The knowledge transfer is achieved by using the hidden layers of the DNN trained on speech data from all languages during the initial training phase [32]. The amount of training data used during the initial training phase can be increased by including more data from high-resourced languages. Retraining these shared hidden layers with a new softmax layer aims to fine-tune the initial model to the target speech data. The retraining step is achieved by using bilingual speech data so that the recognizer can recognize both target languages.

5. EXPERIMENTAL SETUP

We perform ASR experiments to investigate the recognition performance provided by the training scheme described in Section 4. We use English and Dutch speech databases for training purposes and the FAME! speech database which is used for training, development and testing purposes.

5.1. Databases

The Dutch speech databases used for DNN training are the Dutch Broadcast database [39] and the components of the CGN [40] corpus that has broadcast-related recordings. These databases contain 17.5 and 89.5 hours of Dutch data respectively. In addition to this, English Broadcast News Database (HUB4) is used as the main source of English broadcast data. The amount of the English data extracted from both 1996 and 1997 components of HUB4 [41, 42] is approximately 141.5 hours.

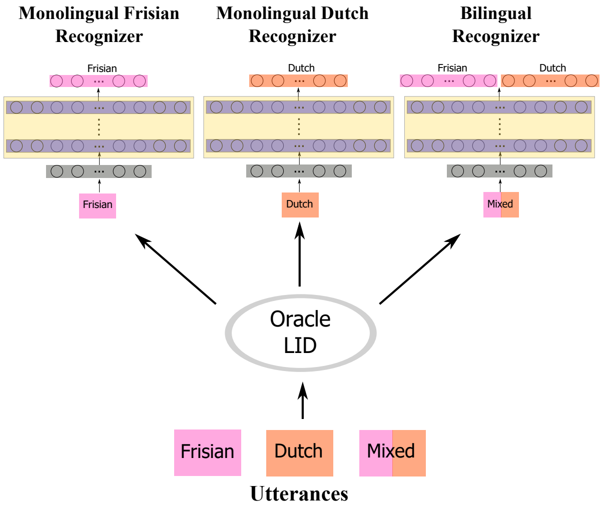

Fig. 2. The baseline CS recognizer using ground truth language information

The training data of the FAME! speech database comprises of 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 total number of word- and sentence-level CS cases in the FAME! speech database is 3837. These switches are mostly performed by the Frisian speakers as they often use Dutch words or sentences while speaking in Frisian. These cases comprise about 75.6% of all switches. The opposite case, i.e., a Dutch speaker using Frisian words or sentences, occurs much less accounting for 2.5% of all switches. This is expected as it is not common practice for Dutch speakers to switch between Dutch and Frisian. In the rest of the cases, the speakers use a *mixed-word* which is neither Frisian nor Dutch, for instance adapted loanwords. The training, development and test sets contain 2756, 671 and 410 language switching cases. There are 542 speakers in the FAME! speech database in total, most of whom are radio program presenters and Frisian celebrities.

5.2. Recognition and CS Detection Experiments

The baseline system, which is depicted in Figure 2, has a conventional CS recognition architecture [2] and uses the ground truth language tags to choose the most appropriate recognition system. Oracle LID information is provided in order to eliminate recognition degradation due to the LID errors. Based on the language tag of each utterance, the recognition is performed by a monolingual Frisian recognizer for Frisian only utterances, a monolingual Dutch recognizer for Dutch only utterances or a bilingual Frisian-Dutch recognizer for mixed utterances. The monolingual Frisian and Dutch recognizers are trained on fr and nl++ (cf. Table 1) data respectively. The bilingual recognizer is trained on the fr-nl++ data which combines the training material used for the monolingual systems. The monolingual systems use a monolingual lexicon and language model, while the bilingual systems uses the bilingual resources detailed in Section 5.3.

For the recognition experiments, we create two different training setups with a single-step and two-step DNN training using varying amounts of speech data. Details of all training setups are presented

Table 2. 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
Training	Retraining								
Oracle LID	-	34.2	27.9	47.1	**34.7**	32.7	23.8	51.0	**33.2**
fy-nl	-	34.8	44.6	48.0	39.8	32.4	39.7	49.9	36.2
fy-nl+	-	35.0	43.6	46.7	39.4	32.1	38.6	48.8	35.7
fy-nl++	-	37.9	31.8	47.1	37.8	34.9	25.5	51.0	35.0
fy-nl++-en	-	39.1	31.2	47.3	38.4	35.8	25.6	51.8	35.7
fy-nl++-en	fy-nl	32.8	38.9	43.8	**36.4**	29.9	35.4	45.9	**33.1**
fy-nl++-en	fy-nl+	33.5	37.3	43.8	**36.4**	29.9	33.4	48.3	**33.0**
fy-nl++-en	fy-nl++	37.2	30.1	45.9	36.8	34.2	25.0	49.9	34.3

in Table 1. The single-step training is performed on fy-nl, fy-nl+, fy-nl++ and fy-nl++-en to assess the influence of varying amounts of the high-resourced languages on the recognition performance. For retraining purposes, we use fy-nl, fy-nl+ and fy-nl++ data for similar purposes. The proposed ASR system is 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 an 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 chose the best performing system for performing word-level CS detection experiments. 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 language model, we have generated the ASR output for each utterance in the mixed (fy-nl) segments and the language tags are aligned after removing the words in the reference transcription and ASR output. The CS detection accuracy is evaluated by reporting the equal error rates (EER) calculated based on the detection error tradeoff (DET) graph [43] plotted for the language tag detection with and without the deletions and insertions introduced due to the ASR errors. The presented code-switching detection results indicate how well the recognizer can detect the switches and hypothesize words in the switched language.

5.3. Lexicon and Language Model

The words in the multilingual lexicon are chosen 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 language model training. This corpus includes Frisian, Dutch and mixed sentences yielding a bilingual language model. 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 [44, 45]. 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 [46] for creating phonetic transcriptions. The out-of-vocabulary (OOV) rates in the Frisian development and test (FR) set are 3.2% and 2.6% respectively. The OOV rates in the complete development and test set (FR-NL) are 2.7% and 2.3%.

The bilingual text corpus contains 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. The Dutch text consists of the orthographic transcriptions of the CGN and the Dutch component of the FAME! training data. The bilingual language models are 3-gram with interpolated Kneser-Ney smoothing trained using the SRILM toolkit [47]. Because the bilingual LM is obtained by mostly concatenating monolingual text data, code switches effectively have to go though unigram back-off during decoding. This language model has a perplexity of 259 on the Frisian development set.

5.4. Implementation Details

The recognition experiments are performed using the Kaldi ASR toolkit [48]. 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. The monolingual and multilingual 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 and retraining 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 [49].

[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

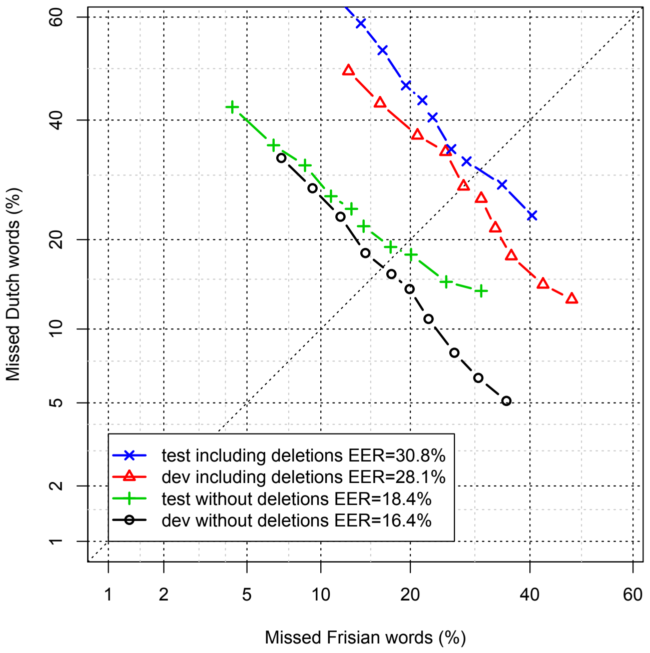

Fig. 3. Word-level code-switching detection evaluation obtained on the code-switching (fy-nl) segments in the FAME! development and test sets

6. RESULTS AND DISCUSSION

6.1. ASR experiments

The recognition results obtained on the development and test sets of the FAME! speech database are given in Table 2. The overall WER of the baseline and the best performing multilingual DNN system is marked in bold. The upper panel of this table presents the number of Frisian and Dutch words in each subset in order to give an impression of the language priors. The baseline recognizer using the oracle LID information has a WER of 34.7% on the development set and 33.2% on the test set. The recognizers with a single-step multilingual DNN training perform significantly worse than the baseline with the best system yielding a WER of 37.8% and 35.0% on the development and test set respectively.

Retraining the multilingual DNN on the target languages improves the recognition performance in all scenarios. The best performing system which is trained on fy-nl+ provides a WER of 36.4% on the development set and 33.0% on the test set. The recognizers retrained on fy-nl and fy-nl+ data provide a similar recognition performance. As expected, further increasing the amount of the included Dutch training data reduces the recognition accuracy on the Frisian only and mixed utterances. It can be concluded that the overall recognition performance of the retrained DNN models do not regularly benefit from increasing the amount of training data belonging to the high-resourced target language. This is presumably due to the high-resource language getting too high a prior in the acoustic modeling.

In general, these results shed light on how well all-in-one CS approaches can perform compared to a recognizer that incorporates a LID system in the front-end and uses the appropriate acoustic modeling based on the LID information. The best performing multilingual DNN recognizer provides comparable recognition accuracies (34.7% vs 36.4% on the development set and 33.2% vs 33.0% on the test set) compared to an ideal recognizer using the oracle LID information.

6.2. CS detection experiments

Evaluating the detection performance at the word level is not as trivial as in whole-utterance detection which is commonly done in language and speaker recognition. For instance, deleted words do not have a language tag, so they may be counted as misses. On the other hand, one might argue that at least these are not false alarms. Therefore we include metrics both ignoring and including the deletions. In our ASR experiments we operated at about 3% insertion and 10% deletion rate.

The DET curves of the best performing multilingual DNN system on the code-switching segments (fy-nl) are plotted in Figure 3. Each point on these curves is obtained for a different language model weight and the EERs for each curve are given in the legend. The CS detection accuracy is higher on the development data with an EER of 16.4% excluding the deletions and 28.1% with the deletions. The EER values obtained on the test set are 18.4% and 30.8% without and with deletions respectively.

Considering the challenging nature of this detection task, the multilingual DNN recognizer has achieved a promising detection accuracy even using a simple bilingual language model which is not tailored for modeling the language switches. The improved bilingual acoustic modeling due to the two-step multilingual training described in Section 4.2 accounts for the accurate detection of the language switches and recognition of the words in the switched language.

7. CONCLUSIONS

In this paper, we use multilingual DNN to recognize code-switching Frisian speech and detect code-switching in the utterances containing both Frisian and Dutch. The multilingual training approach is performed in two steps: i) an initial training phase using the multilingual speech data from target languages and other closely related high-resourced languages and ii) retraining of the shared hidden layers learned in the initial phase on a smaller amount of speech data only from the target languages. The retraining stage is performed using bilingual data taking the bilingual nature of code-switching speech. The recognition results have shown that multilingual training of DNNs provides comparable recognition accuracies on code-switching Frisian speech compared to an ideal recognizer using oracle language identification information. Moreover, the best performing multilingual DNN provides encouraging code-switching detection accuracies using only a primitive bilingual language model.

Future work includes developing language models that can capture code-switching more accurately and investigating the lexicon-free ASR approaches to be able to recognize the *mixed-words* which appear in neither the Frisian nor the Dutch lexicon.

8. 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] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

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

[18] 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.

[19] K. M. Knill, M.J.F. Gales, S.P. Rath, P.C. 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.

[20] Ngoc Thang 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.

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

[22] 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.

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

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

[25] A. P. Versloot, *Mechanisms of language change. Vowel reduction in 15th century West Frisian*, Ph.D. thesis, University of Groningen, 2008.

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

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

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

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

[30] 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 radio broadcast in Frisian designed for code-switching research," in *Proc. LREC*, 2016.

[31] G.E. Dahl, Dong Yu, Li Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE TASLP*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[32] Dong Yu and Li Deng, *Automatic Speech Recognition: A Deep Learning Approach*, Springer-Verlag London, 2015.

[33] G. Hinton, "A practical guide to training restricted boltzmann machines," Tech. Rep. UTML TR 2010003, Department of Computer Science, University of Toronto, 2010.

[34] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural Networks, 1989. IJCNN., International Joint Conference on*, 1989, pp. 593–605 vol.1.

[35] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[36] D. Imseng, P. Motlicek, H. Bourlard, and P. N. Garner, "Using out-of-language data to improve an under-resourced speech recognizer," *Speech Communication*, vol. 56, pp. 142 – 151, 2014.

[37] S. S. Juan, L. Besacier, B. Lecouteux, and M. Dyab, "Using resources from a closely-related language to develop ASR for a very under-resourced language: A case study for Iban," in *Proc. INTERSPEECH*, 2015, pp. 1270–1274.

[38] 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.

[39] D. A. Van Leeuwen and R. Orr, "Speech recognition of non-native speech using native and non-native acoustic models," in *Workshop on Multi-lingual Interoperability in Speech Technology (MIST)*, 1999, pp. 27–32.

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

[41] D. Graff and J. Alabiso, "1996 English Broadcast News Speech (HUB4) LDC97T22," 1997, Philadelphia: LDC.

[42] J. Fiscus, J. Garofolo, M. Przybocki, W. Fisher, and D. Pallett, "1997 English Broadcast News Speech (HUB4) LDC98S71," 1998, Philadelphia: LDC.

[43] 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.

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

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

[46] 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.

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

[48] 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.

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