Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 159 – 166

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Investigating Bilingual Deep Neural Networks for Automatic Recognition of Code-switching Frisian Speech

Emre Yılmaz*, Henk van den Heuvel, David van Leeuwen

CLS/CLST, Radboud University, Nijmegen, Netherlands

Abstract

In this paper, a code-switching automatic speech recognition (ASR) system built for the Frisian language is described. Frisian is mostly spoken in the province Fryslân which is located in the north of the Netherlands. The native speakers of Frisian are mostly bilingual and often code-switch in daily conversations due to the extensive influence of the Dutch language. In the scope of the FAME! Project, the influence of this unforeseen language switching on modern ASR systems will be investigated with the objective of building a robust recognizer that can handle this phenomenon. For this purpose, in this work, we design a bilingual deep neural network (DNN)-based ASR system and investigate the impact of bilingual DNN training in the context of code-switching speech.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: Automatic speech recognition, bilingual DNN, code-switching, low-resourced languages, Frisian

1. Introduction

Language interaction in minority languages spoken in multilingual countries has been researched in the field of linguistics for more than 30 years[1,2,3]. This interaction occurs in the form of phonological, morphological, syntactic and lexical changes consequent to various linguistic phenomena such as word borrowing, interference and relexification. One prominent mechanism induced in the interacting languages is code-switching (CS) which is defined as the continuous alteration between two languages in a single conversation.

CS is mostly noticeable in some minority languages influenced by the majority language or majority languages that have been influenced by globally influential languages such as English and French. Despite the well-established research line in linguistics, robustness of speech-to-text systems against CS and other kinds of language switches have recently received some interest resulting in some robust acoustic modeling[4,5,6,7,8] and language modeling[9,10,11] approaches for CS speech.

* Corresponding author. Tel.: +31243612055
E-mail address: e.yilmaz@let.ru.nl

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.044

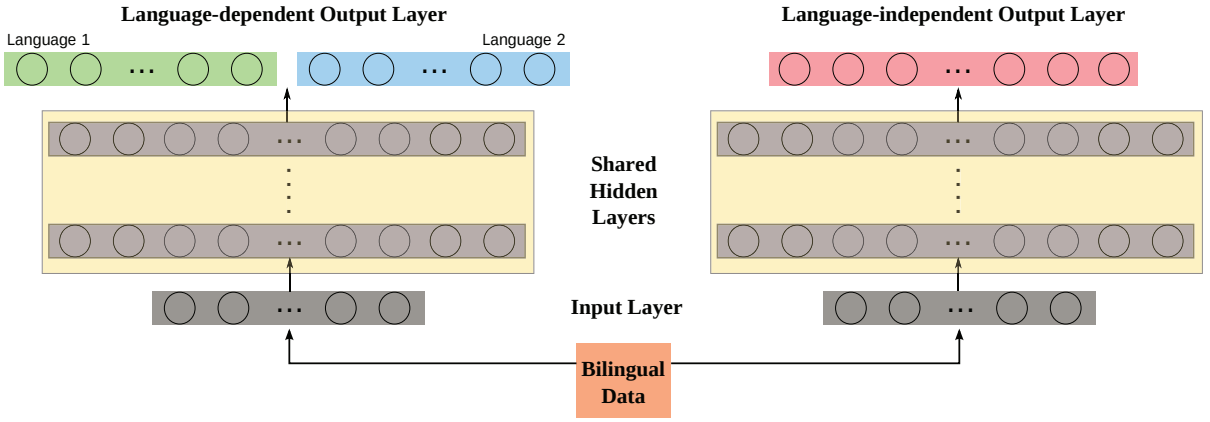

Fig. 1. Language-dependent and language-independent bilingual DNN architecture

Language identification and/or diarization is a relevant task for the automatic speech recognition (ASR) of CS speech [12,13,14,15]. 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, single-pass ASR approaches, which do not directly incorporate a language identification system, have also been proposed yielding promising results [5,8].

Multilingual training of deep neural network-based ASR systems has provided some improvements in the recognition accuracies for both low- and high-resourced languages [16,17,18,19,20,21,22,23,24]. Some of these techniques incorporate multilingual DNNs for feature extraction [25,16,26,21]. Training DNN-based acoustic models on multilingual data to obtain more reliable posteriors for the target language has also been investigated [19,20,24]. ASR systems using DNN-HMM models can employ language-independent or language-dependent phonetic alphabets depending on the purpose of the multilingual training.

In this work, we describe a bilingual deep neural network (DNN)- based ASR system which is designed to recognize both Frisian and Dutch. By investigating different bilingual DNN architecture, we aim to get more insight into developing a more versatile acoustic modeling scheme coping with the language switches. Specifically, both phone-dependent and phone-independent bilingual DNN training approaches are applied on the novel Frisian database and the recognition performance of both systems is compared in order to have a better understanding how far a CS ASR system can benefit from phone merging and cross-language knowledge transfer by using shared hidden layers. To the best of our knowledge, the impact of bilingual DNN training on CS speech data has never been explored. To be able to make a fair comparison of both architecture, the proposed ASR system uses a bilingual lexicon and a bilingual language model trained on a text corpus containing Frisian, Dutch and mixed-language sentences.

This paper is organized as follows. Section 2 summarizes the novel Frisian database that has recently been collected for CS and longitudinal speech research. Section 3 summarizes the DNN-based ASR and Section 4 describes the CS ASR system. The experimental setup is described in Section 5 and the recognition results are presented in Section 6. Section 7 concludes the paper.

2. FAME!: Frisian Radio Broadcast Database

The FAME! speech database has been collected in the scope of the Frisian Audio Mining Enterprise Project. This project aims to build a spoken document retrieval system for the disclosure of the archives of Omrop Fryslân (Frisian Broadcast) covering a large time span from 1950s to the present and a wide variety of topics. Omrop Fryslân is the regional public broadcaster of the province Fryslân with a radio broadcast archive containing more than 2600 hours of recordings. The FAME! speech database contains a small subset of these radio broadcasts and it is the first spoken Frisian database of high recording and annotation quality. The recordings selected for the database

are chosen to include language switching cases, speaker diversity and have a large time span (1966-2015). The content of the recordings are very diverse, including radio programs about culture, history, literature, sport, nature, agriculture, politics and society and languages. The longitudinal and bilingual nature of the material enables research into language variation in Frisian over years, formal versus informal speech, dialectology, CS trends, speaker tracking and diarization over a large time period. This database will be the main resource for learning the acoustic models of the ASR system described below for the Frisian language.

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. To be able to assess the impact of background noise and/or music on the recognition accuracy, the segments containing background noise/music are also labeled. In order to get more precise information about the speaker details, the meta-information of all available radio broadcasts is also provided together with the recordings during the annotation. For further details, we refer the reader to [27].

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 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 after narration (*between-speaker switches*). The former type is in line with the definition of CS phenomenon in linguistics, while the latter occurs due to the broadcast nature of the data. Both type of switches pose a challenge to the ASR systems and have to be handled carefully during recognition.

3. Fundamentals of DNN-based ASR

A single artificial neuron, which is the basic element of the DNN structure, receives N input values $\mathbf{v} = [v_0, v_1, ..., v_{N-1}]$ with weights $\mathbf{w} = [w_0, w_1, ..., w_{N-1}]$ and an offset value b. To compute the neuron output y, a non-linear function f is applied the weighted sum z of all outputs of the previous layer and the offset, i.e., $y = f(z) = f(\mathbf{w}^T\mathbf{v} + b)$. A 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 $\mathbf{v}_l = f(\mathbf{z}_l) = f(\mathbf{W}_l\mathbf{v}_{l-1} + \mathbf{b_l})$ where the dimensions of $\mathbf{v}_l$, $\mathbf{W}_l$, $\mathbf{v}_{l-1}$ and $\mathbf{b_l}$ 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

$$\mathbf{v}_{L+1} = P(q_i|\mathbf{o}) = \frac{e^{z_L^i}}{\sum_m^{M_{L+1}} e^{z_L^m}}, \tag{1}$$

where M_{L+1} is equal to the number of HMM states.

The DNN-HMM training scheme applied in this paper is achieved in three main stages [28,29]. 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 [30] is applied to obtain a robust initialization for the DNN model. Finally, the back-propagation algorithm [31] is applied to train the DNN that will be used as the emission distribution of the HMM states.

4. Bilingual training of DNN

Bilingual training of DNN can be achieved either using language-dependent or language-independent phones as the DNN targets. Both architecture are visualized in Figure 1. The architecture on the left side of the figure shows a language-dependent output layer where the units associated with each language are separated from each other. For this purpose, disjoint phonetic alphabets are used for different languages. The architecture on the right has a single

Table 1. Phonemes of Frisian and Dutch

	Consonants	Vowels			Total
		Mono.	Dip.	Tri.	
Frisian	20	20	24	6	70
Dutch	22	13	3	-	38

output layer where the units are associated with phones from a global phonetic alphabet covering the phone sets of both languages. The shared hidden layers are language-independent in both architecture.

4.1. Using language-dependent phones

As a multitask learning scheme for DNNs, bilingual DNN training with separate phone sets provides better feature representations at the shared hidden layers for both languages by combining the information learned from both information sources[29]. In this setting, the phones of each language are modeled separately. This is achieved by appending a language id to every phone of a word based on the language of its lexicon. The DNNs are trained on either spectral features or features obtained by applying language-independent transformations, e.g., mel-frequency cepstral coefficients (MFCC), that allow the cross-lingual knowledge transfer. Various training stratagies have been described for tuning the DNN models for the target language[18,19,20].

4.2. Using language-independent phones

One goal of using a language-independent output layer is to merge the phonetic units of a low-resource language with a high-resource language and increase the available number of data for each phone[32,33,23,34]. This approach requires mapping the phones of different languages into a common global phonetic alphabet. This mapping can be achieved in multiple ways[35,36,37]. A common approach is to merge data from multiple languages by mapping the phones that are associated with the same International Phonetic Alphabet (IPA) symbol. Moreover, various clustering approaches has also been used for discovering the phones that can be merged by evaluating the distance between the acoustic models learned on these phones.

In this work, we perform phone merging based on the associated IPA symbol. The phonemes of Frisian and Dutch are summarized in Table 1. Frisian phonetic alphabet consists of 20 consonants, 20 monophthongs, 16 falling diphthongs, 8 rising diphthongs and 6 triphthongs. Dutch consonants are similar to the Frisian consonants, while Frisian has more vowels compared to Dutch which has 13 monophthongs and 3 diphthongs. Phonologically, it is reasonable to assume that Dutch vowels are a subset of the Frisian vowels.

5. Experimental setup

We perform ASR experiments on our novel CS Frisian database to investigate the recognition performance provided by bilingual DNN training. We used two databases in the experiments. The FAME! speech database, which is described in Section 2, comprises radio broadcasts in Frisian and Dutch languages. This database will be used for training, development and testing purposes. The other database used in the experiments is the Dutch Broadcast database[38]. This database is used only for training purposes aiming to increase the amount of available Dutch data.

5.1. Databases

The training data of FAME! speech database comprises 8.5 hours and 3 hours of speech from Frisian and Dutch speakers respectively. The development and test sets consist of 1 hour of speech from Frisian speakers and 20 minutes of speech from Dutch speakers each. The Dutch Broadcast database contains 17.5 hours of Dutch data. All data has a sampling frequency of 16 kHz.

The total number of word- and sentence-level Frisian-Dutch CS cases in the FAME! speech database is equal to 3837. These switches are mostly performed by the Frisian speakers as they often use Dutch words or sentences while

Table 2. Word error rates in % obtained on the FAME! development and test sets

	Language-independent				Language-dependent			
	Devel		Test		Devel		Test	
Train. data	WER(%)	CS-WER(%)	WER(%)	CS-WER(%)	WER(%)	CS-WER(%)	WER(%)	CS-WER(%)
FR	39.2	69.2	36.4	65.5	**38.4**	76.0	**36.3**	77.3
FR-NL	38.9	59.7	37.0	**59.3**	40.2	**62.4**	36.9	**70.0**
FR-NL+	**38.8**	**59.5**	**36.1**	60.3	40.3	63.0	36.5	71.5

speaking in Frisian. These cases comprise about 75.6% of the 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. The training, development and test sets contain 2756, 671 and 410 language switching cases.

There are 309 identified speakers in the FAME! speech database, 21 of whom appear at least 3 times in the database. These speakers are mostly program presenters and famous people appearing multiple times in different recordings over years. There are 233 unidentified speakers due to lack of meta-information.

5.2. Lexicon and Language Model

The bilingual lexicon consists of words which are both present in the initial Frisian (340k entries) or Dutch (1.5M entries) lexicons and the bilingual text corpus used for language model training. In pilot experiments, modeling all Frisian vowels on monophthong level has provided the best recognition performance. Therefore, all diphthongs and triphthongs are modeled as a combination of their monophthong constituents.

The bilingual lexicon contains 110k Frisian and Dutch words which appear both in the initial lexicons and the text corpus. The number of entries in the lexicon is 160k due to the words with multiple phonetic transcriptions. The phonetic transcriptions are learned by applying grapheme-to-phoneme bootstrapping[39,40] for the Frisian and Dutch words in the training data which do not appear in the initial lexicons. We use the Phonetisaurus G2P system[41] for learning phonetic transcriptions based on the already existing words in the initial lexicons. The out-of-vocabulary rates in the development and test set are 3.4% and 2.8% respectively.

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[42] and FAME! training data. The bilingual language models are 3-gram with interpolated Kneser-Ney smoothing trained using the SRILM toolkit[43]. This language model has a perplexity of 259 on the FAME! development set.

5.3. ASR experiments

We perform ASR experiments on the development and test data only from the Frisian speakers, since most CS occurs during these recordings. In this way, we aim to get insight about the ASR performance on *within-speaker switches*. The baseline system for this scenario is the monolingual ASR system trained only on the data uttered by Frisian speakers (FR). We compare this system with the bilingual systems trained on the combined Frisian and Dutch data in the FAME! data (FR-NL) and FAME! and Dutch Broadcast databases (FR-NL+). The amount of the Dutch training data for each scenario is 3 and 20.5 hours of speech respectively.

The recognition experiments are performed using the Kaldi ASR toolkit[44]. We train a conventional GMM-HMM system with 25k 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 language-dependent and language-independent bilingual DNNs with 5 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-discriminative training using a state-level minimum Bayes risk (sMBR) criterion[45].

We adopt two performance measures to quantify the recognition performance of the ASR system, namely the Word Error Rate (WER) and Code-Switching WER (CS-WER). The latter performance measure is the ratio of the number of erroneously recognized switched words to the total number of switched words. Here, the switched words include the Dutch and Frisian words spoken in Frisian and Dutch context respectively as well as the *mixed-words* which are defined in Section 5.1.

6. Results and Discussion

The recognition results obtained on the development and test sets of the FAME! speech database are presented in Table 2. For each column, the best results are marked in bold. The WER and CS-WER results obtained using language-independent and language-dependent DNNs are shown on the right and left panels respectively. The results on the development and test data follow a similar trend. The baseline system trained on FR data has a WER of 39.2% and a CS-WER of 69.2% on the development set. The systems trained on FR-NL and FR-NL+ perform better at the recognition of switched words providing a CS-WER of 59.7% and a CS-WER of 59.5% on the development set. The general performance of these systems are also marginally better with a WER of 38.8% and 38.9 compared to 39.2% of the system only trained on FR data. From these results, it can be concluded that merging the phones of both languages helps the recognition of the switched words without significantly reducing the recognition accuracy obtained on the target Frisian language.

The results obtained using language-dependent phones are presented on the right panel. Similarly, adding Dutch training data also improves the recognition of switched words significantly in this scenario. However, the general performance of the system trained on FR is better than the performance of bilingually trained systems. The system trained on FR yields a WER of 38.4% compared to 40.2 of the system trained on FR-NL and 40.3% of the system trained on FR-NL+. This results show that jointly training Frisian and Dutch data using separate phones does not improve the general recognition accuracy. Performing language adaptation is not viable in this case, since the target speech contains words from both languages unlike the previous work using multilingually trained DNNs fine-tuned on a single language. This limitation is expected to reduce the recognizer performance on the target language.

Finally, we will compare the results obtained using language-dependent and language-independent phones. The language-depen-dent system trained on FR provides a much higher CS-WER compared to the language-independent system. This implies that not merging the phones of the switched words with Frisian words in the training data has a detrimental impact on the recognition of switched words. This can be compensated by adding Dutch training data which reduces the CS-WER from 76.0% to 62.4% which is still higher than the 59.7% of the language-independent system trained on the same data.

7. Conclusion

We have presented a bilingual DNN-based ASR system trained for the recognition of code-switching Frisian speech. Recognition of this type of speech is challenging due to unexpected language switches between Frisian and Dutch languages. Two different DNN architecture with language-independent and language-dependent targets are investigated by performing recognition experiments on a novel Frisian broadcast database. The results have demonstrated that bilingual DNN training by merging the phones of both languages provides the best recognition performance on the switched words yielding a CS-WER of 59.5% and a WER of 38.8% on the development set of the FAME! speech database.

Acknowledgements

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

References

1. Auer, P.. *Code-switching in Conversation: Language, Interaction and Identity*. London, Routledge; 1998.
2. Muysken, P.C.. *Bilingual Speech: A Typology of Code-mixing*. Cambridge University Press; 2000.
3. Thomason, S.. *Language Contact - An Introduction*. Edinburgh University Press; 2001.
4. Stemmer, G., Nöth, E., Niemann, H.. Acoustic modeling of foreign words in a german speech recognition system. In: *Proc. EUROSPEECH*. 2001, p. 2745–2748.
5. Lyu, D.C., Lyu, R.Y., Chiang, Y.C., Hsu, C.N.. Speech recognition on code-switching among the Chinese dialects. In: *Proc. ICASSP*; vol. 1. 2006, p. 1105–1108.
6. Vu, N.T., Lyu, D.C., Weiner, J., Telaar, D., Schlippe, T., Blaicher, F., et al. A first speech recognition system for Mandarin-English code-switch conversational speech. In: *Proc. ICASSP*. 2012, p. 4889–4892. doi:10.1109/ICASSP.2012.6289015.
7. Modipa, T.I., Davel, M.H., De Wet, F.. Implications of sepedi/english code switching for ASR systems. In: *Pattern Recognition Association of South Africa*. 2015, p. 112–117.
8. Lyudovyk, T., Pylypenko, V.. Code-switching speech recognition for closely related languages. In: *Proc. SLTU*. 2014, p. 188–193.
9. Li, Y., Fung, P.. Code switching language model with translation constraint for mixed language speech recognition. In: *Proc. COLING*. 2012, p. 1671–1680.
10. Adel, H., Vu, N., Kraus, F., Schlippe, T., Li, H., Schultz, T.. Recurrent neural network language modeling for code switching conversational speech. In: *Proc. ICASSP*. 2013, p. 8411–8415.
11. Adel, H., Kirchhoff, K., Telaar, D., Vu, N.T., Schlippe, T., Schultz, T.. Features for factored language models for code-switching speech. In: *Proc. SLTU*. 2014, p. 32–38.
12. Weiner, J., Vu, N.T., Telaar, D., Metze, F., Schultz, T., Lyu, D.C., et al. Integration of language identification into a recognition system for spoken conversations containing code-switches. In: *Proc. SLTU*. 2012.
13. Lyu, D.C., Chng, E.S., Li, H.. Language diarization for code-switch conversational speech. In: *Proc. ICASSP*. 2013, p. 7314–7318.
14. Yeong, Y.L., Tan, T.P.. Language identification of code switching sentences and multilingual sentences of under-resourced languages by using multi structural word information. In: *Proc. INTERSPEECH*. 2014, p. 3052–3055.
15. Mabokela, K.R., Manamela, M.J., Manaileng, M.. Modeling code-switching speech on under-resourced languages for language identification. In: *Proc. SLTU*. 2014, p. 225–230.
16. Thomas, S., Ganapathy, S., Hermansky, H.. Multilingual MLP features for low-resource LVCSR systems. In: *Proc. ICASSP*. 2012, p. 4269–4272.
17. Swietojanski, P., Ghoshal, A., Renals, S.. Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In: *Proc. SLT*. 2012, p. 246–251.
18. Heigold, G., Vanhoucke, V., Senior, A.W., Nguyen, P., Ranzato, M., Devin, M., et al. Multilingual acoustic models using distributed deep neural networks. In: *Proc. ICASSP*. 2013, p. 8619–8623.
19. Huang, J.T., Li, J., Yu, D., Deng, L., Gong, Y.. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In: *Proc. ICASSP*. 2013, p. 7304–7308. doi:10.1109/ICASSP.2013.6639081.
20. Ghoshal, A., Swietojanski, P., Renals, S.. Multilingual training of deep neural networks. In: *Proc. ICASSP*. 2013, p. 7319–7323.
21. Tuske, Z., Pinto, J., Willett, D., Schluter, R.. Investigation on cross- and multilingual MLP features under matched and mismatched acoustical conditions. In: *Proc. ICASSP*. 2013, p. 7349–7353.
22. Knill, K.M., Gales, M., Rath, S., Woodland, P., Zhang, C., Zhang, S.X.. Investigation of multilingual deep neural networks for spoken term detection. In: *Proc. ASRU*. 2013, p. 138–143.
23. Vu, N.T., Imseng, D., Povey, D., Motlicek, P., Schultz, T., Bourlard, H.. Multilingual deep neural network based acoustic modeling for rapid language adaptation. In: *Proc. ICASSP*. 2014, p. 7639–7643. doi:10.1109/ICASSP.2014.6855086.
24. Das, A., Hasegawa-Johnson, M.. Cross-lingual transfer learning during supervised training in low resource scenarios. In: *Proc. INTERSPEECH*. 2015, p. 3531–3535.
25. Vu, N.T., Metze, F., Schultz, T.. Multilingual bottle-neck features and its application for under-resourced languages. In: *Proc. SLTU*. 2012, p. 1–4.
26. Vesely, K., Karafiat, M., Grezl, F., Janda, M., Egorova, E.. The language-independent bottleneck features. In: *Proc. SLT*. 2012, p. 336–341.
27. Yılmaz, E., Andringa, M., Kingma, S., Van der Kuip, F., Van de Velde, H., Kampstra, F., et al. A longitudinal radio broadcast in Frisian designed for code-switching research. In: *Proc. LREC*. 2016.
28. Dahl, G., Yu, D., Deng, L., Acero, A.. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 2012;**20**(1):30–42. doi:10.1109/TASL.2011.2134090.
29. Yu, D., Deng, L.. *Automatic Speech Recognition: A Deep Learning Approach*. Springer-Verlag London; 2015.
30. Hinton, G.. A practical guide to training restricted boltzmann machines. Tech. Rep. UTML TR 2010003; Department of Computer Science, University of Toronto; 2010.
31. Hecht-Nielsen, R.. Theory of the backpropagation neural network. In: *Neural Networks, 1989. IJCNN., International Joint Conference on*. 1989, p. 593–605 vol.1.
32. Imseng, D., Motlicek, P., Bourlard, H., Garner, P.N.. Using out-of-language data to improve an under-resourced speech recognizer. *Speech Communication* 2014;**56**:142 – 151.
33. Ragni, A., Knill, K.M., Rath, S.P., Gales, M.J.F.. Data augmentation for low resource languages. In: *Proc. INTERSPEECH*. 2014, p. 810–814.
34. Sahraeian, R., Van Compernolle, D., de Wet, F.. Using generalized maxout networks and phoneme mapping for low resource ASR- a case study on Flemish-Afrikaans. In: *Pattern Recognition Association of South Africa*. 2015, p. 112–117.

35. Schultz, T., Waibel, A.. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication* 2001;**35**(12):31 – 51.
36. Köhler, J.. Multilingual phone models for vocabulary-independent speech recognition tasks. *Speech Communication* 2001;**35**(12):21 – 30.
37. Yu, S., Zhang, S., Xu, B.. Chinese-English bilingual phone modeling for cross-language speech recognition. In: *Proc. ICASSP*; vol. 1. 2004, p. 917–920.
38. Van Leeuwen, D.A., Orr, R.. Speech recognition of non-native speech using native and non-native acoustic models. In: *Workshop on Multi-lingual Interoperability in Speech Technology (MIST)*. 1999, p. 27–32.
39. Davel, M., Barnard, E.. Bootstrapping for language resource generation. In: *Pattern Recognition Association of South Africa*. 2003, p. 97–100.
40. Maskey, S.R., Black, A.B., Tomokiyo, L.M.. Bootstrapping phonetic lexicons for new languages. In: *Proc. ICLSP*. 2004, p. 69–72.
41. Novak, J.R., Minematsu, N., Hirose, K.. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering* 2015;:1–32.
42. Oostdijk, N.. The spoken Dutch corpus: Overview and first evaluation. In: *Proc. LREC*. 2000, p. 886–894.
43. Stolcke, A.. SRILM – An extensible language modeling toolkit. In: *Proc. ICSLP*. 2002, p. 901–904.
44. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. The Kaldi speech recognition toolkit. In: *Proc. ASRU*. 2011.
45. Vesely, K., Ghoshal, A., Burget, L., Povey, D.. Sequence-discriminative training of deep neural networks. In: *Proc. INTERSPEECH*. 2013, p. 2345–2349.