Multilingual Neural Network Acoustic Modelling for ASR of Under-Resourced English-isiZulu Code-Switched Speech

Astik Biswas[1], *Febe de Wet*[1], *Ewald van der Westhuizen*[1], *Emre Yılmaz*[2,3], *Thomas Niesler*[1]

[1] Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
[2]CLS/CLST, Radboud University, Nijmegen, Netherlands
[3] Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore

abiswas@sun.ac.za, fdw@sun.ac.za, ewaldvdw@sun.ac.za, e.yilmaz@let.ru.nl, trn@sun.ac.za

Abstract

Although isiZulu speakers code-switch with English as a matter of course, extremely little appropriate data is available for acoustic modelling. Recently, a small five-language corpus of code-switched South African soap opera speech was compiled. We used this corpus to evaluate the application of multilingual neural network acoustic modelling to English-isiZulu code-switched speech recognition. Our aim was to determine whether English-isiZulu speech recognition accuracy can be improved by incorporating three other language pairs in the corpus: English-isiXhosa, English-Setswana and English-Sesotho. Since isiXhosa, like isiZulu, belongs to the Nguni language family, while Setswana and Sesotho belong to the more distant Sotho family, we could also investigate the merits of additional data from within and across language groups. Our experiments using both fully connected DNN and TDNN-LSTM architectures show that English-isiZulu speech recognition accuracy as well as language identification after code-switching is improved more by the incorporation of English-isiXhosa data than by the incorporation of the other language pairs. However additional data from the more distant language group remained beneficial, and the best overall performance was always achieved with a multilingual neural network trained on all four language pairs.
Index Terms: code-switching, under-resourced languages, African languages, speech recognition, DNN, TDNN-LSTM.

1. Introduction

With 11 official languages whose usage patterns overlap geographically, South Africa has a highly multilingual population. As a consequence, it is common to use more than one language during discourse. This phenomenon is known as code-switching (CS) and can occur between sentences, within the same sentence, and even within the same word [1, 2]. In South Africa, English is widespread and can be regarded as a common denominator among languages. It is, however, not the most frequently used mother tongue by some margin. As a consequence, code-switching between English and the other languages permeates the daily conversations of South Africans. Automatic speech recognition (ASR) systems deployed in this environment should therefore be able to process multilingual speech that includes such code-switching.

Although most state-of-the-art ASR systems are monolingual, the automatic recognition of speech including code switching has recently received increased attention [2–5]. In comparison with monolingual speech, code-switching in spontaneous speech is highly unpredictable and difficult to model. Despite recent advances in ASR achieved by the application of neural networks, the success for code-switched speech has been limited by the particular challenges this presents to acoustic [1,6] and language modelling [7]. These challenges are even more acute when the languages concerned are under-resourced [6,8]. In South Africa, code-switching is prevalent between English, a highly-resourced language, and the nine official African languages, which are all under-resourced.

Two main strategies to deal with code-switching in ASR have been described in the literature. The first incorporates language identification (LID) into the speech processing pipeline [9–11]. The LID component first labels speech frames and monolingual ASR is subsequently used to perform decoding. This approach has the advantage of simplicity, since conventional acoustic and language modelling methods, which achieve excellent monolingual performance, can be employed. However, language identification is a difficult task, especially in the presence of intra-word or intra-sentential code switching, and LID error propagation will lead to poor ASR by the monolingual recognisers.

The second strategy is to perform single pass ASR that does not depend on LID [2,12]. This has the advantage of not requiring an explicit a-priori LID and can therefore in principle avoid the errors necessarily associated with incorrect LID. It does, however, require new methods of language and acoustic modelling which explicitly model and allow the abrupt language changes at a code switch during the recognition pass. Training such acoustic and language models requires data, which is scarce for code-switched speech.

In this work, we investigate whether improved acoustic modelling can be achieved by the application of multilingual neural network training approaches to English-isiZulu code-switched speech. We build on a first study on these languages in which it was reported that language dependent acoustic modelling outperformed language independent acoustic modelling, but the reported word error rates were very high ($> 80\%$) [1]. Our aim was to determine whether the recognition performance for English-isiZulu code-switched speech could be improved by leveraging additional code-switched data by means of multilingual neural network architectures. Both deep neural networks (DNNs) and time delay neural network - long short-term memory (TDNN-LSTM) networks were considered for this purpose. Specifically we considered whether code switched data from other languages can be useful for acoustic modelling. In addition, since some of the languages we consider are related, we evaluate the relative merits of multilingual modelling within and across these language families.

2. Corpus Details

A multilingual corpus containing examples of code-switched speech has been compiled from 626 South African soap opera

episodes. The ELAN media annotation tool [13] was used to segment and annotate the data. The soap opera speech is typically fast, spontaneous and often expresses emotion. The spontaneous nature of the speech and the presence of a wide variety of code-switching makes it a challenging corpus to investigate ASR performance.

The corpus is still under development and the version we used corresponds to the 14.3 hour language-balanced set introduced in [14]. The data contains examples of code-switching between South African English, isiZulu, isiXhosa, Setswana and Sesotho. Of the four Bantu languages, isiZulu and isiXhosa belong to the Nguni (N) language family while Sesotho and Setswana are Sotho (S) languages. An overview of the statistics for the training (*Train*), development (*Dev*) and test (*Test*) sets for each language pair is given in Table 1. Each data set is described in terms of its total duration as well as the duration of the monolingual (m) and code-switched (c) segments.

Table 1: *Duration in hours (h) or minutes (m) of English, isiZulu, isiXhosa, Setswana, Sesotho segments in monolingual (emdur, zmdur, xmdur, tmdur, smdur) and code-switched (ecdur, zcdur, xcdur, tcdur, scdur) utterances.*

English-isiZulu (E-Z)					
Set	**emdur**	**zmdur**	**ecdur**	**zcdur**	**tot dur**
Train	93m	93m	45.86m	56.99m	4.81h
Dev	0	0	4.01m	3.96m	8m
Test	0	0	12.76m	17.85m	30.4m
Total	93m	93m	62.40m	78.60m	5.45h
English-isiXhosa (E-X)					
Set	**emdur**	**xmdur**	**ecdur**	**xcdur**	**tot dur**
Train	65.22m	53.55m	18.04m	23.73m	2.67h
Dev	2.86m	6.48m	2.21m	2.13m	13.68m
Test	0	0	5.56m	8.78m	14.34m
Total	68.08m	60.03m	25.81m	34.64m	3.143h
English-Setswana (E-T)					
Set	**emdur**	**tmdur**	**ecdur**	**tcdur**	**tot dur**
Train	40.4m	30.96m	34.37m	34.01m	2.33h
Dev	0.76m	4.26m	4.54m	4.27m	13.83m
Test	0	0	8.87m	8.96m	17.83m
Total	41.16m	35.22m	47.78m	47.24m	2.86h
English-Sesotho (E-S)					
Set	**emdur**	**smdur**	**ecdur**	**scdur**	**tot dur**
Train	49.34m	35.32m	23.02m	34.04m	2.36h
Dev	1.09m	5.05m	3.03m	3.59m	12.77m
Test	0	0	7.80m	7.74m	15.54m
Total	50.43m	40.37m	33.85m	45.37m	2.83h

Two types of code-switching occur in our data:

1. **Intersentential code-switching**: Language alters between utterances of a conversation.
2. **Intrasentential code-switching**: Language alters within a single utterance. This can be further sub-divided into the following three categories, with *English*-isiZulu examples from our corpus:

   - **Alternation**: Structurally independent stretches of English and isiZulu, e.g.:
     "*he is a fighter* ufuzo ubaba wakhe".
   - **Insertion**: An English language element is incorporated into the structure of isiZulu, e.g.:
     "ubekwa yini la *late* kangaka".
   - **Intraword**: When the isiZulu affixes are used with the English items to form a word, e.g.:
     "wena u-*feel*-a kanjani vele ngaye".

A total of 10 343 code-switched utterances and 19 207 intrasentential language switches are observed in the corpus. Note that the test utterances always contain code switching and are never monolingual. There are 734 and 3 199 isiZulu tokens in the development and test sets respectively. The corresponding counts for English are 838 and 2 459.

The duration values in Table 1 show that all the language pairs in our corpus are under-resourced. It has been found that the amount of training data has a significant influence on the robustness of ASR systems [6, 15, 16]. However, very little additional speech data is available for the languages under investigation. The data that is available originates from different domains and is poorly matched to our corpus. Out-of-corpus data was therefore not considered in this investigation. Instead, we evaluated the effect on system performance of combining different subsets of the training sets listed in Table 2.

Table 2: *Training set configurations.*

Training set	**Composition**
1 CS pair	English-isiZulu
2 CS pair N-N	English-isiZulu + English-isiXhosa
2 CS pair N-S	English-isiZulu + English-Sesotho
4 CS pair	English-isiZulu + English-isiXhosa + English-Sesotho + English-Setswana

3. Neural Networks for Acoustic Modelling

Neural networks trained on multilingual data have recently achieved substantial performance improvements over systems based on monolingual acoustic models for various target languages [15–19]. It has been shown that the hidden layers of a neural network can extract acoustic information useful for improved modelling from closely related languages [2, 15, 20, 21]. The ability to effect such cross-lingual information transfer makes multilingual training especially attractive when dealing with under-resourced speech recognition. The following sections provide a brief overview of the neural networks used for acoustic modelling in our experiments.

3.1. Multilingual DNNs

The concept behind multilingual DNNs (MDNN) is to use resources from other languages to develop acoustic models for an under-resourced language. The general MDNN framework is illustrated in Figure 1.

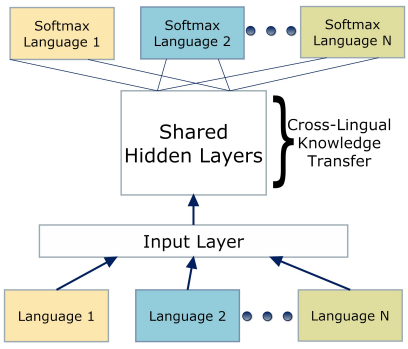

Figure 1: *General framework for Multilingual DNN acoustic modelling.*

The MDNN is trained with a relatively large corpus from multiple languages to compute class conditional hidden Markov model (HMM) posterior probabilities. A DNN can be thought of as a cascaded sequence of feature extractors followed by a

softmax layer. In the MDNN architecture, the hidden layers are shared across all languages, but each language has an individual softmax layer [15].

3.2. TDNN-LSTM Neural Networks

Recently, TDNN [22, 23] and LSTM [24] topologies have been found to yield further improvements in speech recognition performance. The sub-sampling mechanism employed by TDNNs significantly reduces model training time. Further improvements are possible by using a lattice-free maximum mutual information (LF-MMI) training criterion [25]. In addition, the interleaving of temporal convolution and LSTM layers has been shown to effectively model future temporal context [26]. We compared the performance of these TDNN-LSTM acoustic models with multilingual DNNs in our experiments.

4. Experiments

All ASR experiments were performed using the Kaldi ASR toolkit [27, 28] and the data described in Section 2. Throughout, the development set was used to optimize hyper-parameters while test data was reserved for final evaluation.

4.1. Acoustic Modelling

4.1.1. GMM-HMMs

The baseline English-isiZulu ASR system was a conventional triphone Gaussian mixture model (GMM)-HMM system with 15K Gaussians using 39-dimensional MFCC features including dynamic (Δ and $\Delta\Delta$) coefficients. The acoustic features were extracted using 25ms Hamming windows with a 10ms frame shift. The GMM-HMM results are used as a point of reference, not for direct comparison with the neural network results.

4.1.2. Multilingual DNNs

DNN-HMM training was performed in three stages [6, 15]. First, a triphone GMM-HMM model was trained to obtain initial HMM transition probabilities and alignments for DNN training. Then a restricted Boltzmann machine (RBM) pre-training algorithm [19, 29] was used to initialize the DNN model. Finally, the DNNs were trained to optimize a cross-entropy objective function using the standard error back-propagation algorithm.

The multilingual training strategy is illustrated in Figure 2. In this scheme, language specific acoustic models are trained and phonemes are not pooled across isiZulu and English. A language identifier appended to the words in each lexicon allows later analysis of the code-switch detection accuracy.

As shown in Figure 2, we initialized the MDNNs with deep belief networks (DBN) [30] pre-trained on two or four pairs of languages. Filter bank spectral features were applied to the input layer to allow cross-lingual knowledge transfer. The performance of DNNs depends on many hyper-parameters, which are very time consuming to optimize individually. Thus, we decided not to optimize the pre-training hyper-parameters for all configurations. To allow a fair comparison across experiments, all DBNs consisted of 10 layers with 1024 hidden units in each layer. This choice was motivated by the results of preliminary experiments. These pre-trained, shared hidden layers were then fine-tuned on the English-isiZulu speech to obtain a softmax layer for the target language.

The networks were trained using 40-dimensional log mel filter-bank features with appended velocity and acceleration. 3-

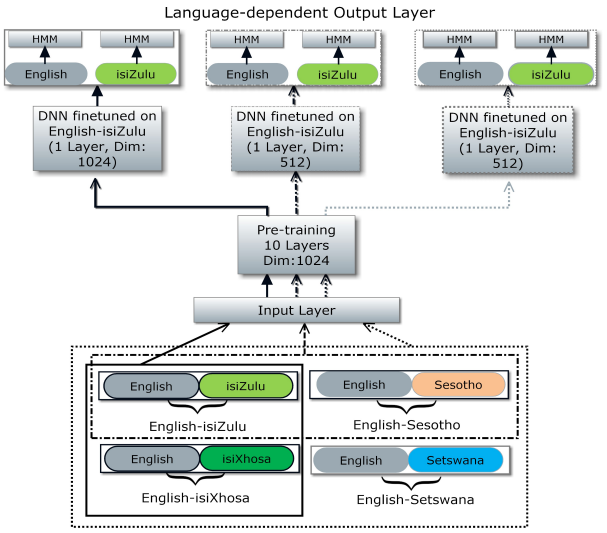

Figure 2: *Multilingual DNN framework for 2 CS pair N-N (solid line _____) , 2 CS pair N-S (dash-dot line _____), and 4 CS pair training sets (dotted line _ _ _ _ _).*

dimensional pitch features were also appended. To include information on the dynamics in the speech signal, a context window of 11 frames (± 5) was used during training.

The DNN training was performed using a stochastic descent algorithm (initial learning rate: 0.008) and a mini-batch size of 256. During training, the data was randomly split into training (90%) and cross-validation (10%) partitions. The learning rate was halved when the accuracy on the cross-validation set between two successive epochs fell below 0.5%.

4.1.3. Multilingual TDNN–LSTMs

LF-MMI [25] and TDNN-LSTM [26] (1 standard, 6 time-delay and 3 LSTM layers) acoustic models were trained for each configuration described for DNNs in the previous section. The standard Switchboard Kaldi (ver. 5.2.99) recipe was used to achieve this. The default 40-dimensional MFCC features combined with i-vectors for speaker adaptation were used in combination with the training parameters provided with the recipe without performing any parameter tuning [31]. Three-fold data augmentation was applied to the training data [32].

4.2. Language modelling

The English-isiZulu vocabulary consisted of 11 269 unique word types and was closed with respect to the development and test sets. The SRILM toolkit [33] was used to train and evaluate a bilingual English-isiZulu language model (LM) using the English-isiZulu training set transcriptions. This model was further interpolated with a monolingual South African English LM trained on newspaper and web text of 471 million word tokens, and a monolingual isiZulu LM trained on newspaper, web and conversational transcription texts of 3.2 million word tokens. The LM interpolation weights were optimised using the development set perplexity as performance measure. The word type statistics for the corpus and LM perplexities as evaluated on the development and test set transcriptions are reported in Table 3.

5. Results and Discussion

The English-isiZulu GMM-HMM system achieved 65.65% and 69.35% word error rate (WER) on the development and test sets

Table 3: *The number of word types for the English-isiZulu corpus, as well as the language model perplexity on the development and test sets.*

Corpus set	English word types	IsiZulu word types	Total word types	Perplexity
Train	3608	6765	10373	–
Development	415	443	858	469.3
Test	871	1420	2291	658.9
Total	3842	7425	11269	–

respectively. The corresponding language-specific WERs on the test set for English and isiZulu were 68.93% and 69.69%. The WERs for the English-isiZulu CS ASR systems derived from the training sets in Table 2 and the configurations described in Sections 4.1.2 and 4.1.3 are reported in Table 4.

The DNN results in the table reveal that additional training data from the same language family (N-N) results in a relative improvement of 2.71% compared to the bilingual baseline. The corresponding gain observed with additional training data from a different language family (N-S) is only 2.09%. Finally, comparing rows 1 and 4 of Table 4 indicates a relative improvement of 4.85% for the 4 CS pair system relative to the baseline DNN system.

Table 4: *WER (%) on the English-isiZulu dev and test sets.*

	Dev		Test		Test English		Test isiZulu	
Train Set	DNN HMM	TDNN LSTM	DNN HMM	TDNN LSTM	DNN HMM	TDNN LSTM	DNN HMM	TDNN LSTM
1 CS Pair (Baseline)	58.46	55.22	64.48	60.53	63.68	56.65	65.28	63.52
2 CS Pair N-N	55.73	52.35	62.73	58.13	61.19	54.05	64.27	61.27
2 CS Pair N-S	58.59	52.16	63.13	58.64	61.76	54.66	64.49	61.71
4 CS Pair	56.87	47.39	61.35	55.76	59.29	50.06	63.40	60.14

When compared with the DNNs, the TDNN-LSTMs show a similar, although more pronounced, trend in WER improvements. The TDNN-LSTM system for 4 CS pairs shows a 7.88% relative improvement in WER compared to the TDNN-LSTM baseline. Despite the observed gains in WER, none of the systems were able to achieve a WER below 55%. This could be due to insufficient training data or because of a weak language model. It should be borne in mind that there are no monolingual sentences at all in the test and development sets (Table 1). Instead, these data sets contain only CS speech and language modelling across switch boundaries is known to be challenging [8].

For better insight, Table 4 includes language-specific WERs derived from the corresponding language-specific deletions, insertions and substitutions. We see that the English WER improves as more English CS data is added for DNN training. However, the isiZulu WER deteriorates when languages from a different family are included during training.

Further analysis of the TDNN-LSTM ASR output is shown in Table 5. We see that that the word correct accuracy improves for both English and isiZulu when different CS language pairs are added to the pool of training data. The analysis also reveals a substantial improvement in word accuracy at the 1 464 CS points occurring in the test data for both the 2-pair N-N and 4-pair TDNN-LSTM systems. Furthermore, it is interesting to note that, relative to the 2-pair N-S model, the 2-pair N-N model shows a greater improvement in word and language correct at the CS points. Although the English-isiXhosa training data is roughly 19 minutes longer than the English-Setswana data, the results seem to suggest that additional data from the same language family (in this case isiXhosa) has a larger impact on recognition accuracy. The same behaviour is seen for the DNN-HMM acoustic models. Thus, we conclude that data from the same language family is most useful for acoustic modelling.

Table 5: *Results analysis for different TDNN-LSTM systems on the English-isiZulu test set described in Table 1. All values are percentage accuracy (%) except the first line which corresponds to the WER (%) from Table 4. (Eng: English, Zul: isiZulu)*

	1 CS Pair (baseline)	2 CS pair N-N	2 CS pair N-S	4 CS Pair
WER	60.53	58.13	58.64	55.76
Eng words correct	44.9	47.74	47.21	51.89
Zul words correct	38.73	41.13	40.01	42.60
Eng insertions	1.13	0.86	0.81	1.24
Zul insertions	0.81	1.27	0.97	1.17
Eng deletions	6.30	5.57	6.47	4.74
Zul deletions	8.83	8.02	9.40	8.11
Eng to Eng substitutions	10.76	9.40	9.38	9.17
Eng to Zul substitutions	6.87	7.74	7.09	7.00
Zul to Zul substitutions	16.56	18.68	17.23	17.80
Zul to Eng substitutions	9.24	6.57	7.28	6.54
Words correct after CS	40.16	43.10	42.89	45.29
Eng Words correct after CS	43.94	45.36	45.88	48.07
Zul Words correct after CS	35.90	40.55	39.53	42.15
Language correct after CS	65.57	69.05	67.89	69.46

6. Conclusions

This paper presents the results of an investigation aimed at improving the automatic recognition of English-isiZulu code-switched speech. Four different MDNN and TDNN-LSTM based systems were developed and evaluated using multilingual code-switched speech extracted from South African soap operas. The recognition systems were trained with language dependent acoustic models and a language independent lexicon. These experiments represent the first multilingual acoustic models trained on multiple code-switched datasets, and the first such investigation for African languages.

We find that additional training data from other code-switched language pairs improves the WER for English-isiZulu when compared with a baseline system trained only on the target language pair. Furthermore, system performance benefits most when the additional training data originates from a closely related language, in our case another Nguni language rather than a Sotho language. However, the best overall performance was achieved when data from all four code-switched language pairs were combined, even though some of the additional languages are more distantly related to the target language. In particular, the accuracy with which the word immediately after a language switch is recognised is observed to improve when data from all four language pairs is included in the acoustic model. This indicates that multilingual neural network training is able to capture acoustic knowledge from other languages pairs that benefits the recognition of code-switched speech.

Despite the fact that our best system shows a WER improvement of 7.88% relative to the baseline, recognition performance still requires further enhancement. Future work will focus on using additional monolingual speech data for training. An attempt will also be made to extend the pool of available data by means of automatic segmentation and transcription.

7. Acknowledgements

We would like to thank the Department of Arts & Culture of the South African government for funding this research and Stellenbosch University for the travel grant that enabled Dr Yılmaz to visit Stellenbosch. We are also indebted to Dr Armin Saeb and Dr Raghav Menon for their valuable insights. Finally, we gratefully acknowledge the support of the NVIDIA corporation for the donation of the GPU equipment used during this research.

8. References

[1] E. van der Westhuizen and T. Niesler, "Automatic Speech Recognition of English-isiZulu Code-switched Speech from South African Soap Operas," *Procedia Computer Science*, vol. 81, pp. 121–127, 2016.

[2] E. Yılmaz, H. van den Heuvel, and D. van Leeuwen, "Investigating bilingual deep neural networks for automatic recognition of code-switching Frisian speech," *Procedia Computer Science*, vol. 81, pp. 159–166, 2016.

[3] Y. Li and P. Fung, "Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints," in *Proc. ICASSP*. IEEE, 2013, pp. 7368–7372.

[4] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, and H. Li, "A first speech recognition system for Mandarin-English code-switch conversational speech," in *Proc. ICASSP*. IEEE, 2012, pp. 4889–4892.

[5] T. I. Modipa, M. H. Davel, and F. De Wet, "Implications of Sepedi/English code switching for ASR systems," in *24th Annual Symposium of the Pattern Recognition Association of South Africa*, 2013, pp. 64–69.

[6] E. Yılmaz, H. van den Heuvel, and D. van Leeuwen, "Code-switching detection using multilingual DNNs," in *Proc. SLT*. IEEE, 2016, pp. 610–616.

[7] H. Adel, N. T. Vu, K. Kirchhoff, D. Telaar, and T. Schultz, "Syntactic and semantic features for code-switching factored language models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 431–440, 2015.

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

[9] 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*, 2014.

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

[11] E. Yılmaz, M. McLaren, H. van den Heuvel, and D. A. van Leeuwen, "Language diarization for semi-supervised bilingual acoustic model training," in *Proc. ASRU*. IEEE, 2017, pp. 91–96.

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

[13] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: a professional framework for multimodality research," in *Proc. LREC*, 2006, pp. 1556–1559.

[14] E. van der Westhuizen and T. Niesler, "A First South African Corpus of Multilingual Code-switched Soap Opera Speech," *Proc. LREC*, 2018.

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

[16] A. Saeb, R. Menon, H. Cameron, W. Kibira, J. Quinn, and T. Niesler, "Very low resource radio browsing for agile developmental and humanitarian monitoring," *Proc. Interspeech 2017*, pp. 2118–2122, 2017.

[17] C. Ni, C.-C. Leung, L. Wang, N. F. Chen, and B. Ma, "Efficient methods to train multilingual bottleneck feature extractors for low resource keyword search," in *Proc. ICASSP*. IEEE, 2017, pp. 5650–5654.

[18] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *Proc. ICASSP*, 2017, pp. 4930–4934.

[19] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language processing*, vol. 20, no. 1, pp. 30–42, 2012.

[20] S. Zhou, Y. Zhao, S. Xu, and B. Xu, "Multilingual Recurrent Neural Networks with Residual Learning for Low-Resource Speech Recognition," *Proc. Interspeech 2017*, pp. 704–708, 2017.

[21] R. Sahraeian and D. Van Compernolle, "Using weighted model averaging in distributed multilingual DNNs to improve low resource ASR," *Procedia Computer Science*, vol. 81, pp. 152–158, 2016.

[22] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, 2015, pp. 3214–3218.

[23] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[24] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.

[25] 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.

[26] 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, 2018.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

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

[29] M. Eastwood and C. Jayne, "Restricted Boltzmann machines for pre-training deep Gaussian networks," in *Proc. IJCNN*. IEEE, 2013, pp. 1–8.

[30] A.-R. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

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

[32] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015.

[33] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Proc. ICSLP*, 2002.