Improving ASR for code-switched speech in under-resourced languages using out-of-domain data

Astik Biswas, Ewald van der Westhuizen, Thomas Niesler & Febe de Wet

Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa

abiswas@sun.ac.za, ewaldvdw@sun.ac.za, trn@sun.ac.za, fdw@sun.ac.za

Abstract

We explore the use of out-of-domain monolingual data for the improvement of automatic speech recognition (ASR) of code-switched speech. This is relevant because annotated code-switched speech data is both scarce and very hard to produce, especially when the languages concerned are under-resourced, while monolingual corpora are generally better-resourced. We perform experiments using a recently-introduced small five-language corpus of code-switched South African soap opera speech. We consider specifically whether ASR of English–isiZulu code-switched speech can be improved by incorporating monolingual data from unrelated but larger corpora. TDNN-BLSTM acoustic models are trained using various configurations of training data. The utility of artificially-generated bilingual English–isiZulu text to augment language model training data is also explored. We find that English-isiZulu speech recognition accuracy can be improved by incorporating monolingual out-of-domain data despite the differences between the soap-opera and monolingual speech.

Index Terms: Code-switching, under-resourced languages, African languages, Bantu languages, acoustic modelling, speech recognition, TDNN-BLSTM, RNN.

1. Introduction

South Africa is a highly multilingual country and code-switching (CS) is common in everyday conversations. Among the 11 official languages, English serves as the *lingua franca*, leading to frequent switching between it and the other languages. Code-switching is usually observed in spontaneous speech, which is generally fast and accented. This presents a challenging scenario for automatic speech recognition (ASR), especially when one or more of the languages are under-resourced.

Automatic speech recognition of code-switched speech is a topic of current research interest [1–5]. For example, a framework for English-Mandarin spontaneous code-switched speech recognition is proposed in [1]. Different phone merging approaches were explored for acoustic modelling, and the application of statistical machine translation (SMT) to the generation of code-switched text for language modelling was evaluated. Large vocabulary English–Mandarin code-switched speech recognition was also considered in [2]. In this case a bilingual acoustic model was developed to reflect the changes in accent associated with code-switching while the points at which language change occurs were predicted by the language model. A study focussing on bilingual Dutch-Frisian code-switched ASR was reported in [4]. Here the application of deep neural network (DNN) architectures in the acoustic model was extensively considered. The English-Sepedi code-switched corpus presented in [3] was the first to include an African language. Subsequently, a GMM-HMM based speech recognizer was presented for English-isiZulu code-switched speech in [5]. Four different system configurations were investigated, with different combinations of language dependent or independent acoustic and language models.

Most broadcast programmes generally contain prepared, monolingual speech. In contrast, South African soap operas exhibit fast, spontaneous speech which contains extensive code-switching. We have recently compiled a corpus containing 14.3 hours of language-balanced code-switched speech drawn from such soap operas. The corpus is arranged into four language pairs: English-isiZulu, English-isiXhosa, English-Setswana, English-Sesotho. Our first attempts at acoustic modelling using this data considered using the different language pairs to improve the recognition of English-isiZulu code-switched speech [6]. We found that using additional training data from other code-switched language pairs improved the performance of a baseline system trained only on the target language pair (English-isiZulu). Furthermore, we found that the acoustic models benefited most from additional training data in closely related languages. However, all these experiments were limited to the soap opera corpus, which is limited in size and uniform in terms of style and character.

In this study we aimed to determine whether further improvements in the ASR performance can be achieved for English-isiZulu code-switched speech by incorporating additional out-of-domain monolingual speech. The speech in question was taken from a substantially larger corpus of prompted monolingual speech and is therefore poorly matched to our soap opera data [7].

2. Speech data

Studies have shown that the amount of training data strongly influences the accuracy and robustness of ASR systems [8–10]. However, code-switched speech data is difficult to collect and annotate. This is in part due to the spontaneous nature of code-switched speech and in part due to the poorly-understood mechanisms underlying the switching between languages. These factors complicate the development of prompts and the elicitation of suitable natural utterances. Furthermore, manual annotation of code-switched speech is difficult, time-consuming and requires more specialised linguistic skills than it does for monolingual speech. Consequently, very little in-domain speech data is available for the language pairs in question.

2.1. South African code-switched soap opera corpus

The English-isiZulu code-switched speech is part of a corpus that has been compiled from 626 episodes of three different multilingual South African soap operas [11]. This corpus includes five South African languages: isiZulu, isiXhosa, Setswana, Sesotho and English. Of these, isiZulu and isiXhosa belong to the Nguni (N) language family while Sesotho and Setswana are Sotho (S) languages. Table 1 shows the total duration as well as the duration of the monolingual and code-switched utterances in the training set of each language pair. It also gives an overview of the training (*Train*), development (*Dev*) and test (*Test*) sets used in this study. The test set comprises approximately 10% of the data selected across all episodes such that there is no overlap between the speakers in the training and test sets. The test set was also constrained to contain no monolingual utterances, but only utterances with intrasentential code-switching. The training sets contained a balanced combination of monolingual and code-switched utterances. As far as possible the data was segmented such that each utterance contains a complete sentence.

Table 1: *Duration in hours (h) and minutes (m) of English, isiZulu, isiXhosa, Setswana, Sesotho in monolingual (Eng, Zul, Xho, Tsn, Sot) and in code-switched (Eng_CS, Zul_CS, Xho_CS, Tsn_CS, Sot_CS) utterances.*

English-isiZulu (EngZul)					
Set	**Eng**	**Zul**	**Engl_CS**	**Zul_CS**	**Total**
Train	93m	93m	45.86m	56.99m	4.81h
Dev	0	0	4.01m	3.96m	8m
Test	0	0	12.76m	17.85m	30.40m
Total	93m	93m	62.40m	78.60m	5.45h
English-isiXhosa (EngXho)					
	Eng	**Xho**	**Eng_CS**	**Xho_CS**	**Total**
Train	65.22m	53.55m	18.04m	23.73m	2.67h
English-Setswana (EngTsn)					
	Eng	**Tsn**	**Eng_CS**	**Tsn_CS**	**Total**
Train	40.4m	30.96m	34.37m	34.01m	2.33h
English-Sesotho (EngSot)					
	Eng	**Sot**	**Eng_CS**	**Sot_CS**	**Total**
Train	49.34m	35.32m	23.02m	34.04m	2.36h

Table 1 shows that all subcorpora are under-resourced. The soap opera speech is also typically fast, spontaneous and expresses emotion. Hence it is a challenging corpus for ASR.

Examples of intrasentential code-switching from the corpus include alternation (structurally independent stretches of English and isiZulu), insertion (an English language element is incorporated into the structure of isiZulu) and intraword switches (isiZulu affixes are used with the English items to form a word). There are a total of 10 343 code-switched utterances and 19 207 intrasentential language switches in the corpus. Note that the test set contains only code-switching utterances. The word type and token counts for the English-isiZulu training, development and test partitions are provided in Table 2.

Table 2: *Number of word types and tokens in the training, development and test sets of the English-isiZulu code-switched corpus.*

	English		**isiZulu**		**Total**	
Data set	Tokens	Types	Tokens	Types	Tokens	Types
Train	28 033	3 608	24 350	6 765	52 383	10 373
Development	838	415	734	443	1 572	858
Test	2 459	871	3 199	1 420	5 658	2 291
Total	31 330	3 842	28 283	7 425	59 613	11 269

2.2. NCHLT corpus

The NCHLT speech corpus contains monolingual wide-band prompted speech in each of the 11 official languages of South Africa. A greedy algorithm was used to select the prompts from a body of text during the compilation of the corpus [12]. Trigram prompts were used for English and the Nguni languages while five-gram prompts were used for the Sotho languages.

The NCHLT corpus contains approximately 50 hours of speech gathered from around 200 speakers in each language.

Table 3: *Statistics of the NCHLT speech data used in this study.*

Language	**Duration (h)**	**# Speakers**	**Tokens**	**Types**
isiZulu	52.22	201	116 319	23 912
isiXhosa	53.15	201	122 236	27 856
English	54.18	202	205 392	8 215

The pre-defined NCHLT English, isiZulu and isiXhosa training data sets, shown in Table 3, were used in our experiments. Although the nature of the NCHLT speech is quite different from the soap opera speech, it is one of the only other annotated speech corpora available for these languages. For this reason we were interested in investigating its potential to improve the performance of our code-switched speech recognizer. We included isiXhosa in our experiments since it is a close relative of isiZulu and may therefore be useful for acoustic modelling.

3. Text data

The three text data sets listed in Table 4 were used for language modelling. The in-domain text was derived from the soap opera corpus training set transcriptions. The English-isiZulu vocabulary of 11 269 word types was closed with respect to the training, development and test sets. Additional out-of-domain text was sourced from monolingual English and isiZulu South African newspaper reports, web text and the transcriptions of conversations.

Table 4: *Text sources used for language modelling.*

Text	**Type**	**English tokens**	**isiZulu tokens**
In-domain	Bilingual	28 033	24 350
English	Monolingual	471M	-
isiZulu	Monolingual	-	3.2M

In a related study on ASR for Frisian-Dutch code-switched speech, it was found that using additional LSTM-generated text reduced the language model perplexity [13]. We therefore considered similarly supplementing our real text data with text generated artificially using an LSTM trained on the in-domain English-isiZulu code-switched text.

4. Acoustic modelling

Recent work in acoustic modelling has established that time delay neural network (TDNN) [14, 15] and long short-term memory LSTM [16] acoustic model topologies can yield substantial improvements in speech recognition accuracy in comparison with deep neural networks (DNNs) [6, 16]. LSTMs use memory cells in the hidden layers instead of the conventional activation functions employed by feedforward networks, allowing training problems associated with vanishing and exploding gradients to be overcome and long-range temporal dependencies to be learnt. The sub-sampling mechanism employed by TDNNs significantly reduces model training time, and further improvements are possible by using a lattice-free, maximum mutual information (LF-MMI) training criterion [17].

The performance benefits of TDNN-LSTM acoustic models for English-isiZulu code-switched ASR have already been established [6]. Here we extend the architecture to bidirectional LSTMs (BLSTMs) which process input data in both time directions using two separate hidden layers [18]. This type of architecture allows the preservation of both past and future context information. Furthermore, the interleaving of temporal convolution and BLSTM layers has been shown to effectively model future temporal context [19].

ASR experiments were performed using the Kaldi ASR toolkit (version 5.2.164) [20]. As a first step the training sets of all the relevant languages were combined to form a single pool of training data. Then, a conventional context-dependent Gaussian mixture model-HMM (GMM-HMM) acoustic model with 25k Gaussians was trained using 39 dimensional mel-frequency cepstral coefficient (MFCC) features including deltas and delta-deltas. This GMM-HMM model was used to obtain the alignments required for neural network training.

The same pool of training data was used to derive acoustic features for neural network training. However, in this case three-fold data augmentation was applied prior to feature extraction [21]. The acoustic features included MFCCs (40-dimensional, without derivatives), pitch features (3-dimensional) and i-vectors for speaker adaptation (100-dimensional).

LF-MMI TDNN-BLSTM acoustic models with 1 standard, 2 time-delay and 3 BLSTM layers were trained for various training set configurations. No parameter tuning was performed for neural net training, but default parameters of the standard Switchboard Kaldi recipe were used [22]. This resulted in a set of multilingual acoustic models, which were subsequently subjected to English-isiZulu code-switched adaptation.

5. Language modelling

The SRILM toolkit [23] was used to train and evaluate a bilingual 3-gram language model trained on the English-isiZulu training data transcriptions. This language model was interpolated with monolingual English and isiZulu LMs derived from the texts in Table 4. The interpolation weights were optimised on the development set perplexity. An analysis of the perplexity of this language model on monolingual and code-swiched text is shown in Table 5.

Table 5: *Detailed perplexity analysis of the bilingual English-isiZulu trigram language model when applied to the development and test sets described in Table 2. Perplexities for language switches indicate the uncertainty of the first word following the switch.*

	Dev set	Test set
All text	425.82	601.69
All language switches in text	2 852.48	3 291.95
All English to isiZulu language switches	3 354.82	3 834.99
All isiZulu to English language switches	2 467.88	2 865.41
All monolingual text	258.28	358.08
All English monolingual text	133.17	121.15
All isiZulu monolingual text	558.00	777.76

We see that the perplexity for monolingual English is much lower than the corresponding value for monolingual isiZulu. This is not surprising, given the difference in size between the two monolingual corpora, as shown in Table 4. The perplexity is particularly high at language switches.

Additional language model training sets containing 2M, 5M and 10M words were generated using an LSTM, as described in Section 3. Two versions of the additional text data were used: one "as is" and the other filtered to contain only code-switched utterances. The filtered versions of the data contained 1.1M, 2M and 7M words respectively. Six language models were derived from the original and filtered versions of the additional text and interpolated with the existing language models using development set perplexity as a performance metric. The perplexity values obtained in this manner are compared to the baseline values in Table 6.

A comparison between the first four rows of the table reveals a substantial reduction in perplexity for both the development and test sets when the in-domain, bilingual language model is interpolated with the two out-of-domain, monolingual language models. The table also shows that interpolation with the language models derived from the artificial data afforded only marginal perplexity reductions, and that most of this reduction was provided by the artificial text containing code-switching.

Table 6: *Perplexity analysis of different language models on the English-isiZulu development and test sets described in Table 1. (CS: code-switched.)*

	Language Model	Dev set	Test set
1	English-isiZulu CS text	539.6	697.5
2	+ Monoligual English	438.1	621.0
3	+ Monolingual isiZulu	513.1	630.7
4	+ Monoligual English + Monolingual isiZulu	425.8	601.7
5	+ Generated text (2M)	418.4	602.1
6	+ Generated CS text (1.1M)	418.1	603.2
7	+ Generated text (5M)	417.9	596.9
8	+ Generated CS text (2.9M)	419.2	597.9
9	+ Generated text (10M)	416.8	596.5
10	+ Generated CS text (7M)	416.3	596.7

6. ASR experiments and results

The data sets described in Section 2 were used in different configurations to train LF-MMI TDNN-BLSTM acoustic models, as described in Section 4. The results that follow were all obtained using the interpolated bilingual trigram language model described in Section 5 (Table 6, row 4). In contrast to the results obtained for Dutch and Frisian code-switched speech, we observed only marginal changes in WER for the language models incorporating the artificially generated text, despite the observed reduction in perplexity. Perhaps this is due to the comparatively small text dataset we have available for training the LSTM-LM [24].

6.1. Adding out-of-domain training data

ASR performance on the English-isiZulu code-switched test set described in Table 2 is summarised in Figure 1. The first set of bars in Figure 1 correspond to the complete code-switched test set while the second and third sets of bars indicate the language specific WERs calculated on the English and isiZulu words respectively. The legend indicates which NCHLT languages were added to the in-domain soap opera training data.

Despite the difference between the NCHLT and soap opera speech, Figure 1 shows that the acoustic models benefit from the additional training data in every case. The word error rates are reduced substantially by the individual incorporation of the

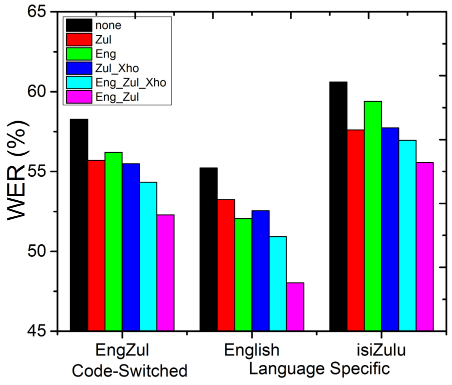

Figure 1: *Word error rates (WER) for English-isiZulu code-switched test data. The legend indicates the NCHLT languages added to the training data.*

NCHLT English and isiZulu data sets, but best results were obtained when both were added together. Compared to the baseline acoustic model trained only on English-isiZulu code-switched speech, this augmentation leads to a 10.27% relative improvement. We also see that the improvement in performance when adding the NCHLT isiZulu and isiXhosa data is almost the same as that achieved when adding only isiZulu. Furthermore, the performance when adding English, isiZulu and isiXhosa from NCHLT is worse than adding only English and isiZulu. We conclude that adding out-of-domain data in a related language (isiXhosa) is not helpful.

6.2. Adding in-domain and out-of-domain training data

In [6] we have shown that additional training data from related code-switched language pairs improved the recognition accuracy of English-isiZulu code-switching speech. The best results were achieved by training the acoustic models on all four code-switched language pairs from the soap opera corpus. This trend is confirmed by the results in Table 7.

Table 7: *WER (%) for English-isiZulu code-switched data: different combinations of in-domain code-switched and out-of-domain monolingual NCHLT data.*

Training data	Dev	Test	English (Test)	isiZulu (Test)
EngZul (baseline)	52.16	58.27	55.23	60.61
+ NCHLT Eng + NCHLT Zul	49.36	52.28	48.03	55.55
EngZul + EngXho	50.06	56.04	51.04	59.89
+ NCHLT Eng + NCHLT Zul	47.08	50.95	45.18	55.39
EngZul + EngXho + EngTsw + EngSot	47.07	53.06	47.62	57.24
+ NCHLT Eng + NCHLT Zul	44.21	50.09	46.24	53.05

The values in Table 7 show that the acoustic model trained on two code-switched language pairs (EngZul + EngXho) yields a relative improvement of 4.02% and 3.82% on the development and test sets respectively compared to the baseline. The table also confirms that, using only in-domain data from the code-switched corpus, the lowest WER for the English-isiZulu test set was achieved when the acoustic models were trained on a combination of all four code-switched training data sets (EngZul + EngXho + EngTsw + EngSot). Similar to the trend observed in Figure 1, the results in Table 7 show that adding additional out-of-domain NCHLT data to the training set improves the recognition performance.

Further analysis of the TDNN-BLSTM ASR output is shown in Table 8. The results confirm that the word correct accuracy improves for both English and isiZulu when more acoustic data is added to the pool of training data. The analysis also reveals a substantial improvement in word accuracy at the 1 464 code-switching points occurring in the test data when additional out-of-domain training data is included in acoustic model training.

Table 8: *Detailed analysis of ASR output when decoding with different acoustic models. All values are percentages (%). (CS: code-switching.)*

	EngZul	+NCHLT Eng_Zul	EngZul +EngXho	+NCHLT Eng_Zul	All four CS pairs	+NCHLT Eng_Zul
English words correct	47.54	54.98	51.08	57.95	54.78	56.53
isiZulu words correct	41.36	45.89	42.08	46.95	45.08	48.73
Words correct after CS	42.96	48.77	44.33	50.55	49.04	51.37
isiZulu Words correct after CS	42.13	45.33	42.34	47.91	44.13	49.38
English Words correct after CS	43.78	50.44	46.61	52.05	50.80	54.16
Language correct after CS	69.81	70.97	68.65	75.07	73.29	74.80

6.3. Balanced addition of in-domain and out-of-domain training data

The results in the previous sections seem to indicate that acoustic modelling for code-switch ASR can be enhanced by using additional, out-of-domain monolingual speech data. However, comparing the values in Tables 1 and 3 reveals that, for most of the languages, the monolingual NCHLT data sets are more than ten times the size of their in-domain code-switched counterparts. To investigate the impact of additional in-domain versus out-of-domain data from a closely related language, we reduced the NCHLT English and Xhosa data to match the EngXho data set in the code-switched corpus. The results of this "balanced" experiment are shown in Table 9.

Table 9: *WER (%) for English-isiZulu test data with additional in-domain English-isiXhosa code-switched data and balanced out-of-domain monolingual NCHLT English and isiXhosa data.*

Training data	Dev	Test	English (Test)	isiZulu (Test)
EngZul	52.16	58.27	55.23	60.61
EngZul + EngXho	50.06	56.04	51.04	59.89
EngZul + NCHLT Eng and Xho	52.35	57.60	55.10	59.52

The results in Table 9 show that the acoustic models benefit more from additional in-domain training data than from an equal amount of out-of-domain training data. The language specific WERs in the last two columns of the table show that, while the English component of the test set benefits substantially from the additional in-domain code-switch data, it does not gain much if a similar amount of NCHLT English data is added to the training set. The corresponding improvement for isiZulu is marginal in both instances. This observation suggests that additional in-domain acoustic data remains a prerequisite for more accurate code-switched ASR.

7. Summary and conclusion

This paper presents a study aimed at improving the automatic recognition of under-resourced English-isiZulu code-switched speech by using the out-of-domain monolingual speech found in the NCHLT speech corpus. TDNN-BLSTM based systems were developed and evaluated using multilingual code-switched speech extracted from South African soap operas and NCHLT speech. The recognition systems were trained with language dependent acoustic models and language independent lexica.

The results of the investigation show that the out-of-domain data has good potential to improve the ASR performance of under-resourced code-switched speech. We found that the addition of out-of-domain speech improves the word error rate substantially for code-switched English-isiZulu when compared with a baseline system trained only on in-domain code-switched speech. Although the out-of-domain speech is monolingual and prompted, and therefore dissimilar in character to the in-domain spontaneous code-switched soap opera speech, acoustic models still benefited from the additional data.

Despite these improvements, error rates remain high and further enhancement is required. Future work will focus on expanding the code-switched data by means of automatic segmentation and transcription, as well as using monolingual speech of both closely and distantly related languages.

8. Acknowledgements

We would like to thank the Department of Arts & Culture of the South African government for funding this research. We also gratefully acknowledge the support of the NVIDIA corporation for the donation of the GPU equipment used during this research.

9. References

[1] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4889–4892.

[2] Y. Li and P. Fung, "Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7368–7372.

[3] 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.

[4] 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.

[5] 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.

[6] A. Biswas, F. de Wet, E. van der Westhuizen, E. Yılmaz, and T. Niesler, "Multilingual Neural Network Acoustic Modelling for ASR of Under-Resourced English-isiZulu Code-Switched Speech," in *Interspeech*, 2018 (Accepted).

[7] E. Barnard, M. H. Davel, C. v. Heerden, F. de Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[8] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7319–7323.

[9] E. Yılmaz, H. van den Heuvel, and D. van Leeuwen, "Code-switching detection using multilingual DNNs," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 610–616.

[10] 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.

[11] E. van der Westhuizen and T. Niesler, "A first South African corpus of multilingual code-switched soap opera speech," *Proc. Language Resources and Evaluation Conference. 2018*, pp. 2854–2859, 2018.

[12] R. Eiselen and M. J. Puttkammer, "Developing text resources for ten south african languages." in *LREC*, 2014, pp. 3698–3703.

[13] E. Yılmaz, H. v. d. Heuvel, and D. A. v. Leeuwen, "Acoustic and Textual Data Augmentation for Improved ASR of Code-Switching Speech," in *Interspeech*, 2018 (Accepted).

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Z. Yu, V. Ramanarayanan, D. Suendermann-Oeft, X. Wang, K. Zechner, L. Chen, J. Tao, A. Ivanou, and Y. Qian, "Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 338–345.

[19] 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.

[20] 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 *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

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

[22] 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.

[23] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[24] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Interspeech*, 2012.