Building a Unified Code-Switching ASR System for South African Languages

Emre Yılmaz\textsuperscript{1,2}, Astik Biswas\textsuperscript{3}, Ewald van der Westhuizen\textsuperscript{3}, Febe de Wet\textsuperscript{3} and Thomas Niesler\textsuperscript{3}

\textsuperscript{1}CLS/CLST, Radboud University, Nijmegen, Netherlands
\textsuperscript{2}Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore
\textsuperscript{3}Dept. of Electrical and Electronic Engineering, Stellenbosch University, South Africa

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

Abstract

We present our first efforts towards building a single multilingual automatic speech recognition (ASR) system that can process code-switching (CS) speech in five languages spoken within the same population. This contrasts with related prior work which focuses on the recognition of CS speech in bilingual scenarios.

The ASR system presented in this work is trained on 4 corpora containing English-isiZulu, English-isiXhosa, English-Setswana and English-Sesotho CS speech. The interpolation of multiple language models trained on these language pairs enables the ASR system to hypothesize mixed word sequences from these 5 languages. We evaluate various state-of-the-art acoustic models trained on this 5-lingual training data and report ASR accuracy and language recognition performance on the development and test sets of the South African multilingual soap opera corpus.

Index Terms: code-switching, automatic speech recognition, multilinguality, South African languages

1. Introduction

South Africa is a multilingual society with 11 official languages and since the majority of South Africans are multilingual, code-switching (CS) occurs commonly in everyday conversations. CS being a part of daily life, the phenomenon also commonly occurs in radio and TV broadcasts which makes broadcast archives valuable sources of CS speech data. We have recently compiled a CS speech corpus which contains 14.3 hours of language-balanced speech compiled from soap opera broadcasts. Our research focuses on designing the acoustic and language models that can operate on this type of multilingual CS speech.

The impact of CS and other kinds of language switches on the performance of speech-to-text systems has recently received research interest, resulting in several robust acoustic modeling \cite{18, 19} and language modeling \cite{10, 11} approaches for CS speech. In previous work, the Radboud team has explored the ASR and code-switching detection performance of various acoustic models applied to CS Frisian-Dutch speech \cite{10, 11, 12, 13}. We further proposed several ways of increasing the amount of available training speech data by applying several automatic transcription strategies \cite{12, 13}.

Meanwhile, the Stellenbosch team has been developing a CS ASR system for isiZulu-English CS speech with a focus on language modeling \cite{12, 13}. The improvements in this system obtained by using speech data from different CS language pairs are explored in another submission to this conference \cite{19}.

In this work, we describe our joint efforts to build a 5-lingual ASR system that can recognize all five languages present in the South African Soap Opera Corpus, namely English, isiZulu, isiXhosa, Setswana and Sesotho. For this purpose, we develop acoustic and language models that enable language switches between these five languages. Unlike the prior work focusing on bilingual CS scenarios, this ASR system can hypothesize CS word sequences containing all five languages.

We first use monolingual and bilingual text to train a language model with words from all target languages. The only source of CS text is the transcriptions of the training speech data which contain 156k words in total. Then, we train several recently proposed acoustic models on the different CS pairs present in the corpus and apply these to the combination of all development and test data. We report 5-lingual ASR accuracies and language recognition (LR) performance of the developed ASR system.

This paper is organized as follows. Section 2 introduces the demographics and the linguistic properties of the target South African languages. Section 3 summarizes the South African Soap Opera Corpus that has recently been collected for CS speech research. Section 4 describes the details of the 5-lingual CS ASR system. The experimental setup is described in Section 5 and the ASR and LR performance of the described ASR system is presented in Section 6. Section 7 concludes the paper.

2. Target South African Languages

The linguistic and phonetic properties of the target indigenous languages are given in the following paragraphs. All four languages belong to Southern Bantu language family. isiZulu and isiXhosa are Nguni languages and linguistically similar. Furthermore, Sesotho and Setswana belong to the Sotho family and are also linguistically similar. All are agglutinative, tonal and click languages written in the Latin alphabet. The information presented in the coming paragraphs is extracted from the Ethnologue\textsuperscript{1}, UCLA Language Materials Project\textsuperscript{2} and Census document\textsuperscript{3}.

The isiZulu language has 11.5M native and 15.5M second language (L2) speakers mostly living in South Africa. As many other indigenous languages in South Africa, it has relatively recently become a written language. The Zulu phonology is characterized by a simple vowel inventory with 5 vowels and highly marked consonantal system with ejectives, implosives

---

1https://www.ethnologue.com/
2http://www.lmp.ucla.edu/
3http://www.statssa.gov.za

---

This work was performed while the first author was on a research visit to the Stellenbosch University.
and clicks [29]. Zulu has borrowed many words from other lan-
guages, especially Afrikaans and English.

The second language, isiXhosa, has 8M native and 11M L2 speak-
ers in South Africa and Lesotho. It has 9 vowels and 39 conso-
nants including ejectives, clicks and uvular trill. Various sound
changes are observed involving vowels and consonants includ-
ing various sorts of assimilation, elision, vowel merging and
devoicing [21]. The fourth and the final language, Setswana, is
spoken by 5M native and 7.5M L2 speakers in South Africa and
Lesotho. It includes 7 vowels and 29 consonants, 3 of which
contain clicks. There are two tones which are orthograp-
ically not marked. Despite a high mutual intelligibility with Sesotho,
they are generally considered to be two separate languages.

3. South African Soap Opera Corpus
A multilingual corpus containing examples of CS speech has
recently been compiled from 626 South African soap opera
episodes. The ELAN media annotation tool [23] has been used
to segment and annotate the data. The spontaneous nature of the
speech and the presence of various CS types makes this type of
speech interesting for designing an ASR system which is ex-
pected to operate on CS speech from South Africa.

The corpus is still under development and the version we
used corresponds to the language-balanced dataset with 14.3
hours of speech introduced in [23]. The data contains exam-
pies of CS between South African English, isiZulu, isiXhosa,
Setswana and Sesotho. The corresponding code-switch lan-
guage pairs are referred to as English-isiZulu, English-isiXhosa,
English-Setswana, and English-Sesotho. 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 dura-
tion of the monolingual (m) and CS (cs) segments.

The soap opera speech is typically fast, spontaneous and
may express emotion, with a speech rate that is between 1.22
and 1.83 times higher than prompted speech in the same lan-
guages. Among the 10 343 code-switched utterances in the cor-
pus, 19 207 intra-sentential language switches are observed. In-
tentional code-switching with English words is observed to be
most frequent. Intra-word CS occurring when English words are
supplemented with Bantu affixes in an effort to conform to
Bantu phonology is also observed. Note that the test utterances
always contain CS and are never monolingual.

Table 1: Duration of English, isiZulu, isiXhosa, Setswana,
Sesotho monolingual (emdur, zmndur, xmdur, tmdur, smdur) and
CS (ecsdur, zcsdur, xsdcur, tsdcur, scsdur) utterances [23]

| Language-pair   | emdur | zmdur | xmdur | tmdur | smdur | scsdur | xsdcur | tsdcur | scsdur | Total |
|-----------------|-------|-------|-------|-------|-------|--------|--------|--------|--------|-------|
| English-isiZulu |       |       |       |       |       |        |        |        |        |       |
| Train (h)       | 1.55h | 1.55h | 45.86m| 56.99m| 4.81h |        |        |        |        |       |
| Dev (h)         | -     | -     | 4.01m | 3.96m | 8m    |        |        |        |        |       |
| Test (h)        | -     | -     | 12.76m| 17.85m| 30.4m |        |        |        |        |       |
| Total (h)       | 1.55h | 1.55h | 1.04h | 1.31h | 5.45h |        |        |        |        |       |
| English-isiXhosa|       |       |       |       |       |        |        |        |        |       |
| Train (h)       | 65.22m| 53.55m| 18.04m| 23.73m| 160.54m|        |        |        |        |       |
| Dev (h)         | 2.86m | 6.48m | 2.21m | 2.13m | 13.68m |        |        |        |        |       |
| Test (h)        | -     | 5.36m | 8.78m | 14.34m|        |        |        |        |        |       |
| Total (h)       | 68.08m| 60.03m| 25.81m| 34.64m| 3.14m |        |        |        |        |       |
| English-Setswana|       |       |       |       |       |        |        |        |        |       |
| Train (h)       | 40.4m | 30.96m| 34.37m| 34.01m| 139.74m|        |        |        |        |       |
| Dev (h)         | 0.76m | 4.26m | 4.34m | 4.27m | 13.83m |        |        |        |        |       |
| Test (h)        | -     | 8.87m | 8.96m | 17.83m|        |        |        |        |        |       |
| Total (h)       | 41.16m| 35.22m| 47.78m| 47.28m| 2.80h |        |        |        |        |       |
| English-Sesotho |       |       |       |       |       |        |        |        |        |       |
| Train (h)       | 49.34m| 35.32m| 23.02m| 34.04m| 141.72m|        |        |        |        |       |
| Dev (h)         | 1.09m | 3.05m | 3.05m | 3.39m | 12.77m |        |        |        |        |       |
| Test (h)        | -     | 7.80m | 7.74m | 15.54m|        |        |        |        |        |       |
| Total (h)       | 50.43m| 40.37m| 33.85m| 45.37m| 2.83h |        |        |        |        |       |

4. 5-lingual CS ASR System
With the ultimate goal of an ASR system that can operate on
all South African languages and correctly process the language
switches, we build a language and an acoustic model that con-
siders the word inventory of the five languages of our corpus.
We apply the ideas that have been shown to be useful in mono-
lingual scenarios to our multilingual CS system to determine the
ASR performance that can be obtained using modern acoustic
and language models.

For language modeling, both monolingual and CS text re-
sources were used in varying quantities based on availability.
CS text is practically non-existent and CS in textual resour-
ces hardly occurs. As is common, we use the transcriptions of the
CS speech data for language model training purposes. Since
the CS text constitutes a relatively small component of all avail-
able text data, we merge all available CS text to train a 5-
lingual CS language model. Secondly, all monolingual text
from the African languages is merged to train a 4-lingual LM.
These models are later interpolated with the English monolin-
gual model that has been trained on a much larger monolingual
corpus. Following this strategy provided the lowest perplexities
on the transcriptions of the development data in pilot experi-
ments.

For training the acoustic models, we have only used the bal-
canced corpora without including any other available monolin-
gual corpora. On this small amount of data, we explore the per-
formance of various NN architectures including conventional
fully connected DNNs, time-delay NN (TDNN) [24, 25] and its
combination with recurrent NN architectures such as long short-
term memory (LSTM) and bidirectional LSTM [26] using dif-
ferent acoustic features. Since most of the target languages are
tonal, we also extract and attach pitch information to the acous-
tic features.

Together with the ASR performance, we also evaluate the LR
accuracies of the described ASR systems to gain an insight
into how welI it can distinguish between the target languages,
especially between those most similar (isiZulu-isiXhosa and
Setswana-Sesotho). To achieve this, we analyze the confusion
in the language tags assigned to each word during recognition.
5. Experimental Setup

5.1. Language Modeling

The available monolingual and CS text in total number of words is presented in Table 2. The texts used for monolingual LM training were collected from various sources which include online newspapers, magazines and newsletters (South African English, isiZulu, isiXhosa, Setswana), web text from the Leipzig Corpus Collection [27] (South African English, isiZulu, isiXhosa, Setswana), and the Babel corpus transcriptions (isiZulu).

The language models used in these experiments are 5-lingual 3-gram and 5-gram with interpolated Kneser-Ney smoothing [28] for recognition and lattice rescoring respectively. We interpolate: (1) a CS 3-gram trained on all CS text, (2) a 4-lingual 3-gram trained on all monolingual text from 4 African languages, and (3) an English 3-gram to obtain the final 3-gram LM. The interpolation weights are learned on the transcriptions of the development data. We have observed that assigning relatively higher weights (0.85-0.9) to the CS LM reduces the perplexities considerably. The final 3-gram model has perplexities of 412 and 617 on the development and test sets respectively.

The language models used in these experiments are 5-lingual 3-gram and 5-gram with interpolated Kneser-Ney smoothing [28] for recognition and lattice rescoring respectively. We interpolate: (1) a CS 3-gram trained on all CS text, (2) a 4-lingual 3-gram trained on all monolingual text from 4 African languages, and (3) an English 3-gram to obtain the final 3-gram LM. The interpolation weights are learned on the transcriptions of the development data. We have observed that assigning relatively higher weights (0.85-0.9) to the CS LM reduces the perplexities considerably. The final 3-gram model has perplexities of 412 and 617 on the development and test sets respectively. For the final 5-gram model, the corresponding figures are 402 and 605.

5.2. Acoustic Modeling

The recognition experiments are performed using the Kaldi ASR toolkit [29]. We train a conventional context dependent Gaussian mixture model-hidden Markov model (GMM-HMM) system with 25k Gaussians using 39 dimensional mel-frequency cepstral coefficient (MFCC) features including the deltas and delta-deltas to obtain the alignments for training the NN models.

As a reference, DNNs with 6 hidden layers and 2048 sigmoid hidden units at each hidden layer are trained on the 40-dimensional log-mel filterbank (FBANK) features with the deltas and delta-deltas. DNN training is performed by minibatch 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 [30].

In addition, we train TDNN (3 standard, 6 time-delay layers), TDNN-LSTM (1 standard, 6 time-delay and 3 LSTM layers) and TDNN-BLSTM (1 standard, 2 time-delay and 3 BLSTM layers) acoustic models with the lattice-free maximum mutual information (LF-MMI) criterion [31] according to the standard recipe provided for the Switchboard database in the Kaldi toolkit (ver. 5.2.99). With these models, we use 40-dimensional MFCCs together with 3-dimensional pitch features (appended when mentioned). The training parameters provided in the recipe are used without performing any parameter tuning. The 3-fold data augmentation [32] is applied to the training data.

5.3. Pronunciation Lexicon

The pronunciation lexicon contains 23,453 words of which 5965 are English, 7448 are isiZulu, 5975 are isiXhosa, 1625 are Setswana and 2437 are Sesotho. Due to the pronunciation variants, the lexicon has 30,489 entries in total. The phonetic alphabet contains a total of 284 phones of which 45 are English, 49 are isiZulu, 66 are isiXhosa, 59 are Setswana and 65 are Sesotho. The ASR experiments are closed vocabulary implying that there are no out-of-vocabulary words in the development and test data.

6. Results and Discussion

We present the results of the ASR experiments in this section. The ASR quality is quantified by calculating the word error rate (WER) with the language tags of the words removed. The language recognition performance is later evaluated in the form of a confusion matrix with an extra row and column to accommodate insertion and deletion errors.

6.1. ASR Results

First, we investigate the best performing NN architecture and then we move to a detailed analysis of the recognition accuracy on each language component. The ASR results obtained on the development and test sets using different acoustic models are presented in Table 3.

A conventional fully connected DNN provides a total WER of 66.5%, while using a TDNN model brings considerable improvements reducing the WER to 60.6%. Adding recurrent layers helps by further reducing the total WER to 58.8%. With a WER of 57.2%, the ASR system using bidirectional recurrent layers together with time-delay layers has the lowest WER among all the aforementioned acoustic models.

We further explore the impact of appending pitch features and lattice rescoring using a larger n-gram language model. Using pitch features is common practice when building ASR systems for tonal languages. In this scenario, using pitch features provides an absolute improvement of 1.7% leading to a WER of 56.1%. 5-gram lattice rescoring brings further marginal improvements reducing the total WER to 55.6%. In the remaining experiments, we analyze the output of this ASR system to gain better insight to language-specific ASR performance and language recognition performance.

The language-specific recognition accuracies for each CS pair and dataset are shown in Table 4. From these results, it can be seen that there is a large performance gap between the gen-
The total number of words in each subcorpus is given in parentheses.

| CS Pair (L₁-L₂) | L₁-L₂ Dev | Test | L₁ Dev | Test | L₂ Dev | Test |
|------------------|----------|------|--------|------|--------|------|
| English-isiZulu   | 44.2 (1572) | 52.6 (5658) | 33.7 (838) | 38.6 (2459) | 57.1 (734) | 63.3 (3199) |
| English-isiXhosa  | 51.9 (2300) | 62.8 (2651) | 39.6 (1153) | 44.9 (1149) | 64.3 (1147) | 76.5 (1502) |
| English-Setswana  | 56.7 (3707) | 52.3 (4939) | 38.0 (1170) | 33.9 (1970) | 65.4 (2537) | 64.6 (2969) |
| English-Sesotho   | 62.6 (3067) | 59.1 (4054) | 41.5 (843) | 41.9 (1794) | 70.6 (2224) | 72.8 (2260) |

Table 5: Confusion matrices of the hypothesized language tags on the development and test data

(a) Dev

|     | ENG | ZUL | XHO | TSN | SOT | DEL |
|-----|-----|-----|-----|-----|-----|-----|
| ENG | 3336 | 152 | 25  | 87  | 72  | 327 |
| ZUL | 84   | 513 | 10  | 11  | 17  | 99  |
| XHO | 166  | 313 | 531 | 25  | 23  | 89  |
| TSN | 373  | 153 | 20  | 832 | 626 | 535 |
| SOT | 361  | 209 | 11  | 528 | 732 | 386 |
| INS | 159  | 57  | 8   | 52  | 63  | 0   |

(b) Test

|     | ENG | ZUL | XHO | TSN | SOT | DEL |
|-----|-----|-----|-----|-----|-----|-----|
| ENG | 5953 | 309 | 62  | 174 | 165 | 694 |
| ZUL | 474  | 2010| 64  | 88  | 120 | 443 |
| XHO | 262  | 496 | 447 | 35  | 41  | 221 |
| TSN | 411  | 138 | 20  | 875 | 810 | 725 |
| SOT | 398  | 180 | 20  | 680 | 457 | 530 |
| INS | 211  | 77  | 16  | 62  | 56  | 0   |

tained. Confusions between the isiZulu-isiXhosa and Setswana-Sesotho language pairs are marked in purple and green respectively. This is done to highlight the cells where higher confusion is expected due to the similarity between the two languages.

Focusing firstly on isiZulu-isiXhosa, we see that the confusion occurs mostly in a single direction, i.e. many more isiXhosa words are identified as isiZulu words. In the second language pair, Setswana-Sesotho, the confusions occur in both directions in both development and test sets. The language recognition performance of the ASR system is significantly worse than any other language couple. This can be explained by the greater linguistic similarity between the languages and their larger intersection in the phoneme set and vocabulary. The lower acoustic and written resources further reduces the LR performance of the ASR system on the Sotho languages.

7. Conclusion

We present a first 5-lingual CS ASR system that is designed to recognize CS speech in 5 South African languages. Using a recently compiled soap opera speech corpus, we explore how well modern NN-based acoustic models can deal with the language switches given the limited availability of resources for the target languages. Language recognition performance implicit in the ASR is also evaluated, especially between the linguistically similar languages. We believe that these first findings are encouraging and provide insight into the challenges in building a unified CS system for multilingual countries such as South Africa.

8. Acknowledgements

This research is funded by the NWO Project 314-99-119 (Frisian Audio Mining Enterprise). The authors would like to thank the Department of Arts & Culture of the South African government for funding this research as well as Stellenbosch University for the travel grant that enabled the first author’s visit to Stellenbosch.

9. References

[1] G. Stemmer, E. Nöth, and H. Niemann, “Acoustic modeling of foreign words in a German speech recognition system,” in Proc. EUROSPEECH, 2001, pp. 2745–2748.
[2] D.-C. Lyu, R.-Y. Lyu, Y.-C. Chang, and C.-N. Hsu, “Speech recognition on code-switching among the Chinese dialects,” in Proc. ICASSP, vol. 1, May 2006, pp. 1105–1108.
[3] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telara, T. Schlippe, F. Blachier, E.-S. Chung, T. Schultz, and H. 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] C. H. Wu, H. P. Shen, and C. S. Hsu, “Code-switching event detection by using a latent language space model and the delta-Bayesian information criterion,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 11, pp. 1892–1903, Nov 2015.

[7] E. 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.

[8] E. Yılmaz, H. Van den Heuvel, and D. A. Van Leeuwen, “Investigating bilingual deep neural networks for automatic speech recognition of code-switching Frisian speech,” in Proc. SLTU, May 2016, pp. 159–166.

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

[10] Y. Li and P. Fung, “Code switching language model with translation constraint for mixed language speech recognition,” in Proc. COLING, Dec. 2012, pp. 1671–1680.

[11] H. Adel, N. Vu, F. Kraus, T. Schlippe, H. Li, and T. Schultz, “Recurrent neural network language modeling for code switching conversational speech,” in Proc. ICASSP, 2013, pp. 8411–8415.

[12] H. Adel, K. Kirchhoff, D. Telaar, N. T. Vu, T. Schlippe, and T. Schultz, “Features for factored language models for code-switching speech,” in Proc. SLTU, May 2014, pp. 32–38.

[13] Z. Zeng, H. Xuy, T. Y. Chongy, E.-S. Chingy, and H. Li, “Improving N-gram language modeling for code-switching speech recognition,” in Proc. APSIPA ASC, 2017, pp. 1–6.

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

[15] E. van der Westhuizen and T. Niesler, “Synthesisising isiZulu-English code-switch bigrams using word embeddings,” in Proc. INTERSPEECH, 2017, pp. 72–76.

[16] E. Yılmaz, H. van den Heuvel, and D. A. van Leeuwen, “Exploiting untranscribed broadcast data for improved code-switching detection,” in Proc. INTERSPEECH, Aug. 2017, pp. 42–46.

[17] 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, Dec. 2017, pp. 91–96.

[18] ——, “Semi-supervised acoustic model training for speech with code-switching,” Submitted to Speech Communication, 2018. [Online]. Available: goo.gl/NKiYAF

[19] 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 Proc. INTERSPEECH, 2018, Accepted for publication.

[20] A. T. Cope, “Zulu phonology, tonology, and tonal grammar,” Ph.D. dissertation, Durban; University of Natal, 1966.

[21] D. P. Kunene, “The sound system of Southern Sotho,” Ph.D. dissertation, Cape Town: University of Cape Town, 1961.

[22] E. van der Westhuizen and T. Niesler, “A first South African corpus of multilingual code-switched soap opera speech,” in Proc. LREC, 2018, pp. 2854–2859.

[23] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, “ELAN: a professional framework for multimodality research,” in Proc. LREC, vol. 2006, 2006, p. 5th.

[24] 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, Mar 1989.

[25] 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.

[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, March 2018.

[27] C. Biemann, G. Heyer, U. Quasthoff, and M. Richter, “The Leipzig Corpora Collection-monolingual corpora of standard size,” Proceedings of Corpus Linguistic, vol. 2007, 2007.

[28] R. Kneser and H. Ney, “Improved backing-off for M-gram language modeling,” in Proc. ICASSP, vol. I, 1995, pp. 181–184.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech recognition toolkit,” in Proc. ASRU, Dec. 2011.

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

[31] 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.

[32] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio augmentation for speech recognition,” in Proc. INTERSPEECH, 2015, pp. 3586–3589.