Automatic Speech Recognition for Humanitarian Applications in Somali

Raghav Menon[1], Astik Biswas[1], Armin Saeb[1], John Quinn[2] and Thomas Niesler[1]

[1]Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
[2]UN Global Pulse, Kampala, Uganda

rmenon@sun.ac.za, abiswas@sun.ac.za, arsaeb@sun.ac.za, john.quinn@unglobalpulse.org,
trn@sun.ac.za

Abstract

We present our first efforts in building an automatic speech recognition system for Somali, an under-resourced language, using 1.57 hrs of annotated speech for acoustic model training. The system is part of an ongoing effort by the United Nations (UN) to implement keyword spotting systems supporting humanitarian relief programmes in parts of Africa where languages are severely under-resourced. We evaluate several types of acoustic model, including recent neural architectures. Language model data augmentation using a combination of recurrent neural networks (RNN) and long short-term memory neural networks (LSTMs) as well as the perturbation of acoustic data are also considered. We find that both types of data augmentation are beneficial to performance, with our best system using a combination of convolutional neural networks (CNNs), time-delay neural networks (TDNNs) and bi-directional long short term memory (BLSTMs) to achieve a word error rate of 53.75%.

Index Terms: automatic speech recognition, BLSTM, under-resourced, acoustic modelling, Somali

1. Introduction

Recent research has established that, in societies with accessible internet infrastructure, social media is popular for voicing views and concerns about social issues [1–3]. However, in a series of surveys by the United Nations (UN), it was found that radio phone-in talk shows become the medium of choice for such communication when internet connectivity is insufficient or inaccessible [4–6]. For this reason, the development of radio browsing systems has been piloted with the UN. Successful systems are in current active use in Uganda and monitor such radio shows in three native languages to obtain information that can inform and support relief and developmental programmes [7, 8].

The success of the Ugandan systems have led to calls for a similar system in Somalia. The acoustic models of the Ugandan system were developed using orthographically annotated speech corpora of between 6 and 9 hours in length. However, the urgency with which the Somali system was required, as well as the difficulty with which the required linguistic expertise could be found, allowed only for the compilation of a 1.57-hour annotated speech corpus. This paper describes our efforts to produce the best possible Somali automatic speech recognition (ASR) system using this limited resource, with the intent of incorporating it in a radio browsing system for the monitoring of UN humanitarian operations in Somalia. Due to the very small corpus, we attempt to capitalise as much as possible on available data from other languages, such as those compiled for the Ugandan systems, by harnessing multilingual acoustic model training strategies. To achieve this, we consider recent architec-
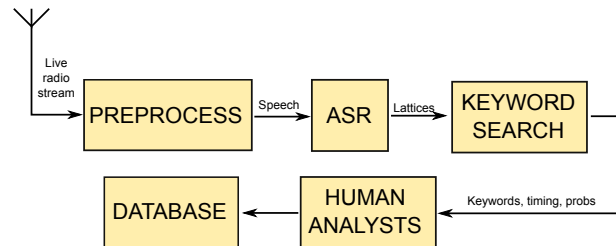

Figure 1: *The radio browsing system [8].*

tures, including feedforward deep neural networks (DNN) and time-delay neural networks (TDNN).

We have been able to locate just one other account of ASR for Somali in the literature [9]. These authors trained a Somali HMM-GMM acoustic model by adapting a French model using a phoneme mapping. A training corpus consisting of 9.5 hours of prepared Somali speech was used. Text-normalisation was achieved using tools prepared in-house to account for text variations. Hence, in contrast to our case, they had a well-built Somali corpus.

There has been a growing interest recently in developing ASR for under-resourced languages [10]. In this paper, we follow the same direction as in [7, 8] starting out by modelling ASR with a HMM-GMM and moving on to more advanced neural net based models. While doing this, we investigate ways in which performance in terms of word error rate (WER) can be improved for an under-resourced target language.

The paper is organized as follows: Section 2 gives a brief description of the radio browsing system, Section 3 provides a detailed account of the data available and language modelling, Section 4 describes the various architectures considered for acoustic model and the features used, Section 5 describes the experimental setup, Section 6 discusses the results and Section 7 concludes the paper.

2. Radio Browsing System

The structure of the existing radio browsing system is shown in Figure 1 [8]. In essence, it consists of an ASR and a keyword spotter. The live audio stream is first pre-processed and then passed on to the ASR system where lattices are generated. These lattices are subsequently searched for the occurrence of the keyword under consideration. This output is passed to human analysts who further processes and format the data for humanitarian decision making and situational awareness. This system is currently deployed and in constant use by the UN in Uganda.

3. Data

The acoustic training and test data used in our experiments, as described in Table 1, consisted of the following.

- **Somali data**, amounting to 1.57 hrs of training data and 10 minutes of test data drawn from broadcast Somali radio phone-in programmes.

- **Ugandan data**, consisting of radio broadcasts recorded in Ugandan English, Luganda, Acholi and Somali and amounting to 6, 9.6 and 9.2 hrs of speech respectively.

- **South African broadcast news data**, amounting to 20 hrs of speech recorded from a popular broadcast news channel [11].

- **South African code-switched data**, consisting of 14.3 hrs of spontaneous speech obtained from broadcast soap opera episodes [12]. This multilingual data contains five languages and frequent code-switching between the pairs English-isiZulu, English-isiXhosa, English-Setswana and English-Sesotho.

Table 1: *Duration in hours (h) and minutes (m) of data available for acoustic modelling. The total English data is obtained by pooling South African English and Ugandan English.*

Set	No of Utterances	Duration
Somali (som)		
Train	1.3k	1.57h
Test	95	10m
Total	1.4k	1.74h
English (eng)		
Train	14.9k	26h
Luganda (lu)		
Train	8.8k	9.6h
Acholi (ac)		
Train	4.9k	9.2h
English-isiZulu (ez)		
Train	8.4k	4.81h
English-isiXhosa (ex)		
Train	7k	2.68h
English-Setswana (et)		
Train	5.3k	2.33h
English-Sesotho (es)		
Train	5.7k	2.36h
Total Train	**56.3k**	**58.55h**

All four datasets described above have been orthographically transcribed by native speakers, and pronunciation dictionaries were available.

The text corpus available for language modelling was limited and is summarised in Table 2. We compiled this corpus by harvesting 59.23k sentences from Somalian news websites, 54.85k sentences from public facebook posts (which are carefully cleaned and filtered), and 215.30k sentences from selected facebook comments (which were not cleaned throughly). Furthermore, we have attempted to augment the language modelling data by artificially generating 775.3k sentences (11.29M words) by using a RNN-LSTM trained on 1320 sentences in the Somali acoustic training set. The incorporation of such artificial data has successfully improved speech recognition performance for other authors [13].

The language model for Somali was generated using the SRILM toolkit [14]. When training only on the language model

Table 2: *Data available for language modelling.*

Corpus	No of Word Tokens	No of Word Types	No of Sentences
Somali (acoustic training transcripts)	15.1k	4.7k	1.3k
Somali (cleaned web harvested text)	1.92M	82.8k	59.2k
Facebook post (cleaned)	1.55M	92.9k	54.9k
Facebook comments (not cleaned)	3.5M	356.7k	215.3k
RNN-LSTM generated text	11.29M	4.7k	775.3k
Total	6.98M	433.9k	1.1M

training corpus, a perplexity of 532.71 was achieved on the transcriptions of the acoustic test set, using Kneser-Ney discounting. We found that this discounting method was not suitable when incorporating the artificially-generated data, however. In this case, Witten-Bell discounting was chosen, leading to a perplexity of 269.80. Language models from different sources were interpolated to obtain the final language model, as shown in Table 3.

Table 3: *Perplexity analysis on Somali test set.*

Corpus	Kneser-Ney discounting	Witten-Bell discounting
Somali (train)	811.03	880.64
Somali (cleaned web harvested text)	502.11	541.86
Facebook post (cleaned)	536.43	571.16
Facebook comments (uncleaned)	1062.94	1161.36
RNN-LSTM generated text	-	1155.51
Interpolated (row 1, 2, 3, 4)	**532.71**	555.90
Interpolated (row 1, 2, 3, 4, 5)	-	**269.80**

4. Acoustic Modelling

The Kaldi speech recognition toolkit has been used for all experiments [15]. This decoder is based on finite state transducers (FSTs) and incorporates the language model, the pronunciation dictionary (lexicon) and context dependency into a single decoding graph [16]. We have built acoustic models using hidden Markov model / Gaussian mixture model (HMM/GMM), subspace Gaussian mixture model (SGMM), deep neural network (DNN), multilingual DNN (MDNN), long short-term memory neural network (LSTM), bidirectional LSTM (BLSTM) and time delay neural network (TDNN) architectures. We describe each briefly in the following.

4.1. Hidden Markov Model

Hidden Markov models (HMMs) with Gaussian mixture models (GMMs) observation probability densities are trained in a step-by-step fashion [17]. Our HMM/GMM system includes the application of linear discriminant analysis (LDA) and maximum likelihood linear transforms (MLLT) as well as speaker adaptive training (SAT) using feature-level maximum likelihood linear regression (fMLLR) for improved acoustic modelling. MFCC features, including velocity and acceleration coefficients (Δ and $\Delta\Delta$), were used as features.

4.2. Subspace Gaussian Mixture Model

Subspace Gaussian mixture models (SGMMs) allow a large collection of GMMs to be represented compactly [18]. This is advantageous especially in later model adaptation steps, such as speaker adaptation [19]. SGMMs are derived from an existing HMM/GMM system by first clustering the Gaussians to obtain a universal background model (UBM). This UBM is refined and
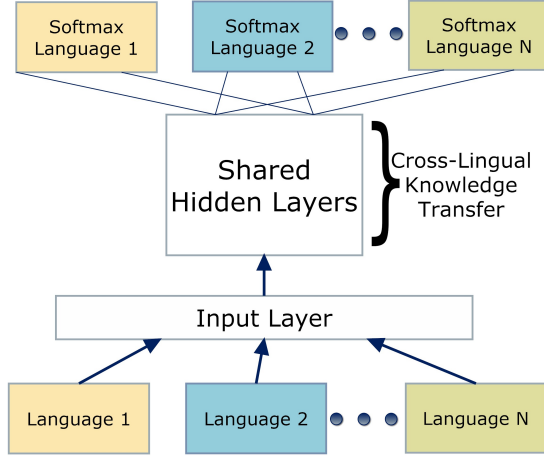
Figure 2: *Multilingual DNN/HMM acoustic model [26].*

Figure 3: *Neural network architectures considered for acoustic modelling.*

used to initialise the SGMM, which is then optimised by EM training. The decoding graph obtained for this acousic model is refined by discriminative training with Boosted Maximum Mutual Information (BMMI). As for the HMM/GMM system, fMLLR transforms were applied and MFCC features with appended Δ and $\Delta\Delta$ were used.

4.3. Deep Neural Networks (DNN)

Deep neural networks (DNNs) have over the last decade emerged as a new and successful acoustic modelling technique for automatic speech recognition [20]. We employ the DNN architecture described in [21]. Pre-training was not performed and the weight initialisation was random. Our DNN uses five 2048-dimensional hidden layers and an output dimension of 1037. The learning rate was allowed to decrease exponentially during training. As with the HMM/GMM and SGMM systems, fMLLR was applied. As features, 23-dimensional log mel-filterbank energies with appended Δ, $\Delta\Delta$ and 3-dimensional pitch were used.

4.4. Multilingual Deep Neural Networks (MDNN)

DNNs perform poorly as acoustic models when the training set is small, as is necessarily the case for under-resourced languages. This can be addressed by multilingual training [22–24]. Here the DNN is trained using a large training corpus containing multiple languages to provide the class conditional posterior probabilities required by HMMs. Figure 2 illustrates one possible multilingual architecture, in which the DNN shares hidden layers across all languages [8, 25]. Each language has its own softmax layer and its own HMM. As features, 23-dimensional log mel-filterbank energies with appended Δ, $\Delta\Delta$ and 3-dimensional pitch were used.

4.5. Long Short-Term Memory (LSTM) and Bi-directional LSTMs (BLSTM)

Acoustic modelling with long short-term memory (LSTM) has been successful and shown to be efficient [27–29]. A LSTM uses a memory block whose behavior is controlled by an input gate, a forget gate and an output gate. This configuration successfully exploits long-term temporal context dependencies. The memory blocks are recurrently connected to form a network. When the information is allowed to flow
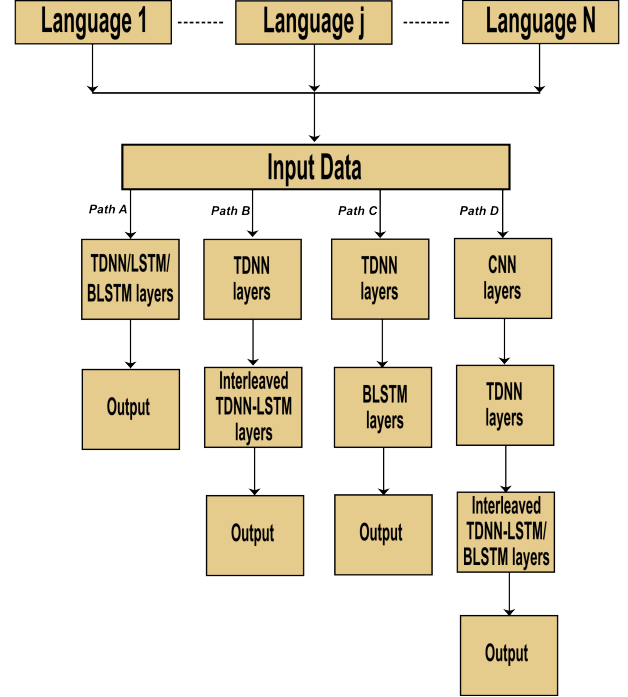
both forwards and backwards in time, a bi-directional LSTM (BLSTM) is obtained. We perform multilingual training by pooling the acoustic data described in Table 1 and presenting it to a LSTM/BLSTM network. The network outputs are phone targets. The trained network is subsequently adapted to the target language. Speaker adaptation is achieved using i-vectors. Features consists of 40-dimensional MFCCs with appended 3-dimensional pitch and 100-dimensional i-vector. This architecture is indicated by Path A in Figure 3.

4.6. Time-delay neural network (TDNN)-based models

Time-delay neural networks (TDNN) were introduced to overcome the limitation of DNNs in dealing with the dynamic nature of speech [30, 31]. A TDNN is able to represent temporal relationships between features by introducing delays on each input and processing each of these delays with different weights. The TDNN architecture is modular and incremental. Small networks are trained to perform limited tasks in order to develop time invariant, hidden abstractions. These are subsequently used to train larger networks [31]. The limitation of TDNN systems has been that they are computationally demanding to the extent that even GPUs could not sufficiently accelerate training. This has been overcome by the introduction of a subsampling procedure that assumes that neighbouring activations are correlated in time [32]. Such subsampling has been demonstrated to not only make the computational complexity of TDNNs practical, but also to improve ASR performance.

We have evaluated TDNN, TDNN/LSTM and TDNN/BLSTM network architectures for acoustic modelling, indicated respectively by Paths A, B and C in Figure 3. We have also considered the incorporation of a convolutional layer at the input of the TDNN/LSTM and TDNN/BLSTM networks, and indicate this by Path D in Figure 3. Multilingual training is achieved by pooling the acoustic data presented

in Table 1 during training and subsequently adapting the trained network to the target language. The features consists of 40-dimensional MFCCs with appended 3-dimensional pitch and 100-dimensional i-vectors for speaker adaptation.

5. Experimental Setup

We have modified the Kaldi 'wsj' recipe for our TDNN, LSTM and BLSTM experiments. For the CNN-TDNN-LSTM experiment we made use of the Kaldi 'ami' recipe. For the CNN-TDNN-BLSTM experiment, we have adapted the Kaldi 'ami' recipe by replacing the interleaved TDNN-LSTM layer with BLSTM layers. While in all cases the number of layers was unchanged, hyperparameters such as the learning rate range, context definitions, number of epochs and chunk widths were optimised. Monolingual acoustic models were built using the Somali data only. To obtain multilingual models, we incorporate the additional languages shown in Table 1. Since the performance of neural networks is known to improve with the amount of training data, speed perturbation (using factors 0.9, 1.0 and 1.1) and dithering were performed before training the TDNN-based models. System performance is reported in terms of word error rate (WER).

6. Results and Discussion

Table 4 indicates WERs achieved using various acoustic models. Decoding was performed using language models built with and without the artificially generated text. From the table we see that, for HMM/GMM and SGMM acoustic models, artificially generated text seems to degrade performance slightly. In contrast, the incorporation of the artificially generated text in the language model results in small improvements for the neural network acoustic models. We note that the parameters used for RNN-LSTM text generation were not optimised and it is possible that further improvements could be achieved if this were done.

Table 4: *Somali word error rates (WER) achieved using different acoustic modelling methods and different training data.*

Acoustic Model	WER (%)	
	No RNN-LSTM-generated text	*With RNN-LSTM-generated text*
HMM/GMM (som)	63.71	63.92
SGMM (som)	60.78	60.98
DNN (som)	59.82	59.48
MDNN (som+eng+lu+ac)	56.75	56.21
MDNN (som+eng+lu+ac+ez+ex)	-	55.93
MDNN (som+eng+lu+ac+az+ex+et+es)	-	57.23

We also see from Table 4 that the DNN/HMM architecture improved on the WER of the HMM/GMM by 4.44% absolute. Furthermore, a multilingual DNN (MDNN) incorporating the additional English, Luganda and Acholi data indicated in Table 1 achieved a further improvement of 3.27% over the DNN/HMM. This indicates that the neural network was able to take advantage of the larger training set afforded by the additional and unrelated languages shown in Table 1. Subsequent adaptation of the MDNN to the Somali data did not yield further gains.

When the code-switched data is added to the training set, the results are mixed. A small improvement in WER (0.28%) is achieved when adding the English-isiZulu and English-isiXhosa data. However, the further addition of English-

Setswana and English-Sesotho is seen to degrade the WER by 1.02%. This indicates that adding to the pool of languages is not always beneficial, and that care must be taken when deciding on the composition of the multilingual training set.

Table 5 shows the speech recognition performance achieved using LSTM, BLSTM and TDNN-based acoustic models. In all cases the training data consisted of the English, Luganda, Acholi and Somali listed in Table 1. All models were subjected to adaptation using the Somali data after training, and in contrast to the MDNN models, this achieved absolute improvements of between 1 and 5%. The LSTM model was outperformed by the TDNN model both with and without adaptation. The BLSTM fared even worse than the LSTM without adaptation, but better than both the LSTM and TDNN after adaptation. The TDNN-LSTM and TDNN-BLSTM both fared better than the LSTM and BLSTM models respectively, with and without adaptation. The addition of the CNN layer led to further consistent improvements for both TDNN-LSTM and TDNN-BLSTM architectures. Our best system used a CNN-TDNN-BLSTM acoustic model and achieved a word error rate of 53.75%.

Table 5: *Somali word error rates (WER) achieved using multilingual LSTM, BLSTM and various TDNN-based acoustic models trained on Somali, English, Luganda and Acholi.*

Acoustic Model	WER (%)	
	No Adaptation	*With Adaptation*
TDNN	60.57	58.05
LSTM	61.80	60.23
BLSTM	62.14	57.91
TDNN-LSTM	59.21	58.05
TDNN-BLSTM	58.12	56.41
CNN-TDNN-LSTM	56.21	55.53
CNN-TDNN-BLSTM	55.12	53.75

7. Conclusion

We have presented our first efforts in building an automatic speech recognition system for Somali as part of a United Nations programme aimed at humanitarian monitoring. A corpus of only 1.57h of in-domain Somali data was available. In addition, approximately 59h of annotated speech in 7 unrelated languages was available for multilingual modelling. Several acoustic model architectures were evaluated, including state-of-the-art neural configurations and a new combination of CNN, TDNN and BLSTM layers. Best performance was achieved by a multilingually-trained CNN-TDNN-BLSTM acoustic model, achieving a word error rate of 53.75%. Augmentation of the language model training data using a generative RNN-LSTM afforded small gains. We also found that BLSTM layers consistently benefited more from subsequent adaptation to the target language than LSTM layers.

8. Acknowledgements

9. References

[1] S. Vosoughi and D. Roy, "A human-machine collaborative system for identifying rumors on Twitter," in *Proc. ICDMW*, 2015.

[2] K. Wegrzyn-Wolska, L. Bougueroua, and G. Dziczkowski, "Social media analysis for e-health and medical purposes," in *Proc. CASoN*, 2011.

[3] P. Burnap, G. Colombo, and J. Scourfield, "Machine classification and analysis of suicide related communication on Twitter," in *Proc. ACM-HT*, 2015.

[4] G. P. P. Series, "Analyzing attitudes towards contraception and teenage pregnancy using social data," *Global Pulse Project Series*, no. 8, 2014.

[5] "Mining citizen feedback data for enhanced local government decision-making," *Global Pulse Project Series*, no. 16, 2015.

[6] "Understanding immunisation awareness and sentiment through social and mainstream media," *Global Pulse Project Series*, no. 19, 2015.

[7] R. Menon, A. Saeb, H. Cameron, W. Kibira, J. Quinn, and T. Niesler, "Radio-browsing for developmental monitoring in Uganda," in *Proc. ICASSP*, 2017.

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

[9] N. Addillahi, N.Pascal, and B. Jean-Francois, "Automatic transcription of Somali language," in *Proc. Interspeech*, 2006.

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

[11] H. Kamper, F. de Wet, T. Hain, and T. Niesler, "Capitalising on North American speech resources for the development of a South African English large vocabulary speech recognition sys-tem," *Computer Speech and Language*, vol. 28, no. 6, pp. 1255–1268, 2014.

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

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

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

[15] 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*, 2011.

[16] M. Mohri, "Finite-state transducers in language and speech processing," *Association of Computational Linguistics*, vol. 23, no. 2, pp. 269–311, 1997.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, South Brisbane, 2015, pp. 5206–5210.

[18] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafit, A. Rastrow, and R. C. Rose, "The subspace Gaussian mixture model-a structured model for speech recognition," *Computer Speech and Language*, vol. 25, no. 2, pp. 204–439, 2011.

[19] P. Motlicek, P. N. Garner, N. Kim, and J. Cho, "Accent adaptation using subspace Gaussian mixture models," in *Proc. ICASSP*, 2013.

[20] D. Yu and L. Deng, *Automatic Speech Recognition : A Deep Learning Approach*. London: Springer-Verlag, 2015.

[21] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech*, 2013.

[22] F. Grezl, M. Karafiat, and K. Vesel, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *Proc. ICASSP*, 2014.

[23] A. Mohan and R. Rose, "Multi-lingual speech recognition with low-rank multi-task deep neural networks," in *Proc. ICASSP*, 2015.

[24] R. Sahraeian and D. Compernolle, "A study of rank-constrained multilingual DNNs for low-resource ASR," in *Proc. ICASSP*, 2016.

[25] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*, 2013.

[26] 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).

[27] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *CoRR*, vol. abs/1402.1128, 2014. [Online]. Available: http://arxiv.org/abs/1402.1128

[28] J. T. Geiger, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling," in *Proc. Interspeech*, 2014.

[29] X. Li and X. Wu, "Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition," *CoRR*, vol. abs/1410.4281, 2014. [Online]. Available: http://arxiv.org/abs/1410.4281

[30] 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.

[31] A. Waibel, "Modular construction of time-delay neural networks for speech recognition," *Neural Computation*, vol. 1, no. 1, pp. 39–46, 1989.

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