Proceedings of the LREC 2020 – 4th Workshop on Computational Approaches to Code Switching, pages 52–56
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Semi-supervised Acoustic and Language Model Training for English-isiZulu Code-Switched Speech Recognition

Astik Biswas, Febe de Wet, Ewald van der Westhuizen, Thomas Niesler
Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
{abiswas, fdw, ewaldvdw, trn}@sun.ac.za

Abstract

We present an analysis of semi-supervised acoustic and language model training for English-isiZulu code-switched ASR using soap opera speech. Approximately 11 hours of untranscribed multilingual speech was transcribed automatically using four bilingual code-switching transcription systems operating in English-isiZulu, English-isiXhosa, English-Setswana and English-Sesotho. These transcriptions were incorporated into the acoustic and language model training sets. Results showed that the TDNN-F acoustic models benefit from the additional semi-supervised data and that even better performance could be achieved by including additional CNN layers. Using these CNN-TDNN-F acoustic models, a first iteration of semi-supervised training achieved an absolute mixed-language WER reduction of 3.4%, and a further 2.2% after a second iteration. Although the languages in the untranscribed data were unknown, the best results were obtained when all automatically transcribed data was used for training and not just the utterances classified as English-isiZulu. Despite reducing perplexity, the semi-supervised language model was not able to improve the ASR performance.

Keywords: code-switched speech, under-resourced languages, semi-supervised training, TDNN, CNN

1. Introduction

South Africa is a multilingual country with 11 official languages, including highly-resourced English which usually serves as a lingua-franca. The largely multilingual population commonly mix these geographically co-located languages in casual conversation. An ASR system deployed in this environment should therefore be able to process speech that includes two or more languages in one utterance.

The study and development of code-switching speech recognition systems has recently attracted increased research attention (Li and Fung, 2013; Yılmaz et al., 2018b; Adel et al., 2015; Emond et al., 2018). Language pairs that are of current research interest include English-Mandarin (Li and Fung, 2013; Vu et al., 2012; Zeng et al., 2018), Frisian-Dutch (Yılmaz et al., 2018b; Yılmaz et al., 2018a) and Hindi-English (Pandey et al., 2018). In South Africa, code-switching most often occurs between highly resourced English and one of the nine under-resourced, officially-recognised African languages.

In previous work, we showed that multilingual acoustic model training is effective for English-isiZulu code-switched ASR if additional training data from closely related languages is used (Biswas et al., 2018a). However, the 12.2 hours of training data provided by combining all our code-switching data is still too little to develop robust ASR systems.

A related study indicated that increasing the pool of in-domain training data using semi-supervised training achieved a significant improvement over the baseline acoustic model (Biswas et al., 2019). These findings motivated us to further optimise semi-supervised acoustic and language modelling training. Specifically, the effect of multiple iterations of semi-supervised training along with the application of a confidence threshold to filter the semi-supervised data was considered. We focus our investigation on one language pair, English-isiZulu, to allow for a detailed analysis of various aspects of the semi-supervised training despite the limited computational resources at our disposal.

2. Multilingual Soap Opera Corpus

The multilingual speech corpus was compiled from 626 South African soap opera episodes. Speech from these soap operas is typically spontaneous and fast, rich in code-switching and often expresses emotion, making it a challenging corpus for ASR development. The data contains examples of code-switching between South African English and four Bantu languages: isiZulu, isiXhosa, Setswana and Sesotho.

2.1. Manually Transcribed Data

Four language-balanced sets, transcribed by mother tongue speakers, were derived from the soap opera speech (van der Westhuizen and Niesler, 2018). In addition, a large but language-unbalanced (English dominated) dataset containing 21.1 hours of code-switched speech data was created (Biswas et al., 2019). The composition of this larger but unbalanced corpus is summarised in Table 2.1.. Note that all utterances in the development and test sets contain code-switching and that the balanced data is a subset of the unbalanced data.

	Language	Mono (m)	CS (m)	Subtotal (m)	Word tokens	Word types
Train	English	755.0	121.8	876.6	194 426	7 908
	isiZulu	92.8	57.4	150.0	24 412	6 789
	isiXhosa	65.1	23.8	88.8	13 825	5 630
	Sesotho	44.7	34.0	78.6	22 226	2 321
	Setswana	36.9	34.5	71.4	21 409	1 525
Dev	EZ	–	8.0	8.0	1 572	858
Test	EZ	–	30.4	30.4	5 658	3711
Total		994.4	271.5	1304.4	283 520	24 933

Table 1: Duration, in minutes (m), word type and word token counts for the unbalanced soap opera corpus. Both monolingual and code-switched (CS) durations are given.

2.2. Manually Segmented Untranscribed Data

In addition to the transcribed data introduced in the previous section, 23 290 segmented but untranscribed soap opera

utterances were generated during the creation of the multilingual corpus. These utterances correspond to 11.1 hours of speech from 127 speakers (69 male; 57 female). The languages in the untranscribed utterances are not labelled. Several South African languages not among the five present in the transcribed data are known to occur in these segments.

3. Semi-Supervised Training

Semi-supervised techniques were used to transcribe the data introduced in Section 2.2. (Yılmaz et al., 2018b; Nallasamy et al., 2012; Thomas et al., 2013), starting with our best existing code-switching speech recognition system. In this study the manually-segmented data was transcribed twice, as illustrated in Figure 1. After each transcription pass, the acoustic models were retrained and recognition performance was evaluated in terms of WER.

We distinguish between the acoustic models used to transcribe data (`AutoT`) and those that were used to evaluate WER (`ASR`) on the test set introduced in Table 2.1.. These two models differ in the composition of their training sets.

The acoustic models indicated by $\texttt{AutoT}_1$ in Figure 1 were trained on all the manually transcribed (ManT) data described in Section 2.1. as well as monolingual data from the NCHLT Speech Corpus (Barnard et al., 2014). These were the best available models to start semi-supervised training. The ManT and NCHLT data were subsequently pooled with the transcriptions produced by the $\texttt{AutoT}_1$ models to train an updated set of acoustic models ($\texttt{AutoT}_2$ in Figure 1) which were used to obtain a new set of transcriptions of the untranscribed data for semi-supervised training. In contrast, the acoustic models $\texttt{ASR}_1$ and $\texttt{ASR}_2$ were trained by pooling only the ManT and AutoT soap opera data; no out-of-domain NCHLT data was used.

Separate `AutoT` and `ASR` acoustic models are maintained because we use only in-domain data for semi-supervised training. This is computationally much easier, since the out-of-domain NCHLT datasets are approximately five times larger than the in-domain sets. However, it was found that better performance can be achieved in the second pass of semi-supervised training if the acoustic models maintain a similar training set composition to that used in the first pass. Hence, $\texttt{AutoT}_1$ and $\texttt{AutoT}_2$ were purpose-built, intermediate systems used solely to generate semi-supervised data.

Figure 1 also shows that each untranscribed utterance was decoded by four bilingual ASR systems. The highest confidence score was used to assign a language pair label to an utterance. In initial experiments, we added only EZ data identified in this way to the pool of multilingual training data. However, it was found that better performance could be achieved when all the AutoT data was added, and this was therefore done in the experiments reported here.

Two ways of augmenting the acoustic model training set with automatically-transcribed data were considered. First, all automatic transcriptions were pooled with the manually-labelled data. Second, utterances with a recognition confidence score below a threshold were excluded. The average confidence score across each language pair was used as a threshold. A larger variety of thresholds was not considered for computational reasons, but this remains part of ongoing work. Confidence thresholds were applied in three ways.

1. No threshold applied in either iteration 1 or 2 of semi-supervised training. The ManT data (21.1 h) was pooled with the AutoT_1 data to train $\texttt{ASR}_1$ and with the AutoT_2 data to train $\texttt{ASR}_2$. The duration of both AutoT_1 and AutoT_2 was 11.1 h.

2. Threshold applied only in iteration 1. In this case only a subset of the AutoT_1 data (4.2 h) was pooled with the ManT data to train $\texttt{ASR}_1$. All 11.1 h of AutoT_2 data was used to train $\texttt{ASR}_2$.

3. Threshold applied in both iteration 1 and iteration 2. This resulted in a 4.2 h subset of AutoT_1 used to train $\texttt{ASR}_1$ and a 4.3 h subset of AutoT_2 used to train $\texttt{ASR}_2$.

These three scenarios are indicated by NT, T_{P1} and T_{P1P2} respectively in Table 3., which shows the number of utterances assigned to each language pair. The total number of utterances and corresponding duration of the data included in the training set is shown in the last column.

Pass		EZ	EX	ES	ET	TOTAL
1	NT	7 951	3 796	11 415	128	23 290 (11.1 h)
2		9 347	2 145	5 415	6 381	23 290 (11.1 h)
1	T_{P1}	3 704	1 731	5 338	58	10 831 (4.2 h)
2		7 888	1 756	8 798	4 869	23 290 (11.1 h)
1	T_{P1P2}	3 704	1 731	5 338	58	10 831 (4.2 h)
2		3 686	834	4 115	2 320	10 955 (4.3 h)

Table 2: Number of utterances assigned to each language pair for automatically transcribed (AutoT) data.

4. Experiments

4.1. Language Modelling

The English-isiZulu vocabulary consisted of 11 292 unique word types and was closed with respect to the training, development and test sets. The SRILM toolkit (Stolcke, 2002) was used to train a bilingual trigram language model (LM) using the transcriptions described in Section 2.1. This LM was interpolated with two monolingual trigrams trained on 471 million English and 3.2 million isiZulu words of newspaper text, respectively. The interpolation weights were chosen to minimise the development set perplexity. The resulting language model was further interpolated with LMs derived from the transcriptions produced by the process illustrated in Figure 1 to obtain a semi-supervised LM.

4.2. Acoustic Modelling

All ASR experiments were performed using the Kaldi toolkit (Povey and others, 2011) and the data described in Section 2. The automatic transcription systems were implemented using factorized time-delay neural networks (TDNN-F) (Povey et al., 2018). For multilingual training, the training sets of all four language pairs were combined. However, the acoustic models were language dependent and no phone merging across languages took place.

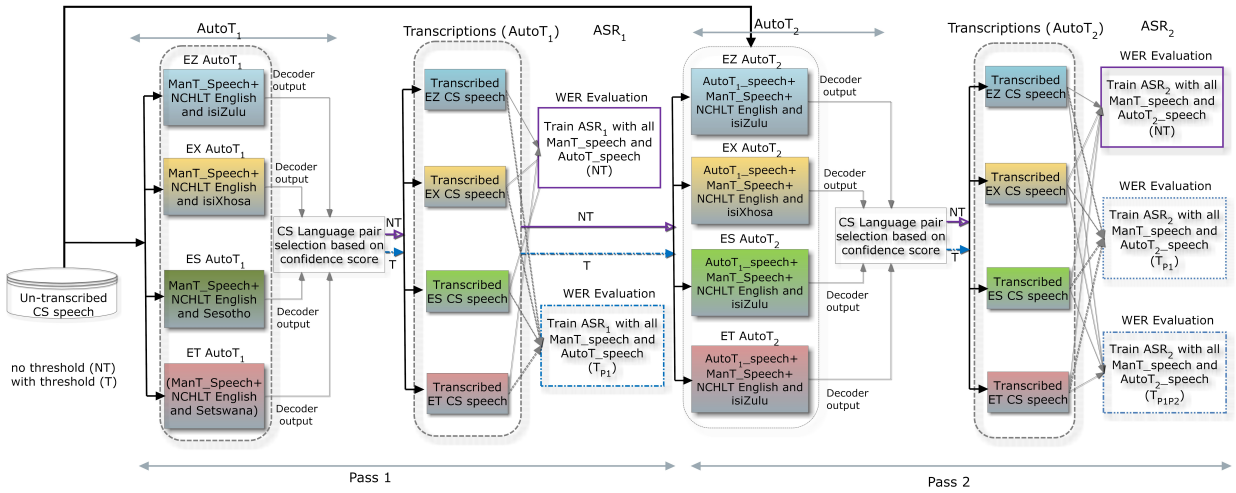

Figure 1: Semi-supervised training framework for English-isiZulu code-switched (CS) ASR.

A context-dependent GMM-HMM was trained to provide the alignments for neural network training. Three-fold data augmentation was applied prior to feature extraction (Ko et al., 2015) and the acoustic features comprised 40-dimensional MFCCs (without derivatives), 3-dimensional pitch features and 100-dimensional i-vectors for speaker adaptation.

We used two types of neural network-based acoustic model architectures: (1) TDNN-F with 10 time-delay layers followed by a rank reduction layer trained using the Kaldi Librispeech recipe (version 5.2.164) and (2) CNN-TDNN-F consisting of two CNN layers followed by the TDNN-F architecture. TDNN-F models have been shown to be effective in under-resourced scenarios (Povey et al., 2018). The locality, weight sharing and pooling properties of the CNNs have been shown to benefit ASR (Abdel-Hamid et al., 2014). The default recipe parameters were used during neural network training. In a final training step the multilingual acoustic models were adapted with English-isiZulu code-switched speech.

5. Results and Discussion

5.1. Language Modelling

Table 5.1. shows the test set perplexities (PP) for the LM configurations described in Section 4.1. The baseline language model, LM_0, was trained on the English-isiZulu acoustic training data transcriptions as well as monolingual English and isiZulu text (Biswas et al., 2018b). LM_0 was also interpolated with trigram LMs trained on the 1-best and 10-best outputs of $AutoT_2$ respectively. MPP indicates monolingual perplexity and is calculated over monolingual stretches of text only, omitting points at which the language alternates. CPP indicates code-switch perplexity and is calculated only over language switch points. Therefore CPP indicates the uncertainty of the first word following a language switch.

Table 5.1. shows that, relative to the baseline, adding automatically generated English-isiZulu transcriptions to the language model training data improves the overall perplexity for both the development and test sets. The per-language results show that this improvement is due to a lower isiZulu perplexity, while English suffers a small deterioration. CPP is reduced when incorporating the 1-best automatic transcriptions but less so when incorporating the 10-best. This indicates that the code-switches present in the 1-best outputs are more representative of the unseen test set switches than those present in the 10-best output.

LM	PP (dev)	PP	MPP_E	MPP_Z	MPP	CPP
LM_0 (baseline)	425.8	601.7	121.2	777.8	358.1	3 292.0
LM_0 + 1-best	416.1	587.4	123.1	743.6	351.1	3 160.3
LM_0 + 10-best	408.2	583.6	124.4	722.8	346.9	3 205.2

Table 3: Perplexity of bilingual English-isiZulu trigram LMs.

5.2. Acoustic Modelling

ASR performance was evaluated on the English-isiZulu test set for various configurations of the $\texttt{ASR}_1$ and $\texttt{ASR}_2$ systems.

5.2.1. ASR_1

Table 5.2.1. reports WER results for different configurations of $\texttt{ASR}_1$. Previously-reported results using a balanced subset of the corpus described in Section 2.1. are reproduced in rows 1 and 2. Language specific WERs are provided for the test set but not the development set.

The results in row 4 of the table show that, when the TDNN-F network is preceded by two CNN layers, test set recognition performance improves by 1.9% absolute. Row 5, on the other hand, shows that the inclusion of the automatically-transcribed English-isiZulu utterances reduces the test set WER of the TDNN-F models by 1.8% absolute. This improvement increases by an additional 0.8% absolute when including all the automatically transcribed data and not just the English-isiZulu utterances, as shown in row 6. Row 7 shows that the performance of the CNN-TDNN-F system is also enhanced by including the automatically transcribed data. In all the above cases, the WER improvements are seen not only overall but also in the English and isiZulu language-specific error rates.

Finally, the results in row 8 illustrate the impact of apply-

ing a confidence threshold to decide which automatically-transcribed utterances to include in the training set. The values in the table indicate that the mixed WER deteriorates marginally and that the English WER improves at the cost of a higher isiZulu WER.

	System configuration	Dev	Test	WER_E	WER_Z
1	ManT (balanced) TDNN-LSTM (Biswas et al., 2018a)	47.4	55.8	50.0	60.1
2	ManT (balanced) TDNN-BLSTM (Biswas et al., 2018b)	47.1	53.1	47.6	57.2
3	ManT (baseline) TDNN-F	41.3	47.4	41.8	51.8
4	ManT CNN-TDNN-F	40.8	45.6	40.0	49.9
5	ManT + AutoT$_1$ (EZ,NT) TDNN-F	41.2	45.7	39.6	50.3
6	ManT + AutoT$_1$ (All,NT) TDNN-F	39.5	44.9	38.9	49.6
7	ManT + AutoT$_1$ (All,NT) CNN-TDNN-F (Biswas et al., 2019)	**38.2**	**44.0**	37.9	48.7
8	ManT + AutoT$_1$ (All,T_{P1}) CNN-TDNN-F	38.8	44.2	36.6	50.1

Table 4: WER (%) on the English-isiZulu development (dev) and test sets for different configurations of $\texttt{ASR}_1$.

5.2.2. ASR_2

The results for the second iteration of semi-supervised training are reported in Table 5.2.2.. In all cases the ManT data was pooled with all the AutoT data and not just the EZ sub-set as was done in row 5 of Table 5.2.1.. Only the results using the CNN-TDNN-F acoustic models are shown, since this gave consistently superior performance in Table 5.2.1..

	Training data	LM	Dev	Test	WER_E	WER_Z
1	ManT + AutoT$_2$ (NT)	LM_0	38.6	42.5	36.2	47.6
2	ManT + AutoT$_2$ (T_{P1})	LM_0	38.0	43.1	37.5	47.4
3	ManT + AutoT$_2$ (T_{P1P2})	LM_0	40.1	43.9	34.2	51.3
4	ManT + AutoT$_2$ (NT, tuned)	LM_0	36.5	41.9	33.0	48.8
5		LM_0 + 1-best	**36.5**	**41.8**	33.9	47.9
6		LM_0 + 10-best	36.7	42.0	34.0	48.1

Table 5: WER (%) on the English-isiZulu development (dev) and test sets for different configurations of $\texttt{ASR}_2$.

A comparison between row 1 in Table 5.2.2. and row 7 in Table 5.2.1. reveals that a second pass of retraining affords a further 1.5% absolute reduction in test set WER. This was found to be statistically significant at more than 95% confidence level using bootstrap interval estimation (Bisani and Ney, 2004). Retraining $\texttt{ASR}_2$ with a threshold applied only to the output of AutoT$_1$ results in a slightly higher WER on the test set (row 2). Applying thresholds in both passes (row 3) improved the English WER but resulted in a substantial deterioration in isiZulu WER. This result suggests that, for the threshold value used here, English benefits from the exclusion of low-confidence automatically transcribed data while isiZulu does not. Thus, further study on the optimum threshold configuration is required.

The results in row 4 of Table 5.2.2. show that a further 0.6% absolute WER reduction can be achieved for the test set by tuning the learning rate during adaptation. Rows 5 and 6 show that retraining the LM on text that includes automatic transcriptions hardly influences recognition performance. Thus, although semi-supervised training led to appreciable improvements in the acoustic models, the corresponding positive effects on the language model were marginal.
A detailed analysis of different ASR outputs is shown in Table 5.2.2.. The analysis confirms that semi-supervised training resulted in substantial improvements in the English and isiZulu word correct accuracy. The results also reveal a substantial improvement in bigram correct accuracy at the 1 464 code-switch points occurring in the test set, where bigram correct accuracy (%) is defined as the percentage of words correctly recognised immediately after code-switch points.

Accuracy (%)	Table 4 (Row 3)	Table 4 (Row 4)	Table 4 (Row 7)	Table 4 (Row 8)	Table 5 (Row 4)
Eng token correct	59.8	61.5	64.5	65.4	68.8
Zul token correct	50.1	51.4	53.2	51.6	53.5
Word correct after switch	53.4	55.6	58.3	57.6	60.9
Zul word correct after switch	49.7	51.4	53.6	51.6	54.4
English word correct after switch	56.7	59.3	62.5	62.9	66.7
Language correct after switch	76.8	76.9	79.1	79.0	81.6
Code-switch bigram correct	29.0	30.8	33.3	32.2	35.6

Table 6: Detailed analysis of ASR accuracy for different acoustic models.

6. Conclusion

We have applied semi-supervised training to improve ASR for under-resourced code-switched English-isiZulu speech. Four different automatic transcription systems were used in two phases to decode 11 hours of multilingual, manually segmented but untranscribed soap opera speech. We found that by including CNN layers, CNN-TDNN-F acoustic models outperformed TDNN-F models on the code-switched speech. Furthermore, semi-supervised training provided a further absolute reduction of 5.5% in WER for the CNN-TDNN-F system. While the automatically transcribed English-isiZulu text data reduced language model perplexity, this improvement did not lead to a reduction in WER. By selective data inclusion using a confidence threshold, approximately 60% of the automatically transcribed data could be discarded at minimal loss in recognition performance. A more thorough investigation of this threshold remains part of ongoing work. We also aim to further extend the pool of training data by incorporating speaker and language diarisation systems to allow automatic segmentation of new audio.

7. Acknowledgements

We would like to thank the Department of Arts & Culture (DAC) of the South African government for funding this research. We are grateful to e.tv and Yula Quinn at Rhythm City, as well as the SABC and Human Stark at Generations: The Legacy, for assistance with data compilation. We also gratefully acknowledge the support of the NVIDIA corporation for the donation of GPU equipment.

8. Bibliographical References

Abdel-Hamid, O., Mohamed, A.-R., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1533–1545.

Adel, H., Vu, N. T., Kirchhoff, K., Telaar, D., and Schultz, T. (2015). Syntactic and semantic features for code-switching factored language models. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(3):431–440.

Barnard, E., Davel, M. H., Heerden, C. v., de Wet, F., and Badenhorst, J. (2014). The NCHLT speech corpus of the South African languages. In *Proc. SLTU*, St Petersburg, Russia.

Bisani, M. and Ney, H. (2004). Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proc. ICASSP*, Montreal, Canada.

Biswas, A., de Wet, F., van der Westhuizen, E., Yılmaz, E., and Niesler, T. R. (2018a). Multilingual neural network acoustic modelling for ASR of under-resourced English-isiZulu code-switched speech. In *Proc. Interspeech*, Hyderabad, India.

Biswas, A., van der Westhuizen, E., Niesler, T. R., and de Wet, F. (2018b). Improving ASR for code-switched speech in under-resourced languages using out-of-domain data. In *Proc. SLTU*, Gurugram, India.

Biswas, A., Yılmaz, E., de Wet, F., van der Westhuizen, E., and Niesler, T. R. (2019). Semi-supervised acoustic model training for five-lingual code-switched ASR. In *Proc. Interspeech*, Graz, Austria.

Emond, J., Ramabhadran, B., Roark, B., Moreno, P., and Ma, M. (2018). Transliteration based approaches to improve code-switched speech recognition performance. In *Proc. SLT*, Athens, Greece.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proc. Interspeech*, Dresdan, Germany.

Li, Y. and Fung, P. (2013). Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints. In *Proc. ICASSP*, Vancouver, Canada.

Nallasamy, U., Metze, F., and Schultz, T. (2012). Semi-supervised learning for speech recognition in the context of accent adaptation. In *Symposium on Machine Learning in Speech and Language Processing*, Portland, Oregon, USA.

Pandey, A., Srivastava, B. M. L., Kumar, R., Nellore, B. T., Teja, K. S., and Gangashetty, S. V. (2018). Phonetically balanced code-mixed speech corpus for Hindi-English automatic speech recognition. In *Proc. LREC*, Miyazaki, Japan.

Povey, D. et al. (2011). The Kaldi speech recognition toolkit. In *Proc. ASRU*, Hawaii, USA.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., and Khudanpur, S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proc. Interspeech*, Graz, Austria.

Stolcke, A. (2002). SRILM – An extensible language modeling toolkit. In *Proc. ICSLP*, Denver, USA.

Thomas, S., Seltzer, M. L., Church, K., and Hermansky, H. (2013). Deep neural network features and semi-supervised training for low resource speech recognition. In *Proc. ICASSP*, Vancouver, Canada.

van der Westhuizen, E. and Niesler, T. R. (2018). A first South African corpus of multilingual code-switched soap opera speech. In *Proc. LREC*, Miyazaki, Japan.

Vu, N. T., Lyu, D.-C., Weiner, J., Telaar, D., Schlippe, T., Blaicher, F., Chng, E.-S., Schultz, T., and Li, H. (2012). A first speech recognition system for Mandarin-English code-switch conversational speech. In *Proc. ICASSP*, Kyoto, Japan.

Yılmaz, E., Biswas, A., van der Westhuizen, E., de Wet, F., and Niesler, T. R. (2018a). Building a unified code-switching ASR system for South African languages. In *Proc. Interspeech*, Hyderabad, India.

Yılmaz, E., McLaren, M., van den Heuvel, H., and van Leeuwen, D. A. (2018b). Semi-supervised acoustic model training for speech with code-switching. *Speech Communication*, 105:12–22.

Zeng, Z., Khassanov, Y., Pham, V. T., Xu, H., Chng, E. S., and Li, H. (2018). On the end-to-end solution to Mandarin-English code-switching speech recognition. *arXiv preprint arXiv:1811.00241*.