Semi-supervised Development of ASR Systems for Multilingual Code-switched Speech in Under-resourced Languages

Astik Biswas[1], Emre Yılmaz[2], Febe de Wet[1], Ewald van der Westhuizen[1], Thomas Niesler[1]
[1] Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
[2] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
{abiswas, fdw, ewaldvdw, trn}@sun.ac.za, emre@nus.edu.sg

Abstract

This paper reports on the semi-supervised development of acoustic and language models for under-resourced, code-switched speech in five South African languages. Two approaches are considered. The first constructs four separate bilingual automatic speech recognisers (ASRs) corresponding to four different language pairs between which speakers switch frequently. The second uses a single, unified, five-lingual ASR system that represents all the languages (English, isiZulu, isiXhosa, Setswana and Sesotho). We evaluate the effectiveness of these two approaches when used to add additional data to our extremely sparse training sets. Results indicate that batch-wise semi-supervised training yields better results than a non-batch-wise approach. Furthermore, while the separate bilingual systems achieved better recognition performance than the unified system, they benefited more from pseudo-labels generated by the five-lingual system than from those generated by the bilingual systems.

Keywords: Code-switched speech, under-resourced languages, multilingual speech, semi-supervised training, ASR

1. Introduction

Much research has already been dedicated to the development of automatic speech recognition (ASR) for code-switching between various languages. English-Mandarin has probably been studied most extensively (Li and Fung, 2013a; Li and Fung, 2013b; Zeng et al., 2018; Vu et al., 2012; Taneja et al., 2019), but other language pairs such as Frisian-Dutch (Yılmaz et al., 2018c), Hindi-English (Pandey et al., 2018; Emond et al., 2018), English-Malay (Ahmed and Tan, 2012) and French-Arabic (Amazouz et al., 2017) have also received some attention. Developing models that are robust to the additional complexity associated with code-switching is challenging. The task becomes even more difficult when the languages in question are under-resourced since small text and acoustic data sets limit modelling capacity.

South Africa has a multilingual population of 57 million people and 11 official languages, including English. Due to the variety of geographically co-located languages, code-switching - the alternation between languages during communication - is a common phenomenon. Code-switching is most prevalent between English, a highly-resourced language, and the South African Bantu languages, which are all under-resourced. A corpus of code-switched speech originating from South African soap operas has recently been compiled to enable the development of ASR for this type of speech (van der Westhuizen and Niesler, 2018).

Previous work demonstrated that multilingual training using in-domain soap opera code-switched speech (Biswas et al., 2018a; Yılmaz et al., 2018a) and poorly matched monolingual South African speech (Biswas et al., 2018b) improves the performance of both bilingual and five-lingual ASR systems when the additional training data is from a closely-related language. Specifically, isiZulu, isiXhosa, Sesotho and Setswana belong to the same Bantu language family and were found to complement each other when combined into a multilingual training set. It has also been shown that, in comparison with in-domain code-switched data, out-of-domain monolingual speech yields relatively little performance improvement in acoustic modelling (Biswas et al., 2018b). However, the in-domain training data that is currently available remains insufficient for robust ASR development and hence obtaining more in-domain data remains key to improve the recognition accuracy of code-switched speech.

Compiling a multilingual corpus of code-switched speech is an extremely labour intensive process, both in terms of effort and time, because manual segmentation and annotation of the data are required. In the absence of manually annotated material, automatically transcribed training material has been shown to be useful in under-resourced scenarios using semi-supervised training (Thomas et al., 2013; Yılmaz et al., 2018c; Guo et al., 2018). For example, this strategy was used successfully to obtain bilingual and five-lingual ASR systems using 11.5 hours of manually segmented but untranscribed soap-opera speech (Biswas et al., 2019). Furthermore, the bilingual systems trained with automatic transcriptions generated by the five-lingual transcription system achieved the best performance.

Motivated by these results, we now investigate a batch-wise semi-supervised technique in which we incorporate additional batches of manually segmented but untranscribed soap opera data for acoustic and language modelling. Initial transcriptions were generated using our best systems trained on the manually transcribed speech. Given the multilingual nature of the data, the transcription systems must not only provide the orthography, but also the language(s) present at each location in the segment. Each utterance was therefore presented to the four individual code-switching systems as well as to the five-lingual system. In both cases two training configurations were considered, the first presenting all the data in one pass and the second presenting the data in smaller batches.

Finally, we also present language modelling experiments that investigate the inclusion of the automatically generated transcriptions and artificially generated text as training material for English-isiZulu.

Language	Mono (m)	CS (m)	Total (h)	Total (%)	Word tokens	Word types
English	755.0	121.8	14.6	69.3	194 426	7 908
isiZulu	92.8	57.4	2.5	11.9	24 412	6 789
isiXhosa	65.1	23.8	1.5	7.0	13 825	5 630
Sesotho	44.7	34.0	1.3	6.2	22 226	2 321
Setswana	36.9	34.5	1.2	5.6	21 409	1 525
Total	994.5	271.5	21.1	100.0	276 290	24 170

Table 1: Duration in minutes (min) and hours (h) as well as word type and token counts for the unbalanced training speech.

2. Multilingual soap opera corpus

This work uses a multilingual corpus including examples of code-switching between South African English and four Bantu languages. The corpus, which was compiled from South African soap opera episodes, contains 21 hours of annotated South African code-switched speech data divided into four language pairs: English-isiZulu (EZ), English-isiXhosa (EX), English-Setswana (ET), and English-Sesotho (ES). Of the Bantu languages, isiZulu and isiXhosa belong to the Nguni language family while Setswana and Sesotho are Sotho-Tswana languages. The speech in question is typically fast and often expresses emotion. These aspects of the data in combination with the high prevalence of code-switching makes it a challenging corpus for ASR experiments.

The corpus is however still under construction. During the first phase of development, more than 600 South African soap opera episodes were manually segmented and a substantial portion of this also manually transcribed. The second phase is currently underway and has thus far contributed manually segmented but still untranscribed data to the corpus.

2.1. Manually transcribed data (ManT)

The version of the soap opera corpus we used to develop our first code-switching ASR systems consisted of 14.3 hours of speech divided into four language-balanced sets, as described in (van der Westhuizen and Niesler, 2018). In addition to the language-balanced sets, approximately 9 hours of manually transcribed monolingual English soap opera speech was also available. This data was initially excluded to avoid a bias toward English. However, pilot experiments indicated that, counter to expectations, its inclusion enhanced recognition performance. These 9 hours of English data were therefore merged with the balanced sets for the experiments described here.

The composition of the unbalanced training speech is reported in Table 1. An overview of the statistics for the development (Dev) and test (Test) sets for each language pair is given in Table 2. The table includes values for the total duration as well as the duration of the monolingual and code-switched segments. The test sets contain no monolingual data. A total of approximately 4,000 language switches (English-to-Bantu and Bantu-to-English) are observed in the test set.

The number of unique English, isiZulu, isiXhosa, Sesotho and Setswana words in the corpus are 8 275, 11 352, 6 169, 2 792, 1 902 respectively. IsiZulu and isiXhosa have relatively large vocabularies due to their agglutinative nature. This property adds to the challenge of developing accurate ASR systems in these languages.

	English-isiZulu				
	emdur	zmdur	ecdur	zcdur	**Total**
Dev	0.00	0.00	4.01	3.96	8.00
Test	0.00	0.00	12.76	17.85	30.40
	English-isiXhosa				
	emdur	xmdur	ecdur	xcdur	**Total**
Dev	2.86	6.48	2.21	2.13	13.68
Test	0.00	0.00	5.56	8.78	14.34
	English-Setswana				
	emdur	tmdur	ecdur	tcdur	**Total**
Dev	0.76	4.26	4.54	4.27	13.83
Test	0.00	0.00	8.87	8.96	17.83
	English-Sesotho				
	emdur	smdur	ecdur	scdur	**Total**
Dev	1.09	5.05	3.03	3.59	12.77
Test	0.00	0.00	7.80	7.74	15.54

Table 2: Duration (minutes) of English, isiZulu, isiXhosa, Sesotho, Setswana monolingual (mdur) and code-switched (cdur) utterances in the code-switching development and test sets.

2.2. Manually segmented data: Batch 1

A set of approximately 11 hours of manually segmented speech representing 127 different speakers was produced in addition to the manually transcribed data introduced in the previous section. Segmentation was performed manually by experienced language practitioners. This data set (**B1**) was automatically transcribed during our initial investigations into semi-supervised acoustic model training (Biswas et al., 2019). Two sets of automatic transcriptions derived from B1 are considered: one obtained using four bilingual systems (AutoT_B_{B1}) and the other using a five-lingual system (AutoT_F_{B1}).

2.3. Manually segmented data: Batches 2 & 3

A subsequent phase of corpus development, currently still underway, has produced two new batches of manually segmented data. Manual transcriptions of this data are not yet available. These data sets will be referred to as Batch 2 (**B2**) and Batch 3 (**B3**), respectively. In contrast to B1, the segmentation was done by trained assistants because no specialist language practitioners were available. Hence, the quality of the segments in B2 and B3 may differ from those in B1.

Batch B2 includes approximately 24 hours of speech produced by 157 speakers, while B3 contains a further 30 hours of speech from 145 speakers. Most speakers occur in both batches and the languages spoken in the segments are not labelled. South African languages other than the five present in the transcribed data are known to occur in these batches, but to a limited extent.

3. Acoustic Modelling

All ASR experiments were performed using the Kaldi ASR toolkit (Povey et al., 2011) and the data described in Sec-

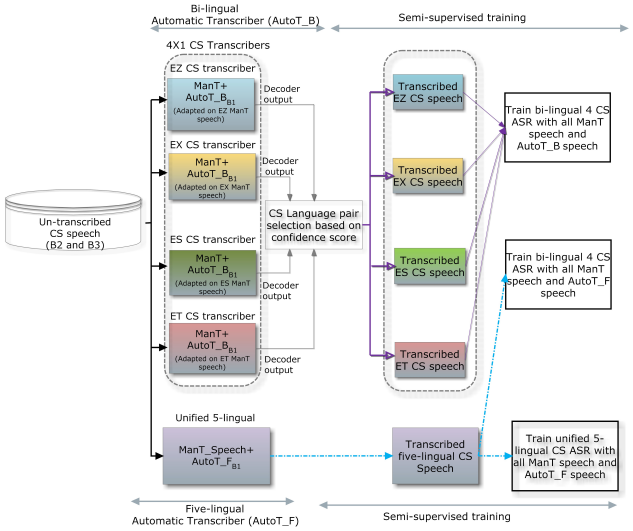

Figure 1: Semi-supervised training framework for bilingual and five-lingual systems.

tion 2. For multilingual training, the training sets of all the relevant languages were pooled. No phone merging was performed between languages and hence all acoustic models are language dependent.

Context-dependent Gaussian mixture model - hidden Markov models (GMM-HMM) were trained to obtain the alignments required for neural network training. Three-fold data augmentation (Ko et al., 2015) was applied prior to feature extraction for neural network training. The feature set used to train the neural network comprised MFCCs (40-dimensional, without derivatives), pitch features (3-dimensional) and i-vectors for speaker adaptation (100-dimensional) (Saon et al., 2013).

The acoustic models of all ASR systems were trained according the standard Kaldi CNN-TDNN-F (Povey et al., 2018) Librispeech recipe (6 CNN layers and 10 time-delay layers followed by a rank reduction layer) using the default hyperparameters. For the bilingual experiments, the multilingual acoustic models were subsequently adapted to the four different target language pairs.

4. Automatic transcription systems

A recent study demonstrated that semi-supervised training can improve the performance of Frisian-Dutch code-switched ASR (Yılmaz et al., 2018c). A similar approach was taken in this study, using the system configuration shown in Figure 1. The figure illustrates the two phases of semi-supervised training for the parallel bilingual as well as five-lingual configurations: automatic transcription followed by acoustic model retraining.

4.1. Parallel bilingual transcription

The first transcription system in Figure 1 consists of four subsystems, each corresponding to a code-switch language pair (4×CS in Figure 1). Acoustic models were trained on the manually transcribed soap opera data (ManT) described in Section 2.1 pooled with the automatically transcribed speech (AutoT_B) introduced in Section 2.2. Because the languages in the untranscribed data were unknown, each utterance was decoded in parallel by each of the bilingual decoders. The output with the highest confidence score provided both the transcription and a language pair label for each segment.

4.2. Five-lingual transcription

The second transcription system was based on a single acoustic model trained on all five languages. The training data consisted of the manually transcribed soap opera speech (ManT) pooled with the transcriptions generated by a five-lingual system (AutoT_F). Since the five-lingual system is not restricted to bilingual output, Bantu-to-Bantu language switching was possible and was observed in these transcriptions. Moreover, the automatically generated transcriptions sometimes contained more than two languages. Although the use of more than two languages within a single utterance is not common, our soap opera data does include a few such examples.

5. Language Modelling

5.1. Baseline language model

The EZ, EX, ES, ET vocabularies contained 11 292, 8 805, 4 233, 4 957 word types respectively and were closed with respect to the training, development and test sets. The SRILM toolkit (Stolcke, 2002) was used to train and evaluate all language models (LMs). Four bilingual and one five-lingual trigram language model were used for the transcription systems as well as for semi-supervised training (Yılmaz et al., 2018c; Biswas et al., 2019). Table 3 summarises the development and test set perplexities. Details on the monolingual and code-switch perplexities are only provided for the test set (columns 3 to 6 in Table 3).

Much more monolingual English text was available for language model development than text in the Bantu languages (471M vs 8M words). Therefore, the monolingual perplexity (MPP) is much higher for the Bantu languages than for English for each language pair.

Code-switch perplexities (CPP) for language switches indicate the uncertainty of the first word following a language switch. EB corresponds to switches from English to a Bantu language and BE indicates a switch in the other direction. Table 3 shows that the CPP for switching from English to isiZulu and isiXhosa is much higher than switching from these languages to English. This can be ascribed to the much larger isiZulu and isiXhosa vocabularies, which are, in turn, due to the high degree of agglutination and the use of conjunctive orthography in these languages. The CPP values are even higher for the five-lingual language model. This is because the five-lingual trigrams allow language switches not permitted by the bilingual models.

5.2. Semi-supervised language models

Only bilingual transcriptions were considered for the semi-supervised language model experiments. The automatically generated transcriptions of data sets B2 and B3 were added to the language model training data, similar to the approach proposed in (Drugman et al., 2019). Semi-supervised bilingual language models for each language pair were obtained by interpolating the baseline trigram with a trigram derived

	Dev	Test	all CPP	CPP_{EB}	CPP_{BE}	all MPP	MPP_E	MPP_Z
Bilingual trigram language model								
EZ	425.8	601.7	3 291.9	3 835.0	2 865.4	358.1	121.1	777.8
EX	352.9	788.8	4 914.4	6 549.6	3 785.6	459.0	96.8	1 355.6
ES	151.5	180.5	959.0	208.6	4 059.1	121.2	126.9	117.8
ET	213.3	224.5	70.2	317.3	3 798.1	160.4	142.1	176.1
Five-lingual trigram language model								
EZ	599.9	1 007.1	6 708.2	17 371.0	2 825.2	561.8	94.4	2 013.0
EX	669.1	1 881.9	15 083.6	50 208.3	5 058.0	1 015.9	87.6	5 590.0
ES	365.5	345.3	3 617.4	2 607.1	5 088.8	207.8	103.9	355.8
ET	237.0	277.5	2 936.6	1 528.4	5 446.3	158.1	99.8	211.2

Table 3: Development and test set perplexities. CPP: code-switch perplexity. MPP: monolingual perplexity.

from the text in the transcriptions of the corresponding target language pair. The interpolation weights were optimised using the development data.

A related study has shown that text data augmentation can be useful in under-resourced scenarios (Yılmaz et al., 2018b). This approach was evaluated on the EZ subset of our data by training a long short-term memory (LSTM) model on the manual and automatic transcriptions to generate additional artificial text data. The model was subsequently used to generate a data set of approximately 11.5 million words. The semi-supervised English-isiZulu language model described in the previous paragraph was interpolated with a trigram trained on this artificial data set.

In a further attempt to strengthen the language model at code-switch points, 1 million of artificial code-switched trigrams were synthesised using the method described in (van der Westhuizen and Niesler, 2019). The perplexity values of the resulting language models are reported in Table 4.

The first row in the table shows that adding the transcriptions of the automatically transcribed data to the LM training set reduces the test set perplexity of the EZ semi-supervised language model by more than 50 relative to the baseline value in Table 3. The ET semi-supervised language model also achieved a significant perplexity reduction on the test set. However, the EX and ES semi-supervised language models did not show any improvement compared to their respective baselines. This may be because there are far fewer isiXhosa and Sesotho segments in the automatically generated transcriptions than isiZulu and Setswana segments (cf. Table 6).

	Resources	Dev	Test	all CPP	CPP_{EB}	CPP_{BE}	all MPP	MPP_E	MPP_Z
EZ	ManT + AutoT_B_{B1} + AutoT_B_{B2B3}	392.2	547.4	2898.8	3297.6	2578.3	328.5	108.2	727.4
	+ 11.5M artificial text	362.8	507.5	2368.8	3005.8	1907.7	315.9	103.7	701.3
	+ 1M synthetic bigrams	358.3	501.9	2139.8	2613.5	1784.2	320.7	108.3	697.5
EX	ManT + AutoT_B_{B1} + AutoT_B_{B2B3}	345.4	787.4	5039.1	7176.3	3654.6	454.5	89.2	1411.4
ES	ManT + AutoT_B_{B1} + AutoT_B_{B2B3}	200.7	206.7	1074.3	347.0	3487.7	144.9	117.3	170.8
ET	ManT + AutoT_B_{B1} + AutoT_B_{B2B3}	138.4	164.8	938.0	214.6	3784.7	108.8	105.5	111.0

Table 4: Development and test set perplexities for different language model configurations. AutoT_B_{B1}: B1 transcribed as described in (Biswas et al., 2019). AutoT_B_{B2B3}: B2 transcribed by system **A** and B3 transcribed by system **C** (Figure 2).

Table 4 also shows that the additional text generated by the LSTM reduced the EZ development and test perplexity values substantially. The additional text also helped to bolster the language model at code-switch points. EZ CPP improved further after the 1M synthesized trigrams were added.

6. Experiments

Bilingual semi-supervised acoustic model training experiments were performed using batches B2 and B3 according to the two approaches illustrated in Figure 2. Similar configurations were used for the five-lingual experiments. In the first approach, both batches were first automatically transcribed using the baseline ASR (System **A** in Figure 2) followed by retraining using automatic transcriptions for both batches (System **B**).

The second approach used batch-wise semi-supervised training. Using System **A**, B2 was transcribed first, followed by acoustic model retraining with the automatically transcribed B2 data included in the training set (System **C**). B3 was then transcribed using the updated models and the retraining process repeated, this time also including the transcriptions of B3 (System **D**). This order was also reversed i.e. B3 first (System **E**) followed by B2 (System **F**). The experimental procedure for the bilingual and five-lingual systems is summarised in Table 5. The manual transcriptions introduced in Section 2.1 were always included in the training set. The composition of the automatic transcriptions included during training is shown in the last three columns of the table, with the last two indicating which systems were used to generate transcriptions for B2 and B3 respectively. Preliminary experiments indicated that the bilingual ASR systems achieved best performance when trained on AutoT_F transcriptions (Biswas et al., 2019). Thus, the bilingual systems considered here were also trained on the AutoT_F transcriptions of B2 and B3 (System N in Table 5).

Type of ASR	System	AutoT		
		B1	B2	B3
Bilingual	A (Baseline)	AutoT_B	-	-
	B		A	A
	C		A	-
	D		A	C
	E		-	A
	F		E	A
5-lingual	G (Baseline)	AutoT_F	-	-
	H		G	G
	I		G	-
	J		G	I
	K		-	G
	L		K	G
Bilingual (5-lingual trans.)	N	AutoT_F	G	I

Table 5: Semi-supervised acoustic model configurations.

7. Results and Discussion

7.1. Automatic transcription

The output of the transcription systems is summarised in Table 6. The first five rows in the table correspond to segments that were classified as monolingual while the

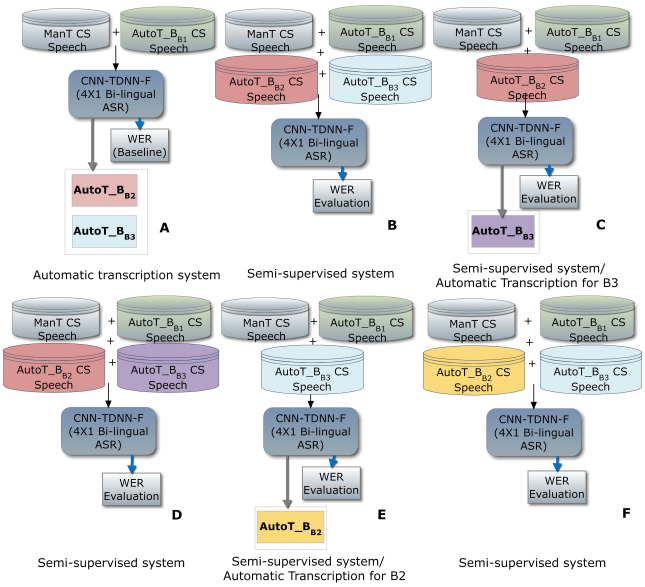

Figure 2: Semi-supervised training configurations for bilingual ASR. System names are given in parenthesis. (A: Baseline, B: Without batches, C to F: Batchwise)

last row shows the number of segments that contain code-switching. The values in this row reveal a high number of code-switched segments in data sets B2 and B3. It should be kept in mind that transcriptions of mixed language utterances produced by configurations A, C and E can only contain code-switching between English and one Bantu language. In contrast, the utterances produced by G, I and K can contain examples of code-switching between two or more Bantu languages.

Language	A		C	E	G		I	K
	B2	B3	B3	B2	B2	B3	B3	B2
English	8 570	11 746	12 027	8 794	12 373	16 333	17 059	12 898
isiZulu	5 955	6 190	6 854	6 604	4 708	4 587	4 641	4 496
isiXhosa	302	244	324	337	332	72	108	174
Sesotho	1 317	1 310	1043	1 889	1 067	678	581	797
Setswana	2 598	2 602	3 042	2 539	1 244	1 009	1 038	1 124
Code-switched	25 824	28 573	27 407	24 257	24 842	27 675	27 077	25 086

Table 6: Number of segments per language for different transcription systems.

7.2. Automatic speech recognition

ASR quality was measured in terms of word error rate (WER) evaluated on the development and test sets for each language pair described in Table 2. Results for the different semi-supervised training configurations are reported in Table 7.

7.2.1. Bilingual semi-supervised training

The upper part of Table 7 shows that, on average, semi-supervised training using 53 additional hours of speech data (B) yields an absolute improvement of 1.3% over the baseline (A) on the test set. It could be argued that this improvement is not large, given how much additional data was added to the training set. However, the segments were not created by language experts and may therefore not be accurate. Improving the quality of the segments might lead to

CS Pair	A (Baseline)		B		C		D		E		F	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Bilingual ASR												
EZ	34.5	40.8	33.4	39.3	32.7	39.6	32.8	38.6	34.6	39.9	32.8	39.0
EX	35.8	42.7	34.8	41.8	35.5	42.0	35.0	41.0	33.7	41.4	34.3	42.2
ES	51.6	48.7	49.3	47.3	49.0	46.5	48.7	45.8	49.5	47.4	49.5	46.4
ET	44.3	41.3	41.8	39.9	41.8	40.3	40.7	39.6	42.3	39.3	42.2	38.6
Overall	41.5	43.4	39.8	42.1	39.8	42.1	39.3	**41.3**	40.0	42.0	39.7	41.6
5-lingual ASR												
CS Pair	G (Baseline)		H		I		J		K		L	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
EZ	37.6	43.6	36.3	41.0	35.7	42.5	34.3	42.1	35.7	42.5	35.0	42.0
EX	40.6	54.5	37.4	50.0	37.7	50.7	38.3	49.0	37.7	50.7	37.9	48.5
ES	54.5	49.3	51.5	48.1	52.8	46.9	51.5	47.8	52.8	49.9	51.6	48.0
ET	47.2	43.9	46.1	42.3	46.4	42.4	44.1	40.9	46.4	42.4	45.6	41.6
Overall	46.5	46.7	44.3	44.4	44.8	44.8	43.5	**44.2**	44.8	44.8	44.1	44.3

Table 7: Mixed WERs (%) for the four code-switched language pairs evaluated using the baseline language model.

better performance. The overall WER values for systems D and F show that batch-wise training results in better performance than processing all the untranscribed data in a single step. The best performing bilingual semi-supervised system (D) achieved an absolute overall WER improvement of 2.1% over the baseline on the test set.

7.2.2. Five-lingual semi-supervised training

The lower half of Table 7 indicates that the five-lingual semi-supervised acoustic models also benefited from the additional data. As for the bilingual systems, the five-lingual system yielded better results when using batch-wise training. The best performing system (J) outperformed the baseline by 2.5% absolute on the test set. Although the WER achieved by the five-lingual system is higher than that achieved by the bilingual system, this remains a promising result. Five-lingual recognition is more difficult since it allows more freedom in terms of the permissible language switches. It does, however, more honestly reflect the large and undetermined mix of languages an ASR system may be confronted with when processing South African speech.

7.2.3. Bilingual semi-supervised training with five-lingual transcriptions & semi-supervised language model

The bilingual acoustic models retrained with transcriptions generated by the five-lingual system (N) achieved the best overall WER: 40.56% which is an absolute improvement of 0.7% over system D, its closest competitor. This improvement may be due to the five-lingual system's ability to transcribe in more than two languages, since the untranscribed soap opera speech is known to contain at least some utterances that do not conform to the four bilingual systems. System N was also evaluated in combination with the semi-supervised language model. The combination of the semi-supervised acoustic and language models (N_{LM1}) reduced the overall WER on the test set by another 0.5% absolute.

Due to computational constraints, additional language model experiments were only conducted on the EZ data set. Table 8 shows that the use of language models derived from the additional text generated by the LSTM model, N_{LM2}, and the synthesised trigrams including additional text generated by the LSTM model, N_{LM3}, resulted in further small reductions in WER for both the development and test sets.

Moreover, the corresponding code-switch bigram accuracy (row 2 in Table 8) also improved when the language model training data included the additional, artificially generated text and trigrams.

	N_{LM1}		N_{LM2}		N_{LM3}	
	Dev	Test	Dev	Test	Dev	Test
WER	31.3	36.9	30.3	36.7	30.5	**36.2**
CSBA	42.3	40.1	45.1	41.0	45.9	**42.1**

Table 8: Mixed WER (%) and code-switched bigram accuracy (CSBA) (%) for EZ using artificial and synthetic text.

7.3. Language specific WER analysis

For code-switched ASR, the performance of the recogniser at the code-switch points is of particular interest. Language specific WERs and code-switched bigram accuracy values for the different systems are presented in Table 9. All values are percentages.

System	EZ			EX			ES			ET		
	E	Z	CSBA	E	X	CSBA	E	S	CSBA	E	T	CSBA
A (baseline)	37.9	48.7	33.3	37.8	54.5	25.8	43.7	61.4	25.2	36.2	51.8	35.6
D	32.1	43.6	39.8	31.3	48.3	33.9	33.0	55.9	34.8	27.6	47.6	42.3
G (baseline)	29.6	54.3	16.3	34.1	70.1	7.3	29.3	65.2	7.5	23.9	57.16	11.3
J	28.2	52.7	16.6	29.2	64.0	11.7	28.8	62.9	7.7	21.5	53.8	13.8
N	30.3	42.9	40.2	30.3	47.2	32.8	33.6	55.8	35.5	28.1	46.0	42.2
N_{LM1}	29.0	42.9	40.1	29.9	47.0	32.7	32.2	55.8	35.8	26.3	46.1	43.6

Table 9: Language specific WER (%) for English (E), isiZulu (Z), isiXhosa (X), Sesotho (S), Setswana (T) and code-switched bigram accuracy (CSBA) (%) for the test set.

The table reveals that five-lingual ASR (systems G & J) is more biased towards English (lower WER) than the bilingual systems (systems A & D), for which the Bantu language WERs are better. As already observed for the language model, the bias of the five-lingual system towards English is due to the much larger proportion of in-domain English training material available when pooling the four code-switched language pairs.

The accuracy at the code-switch points is substantially better when using the bilingual semi-supervised system. This is probably due to the higher ambiguity encountered by the five-lingual system at code-switch points than the bilingual systems. The table also reveals that the improvements observed for systems using the semi-supervised language models are due mostly to improved English recognition (N_{LM1}).

8. Conclusions

In this study we evaluated semi-supervised acoustic and language model training with the aim of improving the ASR performance of under-resourced code-switched South African speech. Two batches (approximately 53 hours in total) of manually segmented but untranscribed soap opera speech, rich in code-switching, were used for experimentation. The new speech was processed both in a single step and in batches by two automatic transcription systems: one comprising four parallel bilingual recognisers and the other a single five-lingual system.

The results indicate that the overall WER of both bilingual and five-lingual systems was reduced substantially and that batch-wise training was the better approach in both instances. However, the overall average performance of the bilingual systems remained better than that of the five-lingual system. This is probably because the five-lingual recognition task is inherently more complex.

The five-lingual system exhibited a bias towards English while the four bilingual recognisers were more accurate for the Bantu languages. Despite the confusability inherent in decoding five languages, the five-lingual system achieved an error rate that was almost as good as that attained by the bilingual systems. Thus, it seems worthwhile developing bilingual and five-lingual code-switched ASR systems in parallel.

The semi-supervised language model resulted in a significant reduction in perplexity that translated into a corresponding decrease in WER. Artificially generated code-switched text and synthetic trigrams also showed potential to further improve the ASR performance.

Future work will focus on incorporating automatic segmentation as well as speaker and language diarisation to extend the pool of training data. The effect of the number and size of training batches on system performance will also be studied more carefully.

9. 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.

10. References

Ahmed, B. H. and Tan, T.-P. (2012). Automatic speech recognition of code switching speech using 1-best rescoring. In *Proc. IALP*, pages 137–140. IEEE.

Amazouz, D., Adda-Decker, M., and Lamel, L. (2017). Addressing code-switching in French/Algerian Arabic speech. In *Interspeech*, pages 62–66.

Biswas, A., de Wet, F., van der Westhuizen, E., Yılmaz, E., and Niesler, T. (2018a). Multilingual Neural Network Acoustic Modelling for ASR of Under-Resourced English-isiZulu Code-Switched Speech. In *Proc. Interspeech*, pages 2603–2607.

Biswas, A., van der Westhuizen, E., Niesler, T., and de Wet, F. (2018b). Improving ASR for code-switched speech in under-resourced languages using out-of-domain data. In *Proc. SLTU*, pages 122–126.

Biswas, A., Yılmaz, E., de Wet, F., van der Westhuizen, E., , and Niesler, T. (2019). Semi-supervised acoustic model training for five-lingual code-switched ASR. In *Proc. Interspeech*, pages 3745–3749.

Drugman, T., Pylkkonen, J., and Kneser, R. (2019). Active and semi-supervised learning in ASR: Benefits on the acoustic and language models. *arXiv preprint arXiv:1903.02852*.

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*, pages 448–455. IEEE.

Guo, P., Xu, H., Xie, L., and Chng, E. S. (2018). Study of semi-supervised approaches to improving English-Mandarin code-switching speech recognition. In *Proc. Interspeech*, pages 1928–1932.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proc. Interspeech*, pages 3586–3589.

Li, Y. and Fung, P. (2013a). Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints. In *Proc. ICASSP*, pages 7368–7372. IEEE.

Li, Y. and Fung, P. (2013b). Language modeling for mixed language speech recognition using weighted phrase extraction. In *Proc. Interspeech*, pages 2599–2603.

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*, pages 1480–1484.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The Kaldi speech recognition toolkit. In *Proc. ASRU*. IEEE Signal Processing Society.

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*, pages 3743–3747.

Saon, G., Soltau, H., Nahamoo, D., and Picheny, M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In *Proc. ASRU*, pages 55–59.

Stolcke, A. (2002). SRILM – An extensible language modeling toolkit. In *Proc. ICSLP*, pages 901–904.

Taneja, K., Guha, S., Jyothi, P., and Abraham, B. (2019). Exploiting monolingual speech corpora for code-mixed speech recognition. *Proc. Interspeech*, pages 2150–2154.

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 *in Proc. ICASSP*, pages 6704–6708.

van der Westhuizen, E. and Niesler, T. (2018). A first South African corpus of multilingual code-switched soap opera speech. In *Proc. LREC*, pages 2854–2859.

van der Westhuizen, E. and Niesler, T. (2019). Synthesised bigrams using word embeddings for code-switched ASR of four South African language pairs. *Computer Speech & Language*, 54:151–175.

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*, pages 4889–4892. IEEE.

Yılmaz, E., Biswas, A., van der Westhuizen, E., de Wet, F., and Niesler, T. (2018a). Building a unified code-switching ASR system for South African languages. In *Proc. Interspeech*, pages 1923–1927.

Yılmaz, E., Heuvel, H. v. d., and van Leeuwen, D. (2018b). Acoustic and Textual Data Augmentation for Improved ASR of Code-Switching Speech. In *Proc. Interspeech*, pages 1933–1937.

Yılmaz, E., McLaren, M., van den Heuvel, H., and van Leeuwen, D. (2018c). 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*.