Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

5th Workshop on Spoken Language Technology for Under-resourced Language, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Automatic Speech Recognition of English-isiZulu Code-switched Speech from South African Soap Operas

Ewald van der Westhuizen*, Thomas Niesler*

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

Abstract

We introduce a new English-isiZulu code-switched speech corpus compiled from South African soap opera broadcasts. isiZulu itself is currently under-resourced, and automatic speech recognition is made even more challenging by the high prevalence of code-switching in spontaneous speech. Analysis of the corpus reflects effects common in conversational isiZulu, such as vowel deletion and cross-language prefixes and suffixes. Baseline monolingual and code-switched automatic speech recognition systems are developed, including a new language model configuration that explicitly includes switching transitions. For code-switched speech, a system with language-dependent acoustic models and language-dependent language models linked by switching transitions leads to best performance, although word error rates overall remain very high.

Peer-review under responsibility of the Organizing Committee of SLTU 2016.

Keywords: multilingual speech recognition; code-switching; isiZulu

1. Introduction

South Africa is a multilingual country with 11 official languages. A high degree of geographical overlap exists between these languages, with English as the *lingua franca*. This leads to the prevalence of code-switching (CS), which is the use of two or more languages in an utterance or conversation [1] and which occurs naturally in multilingual communities. The development of automatic speech recognition (ASR) for code-switched speech is a current research challenge, and is constrained by the difficulty of obtaining representative data for acoustic and language model training.

South African television soap operas exhibit interesting examples of code-switching in a deliberate effort to accommodate a language-diverse viewer base. We have compiled a new corpus of transcribed code-switched speech from such soap opera broadcasts to investigate the effect of code-switching on ASR systems, and for which one of the languages (isiZulu) is under-resourced. We present a set of first baseline experiments for the new corpus, and propose

* Corresponding author. Tel.: +27 (0)21 808 4936
E-mail address: ewaldvdw@sun.ac.za, trn@sun.ac.za

Peer-review under responsibility of the Organizing Committee of SLTU 2016.

Table 1. English-isiZulu code-switched corpus, indicating training, development and test sets.

	Train	*Dev*	*Test*
Duration	15.4h	8min	30.4min
Word tokens	198k	1.6k	5.7k
Word types	14.8k	900	2.5k
Utterance count	40.2k	225	768

a language modelling architecture which explicitly integrates code-switching. Currently, no word error rate (WER) evaluations of ASR systems have been reported for code-switched speech involving South African Bantu languages.

Section 2 reviews existing code-switched corpora. Sections 3 and 4 respectively introduce and analyse our English-isiZulu soap opera code-switched corpus. Section 5 describes the general experimental set-up and Section 6 presents monolingual systems and results. Section 7 proposes modelling architectures for code-switched speech, with associated experimental results in Sections 8. Section 9 concludes the paper.

2. Existing Code-Switch Corpora

A number of code-switch corpora have been described in the literature. We mention a few notable corpora below.

1. The SEAME corpus was compiled at Nanyang Technological University (NTU) in Singapore and at Universiti Sains Malaysia (USM). It contains 63 hours of spontaneous Mandarin-English code-switched conversational and interview speech uttered by Malaysian and Singaporean speakers[2,3,4].
2. The HKUST Mandarin-English Corpus was compiled at the Hong Kong University of Science and Technology (HKUST)[5,6]. It consists of code-switched spontaneous speech recorded from meetings and interviews, of which 5 hours have been transcribed and 15 hours are not fully transcribed.
3. The CECOS Chinese-English Corpus was compiled at the National Cheng Kung University in Taiwan[7]. It contains 12 hours of speech collected from 77 speakers uttering prompted code-switch sentences.
4. The CUMIX Cantonese-English speech corpus[8] was compiled at The Chinese University of Hong Kong. It contains 17 hours of code-switched speech read by 80 speakers.
5. A small English-Spanish (Spanglish) Corpus was compiled at the University of Texas from spontaneous lunch break conversations. The corpus contains 40 minutes of transcribed speech with a vocabulary of 1516 words[9].
6. A corpus of Sepedi-English code-switched speech was compiled by the South African CSIR[10]. It contains 10 hours of prompted speech, sourced from radio broadcasts and read by 20 Sepedi speakers.

3. Corpus Description

We have compiled a corpus of wideband English-isiZulu code-switched spontaneous speech. Each utterance is annotated orthographically and code-switching boundaries are delineated in time. All speech has been collected from broadcast South African soap operas.

Fluent bilingual speakers of isiZulu and English produced and validated the transcriptions. Where possible, soap opera scripts and subtitles, which are often included in view of the multilingual viewers base, assisted with transcription. Discussions with the soap opera producers and comparisons between scripts and uttered speech, confirmed that actors ad-lib heavily and regularly, yielding spontaneous and natural delivery of speech. This distinguishes the corpus from others using prompted and read speech. It also means that spontaneous and disfluent speech elements are common, e.g. contracted word pronunciations containing vowel deletions as a result of fast speech, especially in isiZulu speech segments. These were in all cases manually transcribed, and their complicating effect on ASR is one of the challenges presented by this type of speech. Overlapping speech and noise, such as background music and chatter and the sounds of props or actor movement, have been excluded from the corpus.

Table 2. Duration of corpus in hours. *t.eng*: total English, *t.zul*: total isiZulu, *m.eng*: monolingual English, *m.zul*: monolingual isiZulu, *cs.eng*: English segments in CS sentences, *cs.zul*: isiZulu segments in CS sentences.

Total	*t.eng*	*t.zul*	*m.eng*	*m.zul*	*cs.eng*	*cs.zul*
16.04	13.17	2.87	12.12	1.56	1.05	1.31

Table 3. Number of CS sentences containing different number of switches. *# switches*: number of switches, *Sent.count*: sentence count

# switches	1	2	3	4	5	6	> 7
Sent.count	1977	1107	399	178	66	36	23

(a)

beng'cel' uyek' uk'fak' i-	age	yam' nala ingangen' khona
zul	eng	zul

(b)

k'khon' ey'-	sign	-iw' ezansi	and those are paid for
zul	eng	zul	eng

Fig. 1. Code-switched examples from the corpus. (a) Insertional intrasentential code-switching with isiZulu as the matrix language. (b) Alternational and insertional intrasentential code-switching.

Table 1 lists the composition and sizes of the training, development and test sets for the complete corpus. No speaker overlap exists between the training and test sets. The development and test sets contain only code-switched sentences while the training set includes both monolingual and code-switched sentences.

4. Corpus Analysis

Table 2 shows the duration of the monolingual and code-switched portions for each language. Monolingual English dominates, covering 75% of the corpus. The code-switched portion is more evenly split.

Our corpus contains alternational and insertional code-switching, with two examples shown in Fig. 1. Transcription (a) exhibits insertional intrasentential code-switching with two switches. Here isiZulu is the matrix language and the English word "*age*" is embedded. Transcription (b) illustrates both alternational and insertional intrasentential code-switching with three switches. The word "*sign*" is an insertion, while the English and isiZulu segments are alternational. The alternational segments adhere to the grammar of the respective language. Our corpus contains 2743 isiZulu-English switches, and 2236 English-isiZulu switches.

Table 3 shows that, of the 3786 code-switched sentences, 1977 (52%) contain one language switch. The number of sentences declines rapidly as the number of switches increases. An average of 1.8 switches occur per CS utterance, while a switch occurs on average every 3.96 words.

The examples also show contracted forms of isiZulu words, with apostrophes indicating vowel deletion. Although isiZulu conforms to a regular /CV/ syllable structure, this may be interrupted at high speech rates, leading to two adjacent consonants. Such vowel deletion is common in Bantu languages. It is highly speaker dependent, and tends to increase in spontaneous, quickly-spoken utterances as in our corpus. The prevalence of contracted forms adds a complicating dimension to the development of an ASR system, especially since they do not occur in written form, such as may be used for language modelling. Contracted words are modelled in our pronunciation dictionary and used in our ASR experiments.

Using an automatically generated phoneme alignment, we find that the speech rate in our corpus is 14.61 phones per second (phones/s) for English and 18.45 phones/s for isiZulu. In comparison, the corresponding figures calculated on the NCHLT corpus[11], which was compiled using read prompts, are 10.62 phones/s for English and 9.04 phones/s for isiZulu. The speech rates in our corpus are notably higher than the read speech, especially for isiZulu.

IsiZulu prefixes and suffixes often occur with inserted English nouns, such as *ey*-sign-*iw'* in Fig. 1(b). Furthermore, English verbs may be agglutinated with isiZulu morphemes, forming compounds, such as *ung'-meet-e* and *ung'-treat-a* (not shown in Fig. 1). Among the 3768 transcribed code-switched sentences, there are 1530 instances of isiZulu prefixes preceding a switch to English and 422 instances of suffixes following a switch from English.

Table 4. CS type counts from a random selection of 50 CS sentences, *t-ez*: total switch count *eng* to *zul*, *t-ze*: total switch count *zul* to *eng*. CS types are *alt-ez*: alternate from *eng* to *zul*, *alt-ze*: alternate from *zul* to *eng*, *ie*: *eng* word insertion, *iz*: *zul* word insertion, *pre*: *zul* prefix, and *suf*: *zul* suffix.

CS type	*t-ez*	*t-ze*	*alt-ez*	*alt-ze*	*ie*	*iz*	*pre*	*suf*
count	37	51	9	12	36	7	26	7

Table 5. Word lengths for English and isiZulu segments in CS utterances. *seglen*: number of words in segment.

seglen	1	2	3	4	5	6	7	> 8
eng	2776	783	388	280	214	169	138	279
zul	1859	1502	988	534	292	186	90	124

Table 4 presents an analysis of the code-switching observed in a set of 50 randomly selected sentences from our corpus. We observe alternational switching occurs less often than insertional switching. The insertion of English words into isiZulu is much more common than the insertion of isiZulu words into English (36 versus 7). An isiZulu prefix preceded a switch 26 times among 51 switches from isiZulu to English. An isiZulu suffix followed a switch 7 times among 37 switches from English to isiZulu.

Table 5 shows the length in words of English and isiZulu segments in code-switched utterances for the full corpus. The high counts of single English words, together with higher counts of isiZulu segments with lengths 2 to 6, confirms the prevalence of English word insertions. For segments longer than 8 words, the count is higher for English. This indicates that, when alternation occurs, English segments contain more words than isiZulu segments. This may be attributed to the agglutinative nature of isiZulu, where morphemes are strung together to form long compounds. Compared to English, fewer words are used to convey the same amount of information. Differences in word length are relevant to ASR, because they affect language specific recognition parameters such as the word insertion penalty.

5. Experimental ASR Setup

Pronunciation dictionaries were based on a broadcast news system for English [12,13] and the NCHLT Corpora for isiZulu [11,14]. Unknown pronunciations were generated using the grapheme-to-phoneme (G2P) tools included with the NCHLT corpora.

In all experiments acoustic features were 13 MFCCs, including velocity and acceleration, yielding 39-dimensional feature vectors. Cepstral mean normalisation was applied per utterance.

Acoustic models consisted of standard 3-state left-to-right hidden Markov models (HMMs) as cross-word triphones. Cross-word contexts across languages are not allowed for the CS language dependent experiments, while cross-word contexts are allowed for the monolingual and language independent CS experiments. Decision tree state clustering was performed for each system. Systems parameters were optimised on a development set unless otherwise specified. The HTK toolkit [15] is used for ASR experiments.

A closed vocabulary, consisting of the training, development and test set vocabularies, is used in all experiments. The SRILM tools [16] are used to train language models.

6. Monolingual Systems

Two monolingual ASR systems were developed for English and isiZulu as baselines. For these systems, monolingual subsets of the full corpus were used for training and testing, as shown in Table 6. A development set is deliberately omitted to avoid further reduction in training set sizes, especially considering the limited size of the isiZulu training set. For this reason the test sets are used for recognition parameter optimisation. No speaker overlap exists between the training and test sets.

Bigram language models were trained on the training transcriptions. Monolingual web text corpora were initially included for language model training, but resulted in a considerable deterioration in WER, and were subsequently

Table 6. Monolingual subsets statistics. *wtok*: word token count, *wtyp*: word type count, *utts*: utterance count, *dur*: duration.

Language	*wtok*	*wtyp*	*utts*	*dur*
eng train	159k	6.7k	31.8k	11.7h
eng test	5k	1k	1k	25m
zul train	11k	4.4k	3.4k	1.2h
zul test	2.7k	1.5k	1k	20m

Table 7. Vocabulary sizes, perplexities and test set WER results for the monolingual systems.

Language	*vocab*	*PPL*	*WER*
eng	6.9k	118.8	61.06
zul	5.2k	2400.0	85.00

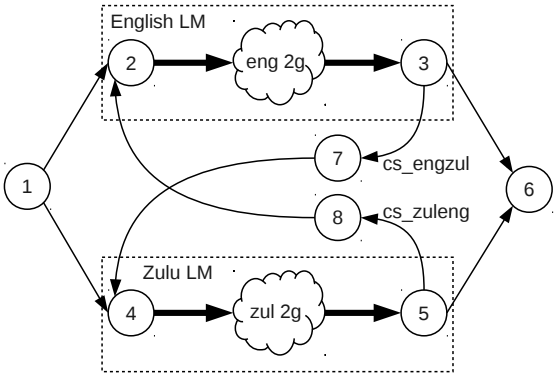

Fig. 2. Recognition grammar containing two monolingual bigram language models and explicit code-switch links, linking the exit node of each monolingual language model to the entry node of the other.

excluded. Table 7 shows the vocabulary sizes, perplexities calculated on monolingual test sets and WERs for the two monolingual systems.

7. Code-Switched Systems

In the following we describe ASR systems developed to process utterances that include code-switching.

7.1. Acoustic Modelling

Two acoustic model configurations were considered.

1. Language independent acoustic models (LIAM) use a universal phoneme set. Phoneme data are pooled across the two languages, and phoneme labels apply across languages, e.g. phoneme model ‘ax’ is trained on both English and isZulu examples of ‘ax’. Decision tree state clustering is performed on this universal phoneme set.
2. Language dependent acoustic models (LDAM) keep separate acoustic models per language. Phoneme data are not pooled across languages and English and isiZulu acoustic models are trained separately. Decision tree state clustering is performed separately on each language specific phoneme set.

7.2. Language Modelling

Language models are again trained on the training set transcriptions. Two configurations are compared.

Table 8. Development and test set WER results for English-isiZulu code-switched ASR baseline systems.

	LILM Dev	*LILM Test*	*LDLM Dev*	*LDLM Test*
LIAM	74.94	83.99	74.68	84.18
LDAM	71.18	81.99	71.79	**80.59**

1. Language independent language models (LILM) are trained on the pooled training set transcriptions, which containing both monolingual and code-switched sentences. The perplexity as calculated on the test set transcriptions is 3596.2.
2. Language dependent language models (LDLM) consists of two monolingual sub-models as depicted in Fig. 2. Each sub-model is trained on monolingual sentences and monolingual segments from code-switched sentences in the respective language drawn from the training transcriptions. The two sub-models are placed in parallel, with code-switch transitions linking the exit node of each to the entry node of the other to enable language switching. The perplexity as calculated on the test set transcriptions is 510.5.

8. Experimental Results and Discussion

Four English-isiZulu code-switching configurations are compared:

1. LIAM-LILM: universal acoustic model set with language independent language model,
2. LIAM-LDLM: universal acoustic model set with language dependent language model,
3. LDAM-LILM: language dependent acoustic models with language independent language model, and
4. LDAM-LDLM: language dependent acoustic models with language dependent language model.

WER results for the development and test sets are show in Table 8. Acoustic model, language model and recognition parameters have been optimised on the development set.

Results show that the WERs for all configurations lie between the WERs of the monolingual systems, and remain high despite the larger training set. This confirms that code-switching adds complexity to the ASR problem. The results also indicate that the LDAM configurations outperform the systems with LIAM on average by 2.8% absolute on the test set. This is despite the smaller training sets for the LDAM. We can deduce that, even when code-switching occurs, the speakers tend towards the phonetic character of the language being spoken at the particular instant in time. Further, we also see that the LDLM configurations outperform the LILM configurations on average by 0.61% absolute, and 1.4% absolute when comparing the LDAM-LILM with LDAM-LDLM, on the test set.

By manual inspection of the recognitions output, it was observed that confusion occurs between phonetically similar words and phrases, e.g. between "*sisi*" and "*see see*", and between:
a) *bengithi ng'*-`send`-*ela wen' i*-`sms` *was' eyayithola*, and
b) *bengithi usello nawe* `mesms` *kwase eyithola*,
where a) is the reference transcription and b) the recognition output. Language models trained on larger corpora may reduce this problem.

Confusion also occurs between compounds derived from similar roots, and between contracted and canonical forms of words, e.g. *yokuthi* with *ukuth'* and *kulungile* with *ulungil'*. Converting contractions to canonical form or using a scoring tool with support for word classes would lead to more optimistic results. Morphological decomposition before language modelling is also a possible course of action.

9. Conclusion

We have introduced a new corpus of spontaneous conversational English-isiZulu code-switched speech. The corpus incorporates phenomena such as vowel deletion and cross-language affixes that pose challenges above those already expected for an under-resources language such as isiZulu. Baseline ASR results were presented for monolingual English and isiZulu ASR systems, as well as for four configurations of code-switched English-isiZulu ASR systems.

A new language model configuration, consisting of sub-models connected by explicit switch transitions, was proposed and evaluated (LDLM). Experiments demonstrated that language dependent acoustic modelling outperforms language independent acoustic modelling on average by 2.8% absolute. Language dependent language modelling outperforms language independent language modelling on average by 0.61% absolute, and 1.4% absolute when comparing the LDAM-LILM with LDAM-LDLM, on the test set. Although the gain is currently minimal, it is promising if one considers that the context of the language model is currently not allowed to extend across the switch transition. For the LILM, such cross-language contexts do exist because the language modelling data is simply pooled. Hence, despite a loss in context across language transition, the strategy of combining sub-language-models with explicit language transitions is successful. In ongoing work, we are trying to address this shortcoming.

Acknowledgements

Computations were performed using Stellenbosch University's Rhasatsha High Performance Computer: http://www.sun.ac.za/hpc.

References

1. Dulm, O.V.. *The grammar of English-Afrikaans code switching : a feature checking account*. LOT Publications; 2007.
2. Vu, N.T., Lyu, D.C., Weiner, J., Telaar, D., Schlippe, T., Blaicher, F., et al. A first speech recognition system for Mandarin-English code-switch conversational speech. In: *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. ISBN 1520-6149; 2012, p. 4889–4892.
3. Dong, M., Chan, P., Cen, L., Li, H., Teo, J., Kua, P.J.. SEAME: A Mandarin-English Code-Switching Speech Corpus in South-East Asia. In: *Proceedings of INTERSPEECH*. 2010, URL: `http://www.isca-speech.org/archive/interspeech_2010`.
4. Adel, H., Vu, N.T., Kirchhoff, K., Telaar, D., Schultz, T.. Syntactic and Semantic Features for Code-Switching Factored Language Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 2015;**23**(3):431–440.
5. Li, Y., Yu, Y., Fung, P.. A Mandarin-English Code-Switching Corpus. In: *Proceedings of LREC*. ISBN 978-2-9517408-7-7; 2012, .
6. Li, Y., Fung, P.. Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. ISBN 1520-6149; 2013, p. 7368–7372.
7. Shen, H.P., Wu, C.H., Yang, Y.T., Hsu, C.S.. CECOS: A Chinese-English Code-Switching Speech Database. In: *Proceedings of Oriental COCOSDA*. 2011, .
8. Chan, J.Y.C., Ching, P.C., Lee, T.. Development of a Cantonese-English Code-Mixing Speech Corpus. In: *Proceedings of INTERSPEECH*. 2005, .
9. Franco, J., Solorio, T.. Baby-steps towards building a Spanglish language model. In: *Proceedings of CICLing*. Springer-Verlag Berlin Heidelberg; 2007, p. 75–84.
10. Modipa, T.I., Davel, M.H., de Wet, F.. Implications of Sepedi/English code switching for ASR systems. In: *Proceedings of PRASA*. 2013, .
11. Barnard, E., Davel, M.H., Heerden, C.V., Wet, F.D., Badenhorst, J.. The NCHLT speech corpus of the South African languages. In: *Proceedings of SLTU*. 2014, .
12. Kamper, H., de Wet, F., Hain, T., Niesler, T.. Resource development and experiments in automatic South African broadcast news transcription. In: *Proceedings of SLTU*. 2012, .
13. Kamper, H., de Wet, F., Hain, T., Niesler, T.. Capitalising on North American speech resources for the development of a South African English large vocabulary speech recognition system. *Computer Speech & Language* 2014;**28**(6):1255–1268.
14. Davel, M., Basson, W., Heerden, C.V., Barnard, E.. NCHLT Dictionaries: Project Report. Tech. Rep.; North-West University; 2013. URL: `https://sites.google.com/site/nchltspeechcorpus/`.
15. Young, S.J., Evenmann, G., Gales, M.J.F., Hain, T., Kershaw, D., Liu, X., et al. *The HTK Book, Version 3.4*. Cambridge, UK: Cambridge University Engineering Department; 2009.
16. Stolcke, A., Zheng, J., Wang, W., Abrash, V.. SRILM at Sixteen: Update and Outlook. In: *Proceedings of ASRU*. 2011, .