*INTERSPEECH 2016*
September 8–12, 2016, San Francisco, USA

# The Effect of Postlexical Deletion on Automatic Speech Recognition in Fast Spontaneously Spoken Zulu

*Ewald van der Westhuizen, Thomas Niesler*

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

ewaldvdw@sun.ac.za, trn@sun.ac.za

## Abstract

We consider the phenomenon of postlexical deletion in fast spontaneously spoken isiZulu speech and its implication for automatic speech recognition (ASR). Analysis of hand-crafted transcripts of fast spontaneous speech recorded from broadcast media indicates that postlexical deletion, especially of vowels, is common in isiZulu. We show that ASR performance can be increased by inclusion of pronunciation variants that model such deletions. We also apply a sequence modelling approach normally used for grapheme-to-phoneme (G2P) conversion to generate orthography containing synthetic deletions. These synthetically generated contacted words are subsequently used to generate accompanying pronunciations using conventional G2P conversion. We evaluate an ASR system using these synthetically generated pronunciations, and compare it to a baseline system without such variants as well as an oracle system. Augmentation with synthetically generated pronunciations leads to an absolute improvement in word error rate (WER) of 2.36% relative to the baseline. Furthermore, the augmented system performs almost as well as the oracle system, with an absolute difference in WER of 0.38%.

**Index Terms**: spontaneous speech, pronunciation alternatives, vowel deletion, elision, ASR

## 1. Introduction

IsiZulu is one of 11 official languages in South Africa. Although a small number of speech corpora are available for this language, the issue of fast spontaneously spoken isiZulu speech has not been addressed. Fast, spontaneous speech often deviates from the regular phonotactical constraints, and such deviations are observed in a speech corpus we have compiled from South African soap opera broadcasts.

One may distinguish between lexical and postlexical phonological phenomena. The former typically apply to words in isolation, as they appear in the lexicon. The latter apply to words in the context of a sentence, for instance where the pronunciation of a word is influenced by adjacent words. Postlexical rules are commonly found in text-to-speech systems, to ensure natural sounding reproduction of words in context.

Fast, spontaneous speech often exhibits optional deletions as a postlexical phenomenon. Such deletions may take the following forms:

- aphesis, dropping initial sounds from a word, e.g. *'cept* for *accept* or *except*;
- apocope, dropping one or more sounds from the end of a word, e.g. *Vince* for *Vincent*; and
- syncope, dropping sounds internal to the word, e.g. *fo'c's'le* for *forecastle* is a classic example.

Unstressed vowels are most commonly deleted. These postlexical deletions occur as a matter of course in fast spontaneously spoken isiZulu, and are witnessed in our corpus.

The effect of postlexical deletion on automatic speech recognition (ASR) performance with fast spoken isiZulu has not been investigated before. We perform such an investigation. Futhermore, we investigate the modelling of postlexical deletion with a sequence modelling tool. The sequence models are used to predict contracted orthographic alternatives which are better matched to fast, spontaneous speech, leading to improved ASR performance.

Section 2 presents background information, followed by a description of the isiZulu speech corpus in Section 3. An analysis of postlexical deletion is discussed in Section 4, followed by deletion modelling and prediction in Section 5. Section 6 discusses the experimental ASR setup with results in Section 7. Section 8 discusses the results and concludes the paper.

## 2. Background

Although the occurrence of postlexical deletion has received very little attention in the literature, it was already mentioned in [1], published in 1857. This text describes the use of apostrophes to indicate deletions in spoken isiZulu. More recently, vowel deletions are discussed in a formally structured lexical context, when morphemes are agglutinated to form compound words [2]. In contrast, our data exhibits apocope and syncope, where sounds are deleted as a result of fast speech. The most explicit mention of the phenomenon was found in [3, p. 156-157], where it is said that "*in fast spoken speech only, Post-Lexical Deletion ... appl[ies] optionally.*"

IsiZulu predominantly follows a /CV/ syllable structure, e.g. *uyakuthanda* with /V/CV/CV/CV/CV/ and *isikole* with /V/CV/CV/CV/. In fast speech, unstressed vowels may be deleted, causing a succession of consonants, e.g. *uyakuthanda* and *isikole* pronounced as *uyak'thand'* and *is'kol'*, violating the regular syllable pattern.

Contracted word forms, using the apostrophe as in our examples above, are used in isiZulu poetry as a literary device [4]. Postlexical deletion is both speaker and domain dependent, and the tone and environment of a conversation can dictate its occurrence, e.g. a formal discourse versus informal conversation.

The agglutinative morphology of isiZulu presents another set of challenges not specifically addressed in this paper. It relates to the vocabulary growing without bound as the size of the data set grows. This also happens in other agglutinative languages such as Finnish, Estonian and Turkish [5]. Agglutination results in high language model (LM) perplexities compared to languages with analytic properties.



http://dx.doi.org/10.21437/Interspeech.2016-820

Table 1: Corpus training and test sets, indicating *spk*: (speaker count), *wtok*: (word token count), *wtyp*: (word type count), *utts*: (utterance count) and *dur*: (duration).

| *Set* | *spk* | *wtok* | *wtyp* | *utts* | *dur* |
|---|---|---|---|---|---|
| train | 50 | 11k | 4.4k | 3.4k | 1.2h |
| test | 4 | 2.7k | 1.5k | 1k | 20m |

Table 2: Counts of graphemes deleted to yield contracted word forms in the training transcriptions.

| count | graphemes | count | graphemes |
|---|---|---|---|
| 1842 | i | 15 | ke |
| 816 | a | 12 | k,in |
| 643 | e | 7 | w,wu |
| 611 | u | 4 | wa |
| 231 | o | 2 | ia,ne,to,be |
| 61 | ku | 1 | thi,khu,kho,ing,yo,si,to,lu |
| 18 | ow | | we,phuku,ma,ile,mi,n,deni |

## 3. Corpus Description

Our corpus contains 1.5 hours of spontaneously spoken monolingual speech recorded from South African soap opera broadcasts, and forms part of an on-going transcription effort. First language isiZulu speakers manually transcribed all speech orthographically. Word labels include apostrophes indicating deletions in utterances, together with the canonical spelling. For example, the sentence:
*Ngiyabonga futhi ukuthi unakekele uThandeka*,
when uttered at speed, is transcribed as:
*Ng'yabonga futh' ukuth' unakekel' uThandeka*,
with apostrophes indicating the deletion of *'i'*, *'i'*, *'i'* and *'e'*, respectively.

Table 1 shows the corpus size and counts. The allocation of a development set is omitted to avoid a further reduction in training set size. Therefore, the test set is used to optimise recognition parameters in ASR experiments. No speaker overlap exists between the training and test sets.

The speech rate is calculated at 18.45 phones/s from automatically generated phoneme alignments. This is more than twice as fast as the 9.04 phones/s speech rate calculated for the prompted isiZulu speech in the NCHLT corpus [6].

## 4. Analysis of Deletions

Table 2 shows the counts of graphemes deleted in the training transcriptions. It is clear that vowels are deleted more often than consonants. Grapheme *'i'* is deleted most, more than twice as often as *'a'*, and about three times as often as *'e'* and *'u'*.

Figure 1 shows counts of start positions of deletions in the canonical forms of the affected words. Positions 0, 1, ..., $i$ denote the first, second, to $(i+1)^{\text{th}}$ letter positions from the beginning of a word, while positions $-1$, $-2$, ..., $-i$ denote the last, second to last, to $i^{\text{th}}$ to last letter positions in a word, respectively. We see most deletions occur at the end of words (position $-1$), with only 8 deletions occurring at the start of words (position 0). The second largest number of deletions occur at the third letter from the word beginning. This trend is also illustrated by the transcription examples in Sections 1, 3 and 5.

## 5. Automatic Prediction of Deletions

We use the *Sequitur G2P* sequence modelling tool [10] to model and predict the occurrence of postlexical deletion. We first analyse prediction errors in a cross-validation framework performed on an isolated list of contracted words. Thereafter, we train and

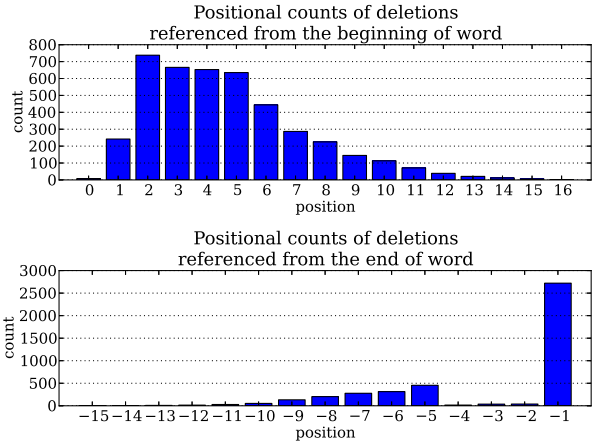


Figure 1: Counts of deletion positions.

Table 3: Analysis of sequence model predicted deletions using isolated contracted word list (Section 5.1).

| | | Model order | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| 1 | % word errors | 76.72 | **51.22** | 53.77 | 57.23 |
| 2 | % correct dels | 41.35 | 82.26 | **83.27** | 77.36 |
| 3 | ratio dels gen vs. ref | 0.79 | 1.20 | 1.23 | 1.15 |
| 4 | % words excess dels | **17.76** | 31.49 | 34.76 | 29.00 |
| 5 | % words lack dels | 42.63 | 10.80 | **9.07** | 13.68 |
| 6 | % words zero dels | 33.70 | 6.43 | **5.42** | 8.45 |

evaluate sequence modelling on the complete corpus, investigating two configurations: (i) word internal context training, evaluated on the corpus test set; and (ii) full sentence training with cross-word context, evaluated on the corpus test set. Finally, we describe the process of creating synthetically generated contracted pronunciations for use in an ASR pronunciation dictionary.

### 5.1. Cross-validation analysis on isolated contracted words

A list of contracted words is extracted from the corpus transcriptions, together with the respective canonical spelling, comprising 2097 unique entries. Whitespaces are inserted to separate the graphemes of the contracted reference spellings, yielding a graphemic dictionary with entries of the form:

```
akakujabeleli   a k a k ' j a b ' l e l '
```

The list is now split randomly into 10 approximately equal partitions. In a 10-fold cross-validation framework, sequence models are trained on 9 partitions, reserving the $10^{\text{th}}$ for testing. The model order is increased step by step during training. Trained sequence models are applied to each canonical spelling in the held-out test set to produce a hypothesised contracted spelling. The output hypotheses are scored against the reference contracted spellings to calculate error rates.

Table 3 reflects accuracy and error statistics, averaged over the ten folds, for model orders 1 to 4. We note that the balance between best accuracy and smallest number of errors lies somewhere between orders 2 and 3. Order 2 generates the most correctly contacted word forms (row 1), while order 3 is able to place deletions at the correct positions most reliably (row 2).

Row 3 presents a ratio giving the total number of predicted versus reference deletions. A ratio of 1.0 represents an equal number of predicted and reference deletions. Order 1 on average predicts too few deletions, while orders 2 to 4 are too generous.

Table 4: Word and sentence error rates for sequence model predicted transcriptions evaluated against reference transcriptions.

| Order | SLWI | | SLXW | |
|---|---|---|---|---|
| | WER | SER | WER | SER |
| 1 | 36.0 | 62.9 | 36.0 | 62.9 |
| 2 | 37.5 | 64.2 | 35.5 | 61.3 |
| 3 | 35.0 | 64.5 | **23.9** | **48.7** |
| 4 | 42.0 | 70.8 | 33.8 | 55.6 |

The errors shown in rows 4, 5, and 6, indicate the percentage of words containing an excess of, a lack of, and zero deletions compared to their reference words, respectively. Order 1 has the fewest excess errors due to fewer predicted deletions (row 4), while order 3 has the fewest errors for predicted words which both lack (row 5) and contain zero (row 6) deletions.

Considering rows 1, 2 and 4, we confirm that the high accuracy of correctly predicted deletions of order 3 can be attributed to the generous allocation of deletions, but this also increases the total word errors compared to order 2, since excessive deletion cause more word errors.

### 5.2. Corpus sentence level word internal context (SLWI)

We now evaluate deletion prediction at sentence level for the complete corpus test set, i.e. not just the contracted forms. Sequence model training is performed at word level (no cross-word context) using all word types in the corpus training set transcriptions. The graphemic training dictionary contains transcribed words with and without deletions, e.g.:

```
akakujabeleli   a k a k u j a b u l e l i
akakujabeleli   a k a k ' j a b ' l e l '
akudonse   a k u d o n s e
akudonse   a k ' d o n s e
```

Table 4 shows word and sentence error rates (WER, SER) of the predicted transcriptions when evaluated against the corpus test set sentences (columns SLWI). The first order model fails to predict any deletions at all, and thus reflects the actual percentage of words in the test set containing deletions (36%). The SER increases as model order increases, as a result of the SER being a coarse-grained evaluation metric. A singly misplaced deletion causes a complete sentence error. Sentence errors as a result increase as deletions are predicted too liberally.

### 5.3. Corpus sentence level cross-word context (SLXW)

In an effort to improve the performance of the sequence models, and to test whether deletions are influenced by graphemes from adjacent words, we extend the context used by these models by performing sequence model training on full sentences of the complete corpus training set. Hence, now the graphemic training dictionary consists of full sentences and their graphemic transcriptions. Word boundaries are labeled with an '@', enabling sequence models to span word boundaries, e.g.:

```
ngicela@ungilalele n g ' c e l ' @ u n g ' l a l e l e
```

The cross-word sequence models operate at grapheme level and should not be confused with the concept of cross-word context-dependent phonemes as used in ASR.

Deletion prediction is performed on full sentence input rather than on words in isolation. Using this configuration, sequence models with order 3 gave the lowest WER of 23.9% (see Table 4). The improvements in WERs compared to the SLWI experiments, indicate that useful information exists at cross-word boundaries, and is learned by the sequence models. Manual side-by-side inspection of the generated output of SLWI and SLXW (orders 3), indicates that the majority of improvements occur at the ends of words. This also confirms that the graphemes located at the beginning of a word, influence deletions that may occur at the end of the preceding word.

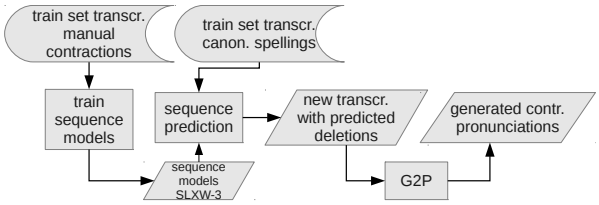


Figure 2: Process of training sequence models and generating contracted pronunciations via deletion prediction.

### 5.4. Synthetically generated pronunciations

Finally, we use model SLXW order 3 to predict deletions, resulting in synthesised contracted forms from which pronunciations can be generated. This process and its resources are depicted in Figure 2. Training set sentences with canonical orthography are passed through the sequence model generator, producing orthographic transcriptions containing predicted deletions. Note that not every input word necessarily has a transformed output variant. Even though we are using a closed vocabulary, we opt to generate the predicted transcriptions from the training set vocabulary only. This avoids an optimistic bias towards the transcribed test set words containing deletions. Contracted pronunciations used in the ASR pronunciation dictionary are generated only from the predicted contracted words in the generated transcriptions using the isiZulu NCHLT grapheme-to-phoneme (G2P) tool [11]. Canonical pronunciations are used for any out-of-dictionary test set words absent from the training set.

## 6. Experimental ASR Setup

The ASR system is built using HTK [12]. The system is similar to the monolingual ASR system described in [13]. Features used are 13 MFCCs, with velocity and acceleration included, yielding 39-dimensional vectors, with cepstral mean normalisation applied per utterance. Acoustic models consist of standard 3-state left-to-right hidden Markov models as cross-word triphones. Decision tree state clustering is performed. Acoustic models are trained and subsequently remain unchanged between recognition experiments below. System parameters are optimised on the corpus test set, for reasons given in Section 3. A closed vocabulary, consisting of the training and test set vocabularies, is used throughout.

The pronunciation dictionary always uses canonical pronunciations for canonical spellings. Three configurations are evaluated for words containing deletions.

- Contracted pronunciations (**p-contr**) are derived from the manually transcribed words with deletions using G2P, using one pronunciation per spelling.
- Canonical pronunciations (**p-canon**) are full word pronunciations without deletions, sourced from trusted pronunciation dictionaries or generated from canonical spellings with G2P, using one pronunciation per spelling.
- Synthetically generated pronunciations (**p-gen**) are generated via the process described in Section 5.4.

The SRILM tools [14] are used to train bigram LMs. Three LM configurations are used in our experiments.

- **LM1** is trained on the training set transcriptions containing orthography with and without deletions. The perplexity

Table 5: WERs for the ASR system configurations. *Trans.*: transcriptions either contracted (contr) or canonical (canon), *Pron.dict*: pronunciation dictionary, *Rec.par*: recognition parameters either optimised (optm) or as baseline system A (as A). Other abbreviations are defined in Section 6.

| | *Trans.* | *Pron.dict* | *LM* | *Rec.par* | *WER* |
|---|---|---|---|---|---|
| A | contr | p-contr | LM1 | optm | 85.00 |
| B | canon | p-canon | LM2 | as A | 83.53 |
| B.1 | canon | p-canon | LM3 | as A | 81.05 |
| B.2 | canon | p-canon | LM3 | optm | 80.77 |
| C | canon | p-canon, p-contr | LM2 | as A | 80.85 |
| C.1 | canon | p-canon, p-contr | LM3 | as A | 78.30 |
| C.2 | canon | p-canon, p-contr | LM3 | optm | 78.03 |
| D | canon | p-canon, p-gen | LM2 | as A | 81.37 |
| D.1 | canon | p-canon, p-gen | LM3 | as A | 78.86 |
| D.2 | canon | p-canon, p-gen | LM3 | optm | 78.41 |

(PPL) evaluated on the test set is 2400, with vocabulary size of 5258. We note that words containing deletions contribute to increased PPL.

- **LM2** is trained on the training set transcriptions containing only canonical orthography. A lower PPL of 1404.5 is obtained when evaluated on the test set, with vocabulary size of 4673, resulting from greater consistency of the canonical word forms.
- **LM3** is an interpolation of LM2 with an LM trained on 100k isiZulu web text sentences. We improve LM2 by LM interpolation, since canonical word forms are used. The interpolation factor is optimised on ASR System B and used in all system configurations. PPL is reduced to 716.5, with vocabulary size of 4673.

A number of ASR system configurations are investigated.

- System A is a first baseline system, where manually transcribed contracted spellings are used. As a baseline, this is the rawest form of the corpus. The p-contr pronunciations are used together with LM1 which supports contracted transcriptions.
- System B is second baseline which uses canonical spellings only. B employs a cleaner representation of the corpus compared to A. The p-canon pronunciations are used, together with LM2 and LM3, which support canonical spellings.
- System C uses canonical spellings only. The pronunciation dictionary includes p-canon and p-contr, i.e. all transcribed pronunciation variants, both canonical and with deletions. Both LM2 and LM3 are evaluated.
- System D uses canonical spellings only. The pronunciation dictionary includes p-canon and p-gen, i.e. both canonical pronunciations and synthetically generated pronunciation variant as generated by the sequence models. Both LM2 and LM3 are evaluated.

The numeral suffix .1 appended to the system label, e.g. B.1, indicates that LM3 was used and recognition parameters are similar to those of the baseline system. Numeral suffix .2 indicates that further recognition parameter optimisation was performed to achieve an optimal WER for the particular system. This enables direct comparison as to which system component led to a change in WER.

## 7. Results

Table 5 shows ASR results for the various system configurations. All WERs are very high. This is due to the small size of the corpus, and also to the difficult nature of the speech (fast and spontanous). For comparative purposes, we note that isiZulu ASR systems trained on read or prompted speech, have been reported to yield WERs of 62.99% (17 hours) [7] and 33.6% (50 hours) [6]. More in line with our results, other researchers have been met with similarly high error rates, such as the WERs of 70-76% and 61.2% achieved on the respective 10-hour and 80-hour portions of the Zulu contribution to the Babel corpora [8, 9], which consists of conversational speech.

The associated WER for System A is the highest, as is to be expected, since LM1 contains word forms with deletions which increase perplexity. System B displays an absolute improvement of 1.47% and uses LM2, which includes exclusively canonical word forms. Pronunciations are strictly canonical, and as a result may be suboptimal. In particular, deletions are not modelled in the p-canon pronunciations. System C improves on System B by 2.68%, which is exclusively attributed to the addition of pronunciations that model deletions. The incorporation of LM3 and further recognition parameter optimisation, results in the lowest WER of 78.03%; a nearly 7% absolute improvement compared to System A. We note that System C is an oracle experiment, since manually transcribed test set vocabulary pronunciations are modelled in the pronunciation dictionary. We therefore regard the results from Systems C, C.1 and C.2 to be optimistic. Finally, System D uses pronunciations generated synthetically by the deletion prediction sequence modelling process shown in Figure 2. The WER presents a 2.16% absolute improvement over System B, and a 2.36% absolute improvement when comparing the parameter optimised systems of B.2 and D.2. Furthermore, the optimistic performance of System C.2 is only 0.38% absolute better than the synthetic System D.2.

## 8. Discussion and Conclusion

We have established the prevalence of postlexical deletion in fast spontaneously spoken isiZulu. We have further shown that the inclusion of pronunciation alternatives modelling postlexical deletion in isiZulu leads to improved ASR performance. This stands in contrast to experiences in some cases of ASR, that the use of alternative pronunciations comes at a cost of increasing confusion during speech recognition, and therefore may lead to deteriorated performance. Besides the successful modelling of pronunciations containing deletions from manual orthographic transcriptions, we have shown that the phenomenon of postlexical deletion itself can be modelled statistically. Sequence models can be used to generate synthetically contracted forms of words to model postlexical deletion. These contracted forms can in turn be used to derive pronunciations using G2P. The inclusion of such synthetically generated pronunciations is demonstrated to lead to improved recognition performance.

Future work include the postlexical deletion modelling using decision trees, applying the modelling of postlexical deletion to other spontaneous speech corpora to establish possible ASR performance improvements, and the investigation into using an isiZulu morphological decomposer to assess whether language modelling at morpheme level will lead to further improvements in ASR performance.

## 9. Acknowledgements

We would like to thank e.tv and Yula Quinn at Rhythm City for assistance with data compilation. Computations were performed using the University of Stellenbosch's Rhasatsha HPC: http://www.sun.ac.za/hpc

# 10. References


[1] L. Grout, *The IsiZulu: A Grammar of the Zulu Language*. Pietermaritzburg: May and Davis, 1859.

[2] G. Poulos and C. T. Msimang, *A Linguistic Analysis of Zulu*, 1st ed. Pretoria: Via Afrika, 1998.

[3] J. S. M. Khumalo, "An Autosegmental Account of Zulu Phonology," Ph.D. dissertation, University of Witwatersrand, 1987.

[4] T. P. Ngwenya, M. B. Shongwe, and P. G. Shabalala-Strydom, *Sisikha Embezeni isiZulu*. Cape Town: New Africa Books, 2003.

[5] M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraçlar, "Unlimited vocabulary speech recognition for agglutinative languages," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 2006.

[6] E. Barnard, M. H. Davel, C. van Heerden, F. D. Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *4th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, May 2014.

[7] D. Henselmans, D. van Leeuwen, and T. R. Niesler, "Baseline Speech Recognition of South African Languages using Lwazi and AST," in *Proceedings of the twenty-fourth annual symposium of the Pattern Recognition Association of South Africa (PRASA)*, 2013.

[8] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low resource languages: Babel project research at CUED," in *Spoken Language Technologies for Under-Resourced Languages (SLTU)*, 2014.

[9] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[10] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 5 2008.

[11] M. Davel, W. Basson, C. van Heerden, and E. Barnard, "NCHLT Dictionaries: Project Report," North-West University, Tech. Rep., 2013. [Online]. Available: https://sites.google.com/site/nchltspeechcorpus/

[12] S. J. Young, G. Evenmann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, Version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2009.

[13] E. van der Westhuizen and T. R. Niesler, "Automatic Speech Recognition of English-isiZulu Code-switched Speech from South African Soap Operas," in *Workshop onSpoken Language Technologies for Under-Resourced Languages (SLTU)*, 2016.

[14] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "SRILM at Sixteen: Update and Outlook," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.