# Evaluating acoustic modelling of lexical stress for Afrikaans speech synthesis

Daniel R. van Niekerk
Multilingual Speech Technologies, North-West University, South Africa.
CAIR, CSIR Meraka, South Africa.
Email: daniel.vanniekerk@nwu.ac.za

***Abstract*—An explicit lexical stress feature is investigated for statistical parametric speech synthesis in Afrikaans: Firstly, objective measures are used to assess proposed annotation protocols and dictionaries compared to the baseline (implicit modelling) on the Lwazi 2 text-to-speech corpus. Secondly, the best candidates are evaluated on additional corpora. Finally, a comparative subjective evaluation is conducted to determine the perceptual impact on text-to-speech synthesis. The best candidate dictionary is associated with favourable objective results obtained on all corpora and was preferred in the subjective test. This suggests that it may form a basis for further refinement and work on improved prosodic models.**

***Index Terms*—pronunciation dictionary, under-resourced language, syllable-stress, lexical stress, Afrikaans, speech synthesis, text-to-speech**

## I. Introduction

Increasingly sophisticated machine learning (ML) techniques are currently state-of-the-art for acoustic modelling and speech generation in corpus-based text-to-speech (TTS) synthesis [1], [2]. In principle, some techniques are flexible enough to perform acoustic modelling directly from the natural language text, thereby potentially avoiding explicit intermediate signal or linguistic representations [3], [4]. Some advantages of this methodology are (1) avoiding uncertainties or errors in explicit intermediate representations, (2) the possibility of optimising larger parts of the (or complete) system directly with respect to the final output, and (3) reducing (or eliminating) the costs associated with the development of intermediate resources such as pronunciation dictionaries.

However, in practice, certain features or intermediate representations are usually explicitly defined; acoustic modelling for TTS typically involves several components and representations of both the audio signal (such as spectral envelope, fundamental frequency and segmental duration) and text (typically linguistically motivated words, syllables and phonemes) [1], [5]. When these representations have a *well-defined function* they may also serve as an additional interface for *input specification*. Thus, while potentially complicating the task of system-wide performance optimisation, such representations offer advantages in the following cases: (1) The scope of application may be increased by allowing different parts of the input to originate from different information sources. For example, in a dialogue system the context in which a sentence is generated may provide additional semantic information. (2) The tractability of modelling certain aspects of the output may be improved by reducing the sparsity of training contexts. For example, when different graphemes map to a single phone due to contextual differences or spelling variants. Both of these cases involve the inclusion of information originating externally to the immediate training data (i.e. speech corpus), with the former involving information obtained at run-time and the latter at the time of construction of the system.

In the context of technology development for so-called under-resourced languages, the benefits of both of the above scenarios are desirable. The possibility of reusing an existing system in new application domains is a welcome prospect given constraints on language or speech data and technical expertise. Furthermore, sparsity is a serious consideration during system design and development with limited training data (e.g. in TTS corpus development [6], [7]).

This paper aims to initiate work on an explicit specification of a lexical stress feature, which is part of an ongoing effort to enhance lexical pronunciation resources in Afrikaans specifically for TTS system development [8]. This is directly motivated by the two points made above and by potential future work on synthesis of higher-level prosody: (1) Regarding the scope of application, since the lexical stress pattern is sometimes involved in word-disambiguation (see Section II), even correctly synthesising isolated words may be difficult without explicit control over this aspect. (2) Regarding tractability, a lexical stress feature could improve over the situation where only simple "positional features" and extended phonetic contexts are used; such a feature is standard in English systems [5]. (3) It has been shown that the realisation of higher-level acoustic phenomena such as prosodic prominence may be dependent on lexical stress patterns [9].

However, whether the inclusion of a lexical stress feature will lead to better (or even comparable) acoustic models for typical TTS applications depends on a number of factors. The immediate corpus is the most direct factor, but others include the accuracy with which the feature can be specified, the nature of the associated "acoustic pattern" being modelled, and the training techniques and models used (especially with respect to built-in assumptions about the data or pattern). It is often difficult to justify including explicit linguistic features in complex speech and language systems, especially given a relatively narrow application domain and an appropriate training corpus [10] (see the advantages of ML approaches above).

Thus the specific contributions of this work are:

1) An objective evaluation of acoustic models over different protocols for including a lexical stress feature compared to implicit stress modelling. Proposals are made based on existing work and evaluated using the Lwazi 2 Afrikaans TTS corpus [11] (Table I).
2) Objective results on additional corpora with the aim of determining whether results in (1) hold on different corpora and to further analyse the scope of the results.
3) A subjective evaluation to determine whether the results have a significant impact on the perceived quality of a TTS system.

In the following section a brief background on lexical stress, especially descriptions and previous work relevant to TTS in Afrikaans, is presented. This is followed by proposals for explicit stress features in Section III and objective evaluations in Section IV. Section V presents a subjective TTS evaluation and Section VI is the conclusion with proposals for further work.

## II. Background and previous work

Stress is a broad concept in linguistics referring to the relative emphasis or prominence of a specific instance of a linguistic unit relative to another in a specified linguistic domain. For example, a stressed syllable in a word (lexical stress) or word in a phrase or sentence (prosodic stress or prominence). Lexical stress may be characterised in terms of how it is exhibited in different languages: so-called *fixed stress* patterns are highly regular and simple to predict, while *variable stress* patterns are tied to specific lexical items and cannot easily be generalised across the vocabulary. However, it is most relevant to note that in variable stress languages, the feature is usually *phonemic* (i.e. it has a word-disambiguating *function*).

Afrikaans, like other Germanic languages such as English and Dutch, can be described as a *variable stress* language; stress patterns depend on etymology and morphology to some extent [12]. Nevertheless, it is possible to predict stress patterns for words by considering orthographic patterns indicative of underlying morphological structure or etymology. The work of Mouton developed linguistically motivated rules for stress assignment in this way, assigning a single (primary) stress label to a syllable in simplex words and derivations [13]. Another description motivates the possibility of assigning a secondary stress label in compound words [12].

At present, the only freely available TTS system for Afrikaans that relies on an explicit stress feature is the *eSpeak* formant synthesiser.[1] This system uses a small pronunciation lexicon and large set of manually constructed letter-to-sound rules to simultaneously determine phonetic and stress features. Since the system is not corpus-based, this pronunciation specification is directly related to acoustic realisation (i.e. *phonetic*); the stress feature uses primary and secondary stress labels and not only serves to disambiguate words, but also to control other aspects of speech prosody.

TABLE I
TTS corpora

| Corpus | Utterances | Words | Speech duration |
|---|---|---|---|
| Lwazi 2 | 1005 | 9217 | 53.45 min. |
| Lwazi 3 (subset) | 3068 | 50150 | 252.81 min. |
| Multi-speaker (subset) | 1621 | 15337 | 93.71 min. |

## III. Lexical stress feature for TTS

This section presents some proposals for an explicit stress feature and identifies baseline resources which were considered for application in TTS.

From the motivation in Section I, an explicit stress feature should ideally improve the tractability of acoustic modelling and have a well-defined function. The former implies that the feature should accurately correspond to the acoustic reality in the data and the latter may guide decisions towards a representation that has a greater prospect of reuse in different contexts.

An immediate candidate is the scheme followed in *eSpeak* which, by design, should accurately describe the acoustic reality. However, since it is "functionally overloaded" it is not ideal in terms of the second criterion and furthermore may conflict with positional features that may be more suited to capturing certain (more predictable) phonetic regularities from the data. Nevertheless, to test this scheme, the algorithm was employed to add primary and secondary stress labels to the syllabified dictionary described in [8] and evaluated below (see the system labelled `ESPEAK` in Table II).[2]

Another proposal is to start with the concrete work of Mouton [13] (which largely coincides with the description in [12]) and extend the scope thereof to compound words by annotating the same existing dictionary. For this work the following procedure was followed: (1) The original rule-set was implemented as directly as possible and applied appropriately to the dictionary words (e.g. words were pre-processed with a simple compound splitter). And (2) the output was manually reviewed for any processing errors and cases outside of the scope of the design of the original rule-set (e.g. phrasal verbs, compounds and recent English influences) according to the protocol presented below. To simplify the manual review process and adhere to the above criteria for a useful stress feature, the following simple protocol was formulated (with the idea that systematic defects can be adressed in future work):

- Unstressed (`0`), primary (`1`) and secondary (`2`) stress levels are defined.
- All monosyllabic words are marked as unstressed.
- All polysyllabic words have at least one primary stressed syllable.

[1] http://espeak.sourceforge.net/

[2] This version of the dictionary is available here (commit: `3a14591`): https://github.com/NWU-MuST/za_lex/blob/master/data/afr/pronundict.txt

TABLE II
OBJECTIVE EVALUATION OVER UTTERANCES IN THE LWAZI 2 CORPUS

| Measure | Pronunciation resource | | | | | | |
|---|---|---|---|---|---|---|---|
| | RCRL | ESPEAK | NEW0 | NEW1 | NEW2 | NEW1rs | NEW1rts |
| $MCD$ (dB) | 4.0745 (0.2720) | 4.0614 (0.2699) | 4.0638 (0.2695) | **4.0614 (0.2690)** | **4.0595 (0.2688)** | 4.0739 (0.2668) | 4.0656 (0.2701) |
| $\Delta T$ (%) | 5.8750 (4.6695) | 5.9913 (4.8847) | **5.8099 (4.6964)** | **5.7514 (4.7473)** | 5.9327 (4.7688) | 5.8729 (4.7222) | 5.8232 (4.7093) |
| $RMSE_t$ (s) | 0.0837 (0.0404) | 0.0835 (0.0416) | **0.0833 (0.0412)** | **0.0822 (0.0409)** | 0.0843 (0.0424) | 0.0836 (0.0414) | 0.0833 (0.0419) |
| $RMSE_{f0}$ (st) | 2.5833 (0.5904) | **2.5476 (0.5643)** | 2.5813 (0.5820) | **2.5515 (0.5826)** | 2.5661 (0.5877) | 2.6287 (0.5740) | 2.5786 (0.5608) |
| $Corr_{f0}$ | 0.6846 (0.1524) | **0.6943 (0.1409)** | 0.6872 (0.1456) | **0.6939 (0.1467)** | 0.6928 (0.1425) | 0.6750 (0.1477) | 0.6883 (0.1388) |
| $RMSE_i$ (dB) | 6.1679 (0.7759) | 6.1317 (0.7818) | **6.1196 (0.7747)** | 6.1361 (0.7735) | **6.1264 (0.7796)** | 6.1619 (0.7717) | 6.1317 (0.7715) |

- More than one stressed syllable is only marked in compound words (although not all compounds necessarily have more than one stressed syllable).
- Stressed syllables are conservatively assigned; words are rarely assigned more than two stressed syllables.
- Words with fewer than 3 syllables only have a single stressed syllable and, more generally, the upper bound on number of stressed syllables (in short words) is assumed to be half the total number of syllables. Thus a four-syllable word will never be assigned more than two stressed syllables.
- Complex stress patterns in compound words are currently not considered, the resulting simplification is that the first stressed syllable is always marked as primary and all subsequent stressed syllables as secondary.

The resulting dictionary is referred to as NEW2 in Table II.[3]

Finally, as a baseline, implicit acoustic modelling of stress using simple positional features is currently the most prudent approach given the uncertainties expressed in Section I. Two resources should be considered: Firstly, the *Resources for Closely Related Languages Afrikaans Pronunciation Dictionary* (RCRL) initially developed for automatic speech recognition [14] with rule-based syllabification (adapted from [15], as described in [8]). And secondly, the work-in-progress dictionary without stress labels (NEW0).

## IV. OBJECTIVE EVALUATION

To evaluate the proposals in Section III, TTS systems were constructed based on the different resources and objective measurements were obtained by comparing synthesised utterances with actual speech.

For the TTS systems the *HTS toolkit* version $2.3_{beta}$ and associated demonstration scripts[4] [16], [17], [18] in combination with the *HTS engine* version 1.09[5], modified to perform mixed-excitation synthesis [19], was used to train multi-space probability distribution Hidden Markov Models (HMMs). The acoustic model contexts and tree tying questions were similar to those used in the HTS demonstration scripts (as in [5]), with the only exceptions being that no "accented" word or ToBI features were used and in the case of the "gpos" feature a simplified label-set marks only content and function word distinctions.

The measurements taken by aligning synthesised (reference) and actual utterances using dynamic time warping (DTW) are intended to include both segmental (spectral envelope) and prosodic (tempo, pitch and loudness) aspects of speech (features were extracted using the *Edinburgh Speech Tools* and *Praat* software packages [20], [21]):

- *Mel-cepstral distortion* ($MCD$): This represents the pronunciation clarity by measuring the mean distance between frames representing the spectral envelope [22]. The euclidean distance measure is used and the units are decibels relative to 1 (dB).
- *Relative absolute duration difference* ($\Delta T$): Is the absolute duration difference over a specific unit as a percentage of the reference duration.
- *Temporal root mean squared error* ($RMSE_t$): Is the RMS of the instantaneous temporal difference derived using the DTW alignment (measured in seconds).
- *Fundamental frequency (F0) RMSE* ($RMSE_{f0}$): Measures the error over the aligned pitch contours (in semitones (st) relative to 1 Hz).
- *Pearson correlation of F0* ($Corr_{f0}$): Measures the Pearson correlation (linear similarity in pitch movement) over the aligned pitch contours (a unitless value ranging from 0.0 to 1.0).
- *Intensity RMSE* ($RMSE_i$): Measures the error over aligned intensity (loudness) contours in decibels.

### A. Evaluating protocols on Lwazi 2

Table II presents a summary of the objective results for the systems built with different resources on the Lwazi 2 corpus. The values reported are calculated and averaged over all utterances using ten-fold cross-validation with standard deviations in parentheses. Acoustic models were constructed and applied on phone alignments with automatic pause insertion to reduce the effect of phrase breaks on the measurements. The RCRL, ESPEAK and NEW2 resources are as described in Section III, with all the dictionaries being equivalent in terms of segmental and syllable descriptions, except RCRL, and the following derivations included for this experiment:

- NEW0 is the new dictionary without stress labels.
- NEW1 is the NEW2 dictionary where all secondary stress labels are changed to primary.

[3]This version of the dictionary is available here (commit: 852cfa5): https://github.com/NWU-MuST/za_lex/blob/master/data/afr/pronundict.txt

[4]http://hts.sp.nitech.ac.jp/

[5]http://hts-engine.sourceforge.net/

TABLE III
OBJECTIVE EVALUATION ON ADDITIONAL CORPORA

| Measure | Pronunciation resource | | |
|---|---|---|---|
| | RCRL | NEW0 | NEW1 |
| Lwazi 3 corpus | | | |
| $MCD$ (dB) | 4.5007 (0.3813) | 4.4996 (0.3828) | **4.4957 (0.3819)** |
| $\Delta T$ (%) | **6.0074 (5.9185)** | 6.0314 (5.9960) | 6.0197 (5.9350) |
| $RMSE_t$ (s) | 0.1042 (0.0819) | 0.1041 (0.0805) | **0.1034 (0.0809)** |
| $RMSE_{f0}$ (st) | 3.8044 (1.7913) | 3.8182 (1.8281) | **3.7783 (1.7847)** |
| $Corr_{f0}$ | 0.5750 (0.2643) | 0.5759 (0.2602) | **0.5781 (0.2584)** |
| $RMSE_i$ (dB) | 7.6637 (1.2022) | 7.6758 (1.1911) | **7.6487 (1.1991)** |
| Multi-speaker corpus | | | |
| $MCD$ (dB) | 5.0910 (0.2972) | 5.0974 (0.2941) | **5.0877 (0.2974)** |
| $\Delta T$ (%) | **8.6001 (7.3877)** | 8.8687 (7.4196) | 8.8432 (7.5016) |
| $RMSE_t$ (s) | **0.1100 (0.0724)** | 0.1138 (0.0757) | 0.1123 (0.0743) |
| $RMSE_{f0}$ (st) | 2.9643 (0.9955) | 3.0643 (0.9978) | **2.9627 (1.0096)** |
| $Corr_{f0}$ | 0.5645 (0.2265) | 0.5386 (0.2293) | **0.5648 (0.2312)** |
| $RMSE_i$ (dB) | 8.2384 (1.2940) | 8.2348 (1.3206) | **8.2059 (1.3325)** |

- NEW1rs and NEW1rts are versions of NEW1 where the position of stress features were randomised. In the former, acoustic modelling proceeded with NEW1 and stress was randomised before synthesis and in the latter randomisation was applied before acoustic modelling.

In the table, the best and worst two measurements are marked using bold and red colour respectively to highlight the general trends. Some interesting observations are:

- The MCD values are close for all instances of the new dictionary; RCRL generally has the highest value and NEW2 the lowest (interestingly).
- Over the prosodic measures NEW1 is the most consistent with ESPEAK comparable (or slightly better) on F0.
- NEW0 and NEW2 were best-performing on the intensity measure.
- Comparing NEW1rs with RCRL and NEW0, it is expected that training with an explicit stress feature and randomising the stress placement during synthesis would give an approximation of the worst-case performance (on prosodic measures). This is confirmed especially on the F0 measures where this system has the worst performance and both RCRL and NEW0 clearly does better, however, on the intensity and duration measures these systems do not appear to perform significantly better.
- A comparison of NEW1rts with NEW1 shows that most measures, except intensity, are positively affected by the accuracy of the annotation (compared to possible improvement associated with an increase in the degrees of freedom).

### B. Evaluation on additional corpora

The above experiment was partially repeated on two additional corpora to determine whether the trends can be expected to hold more generally. Subsets of the single-speaker Lwazi 3 corpus [23] and a recently developed multi-speaker TTS corpus were used [7] (these subsets exclude foreign language parts – see Table I). The results using RCRL, NEW0 and NEW1 display a similar trend, with the possible exception of the temporal measure – see Table III. Table V further presents the results over three different classes of words: monosyllabic (involving no primary stress labels), words containing only a single stressed syllable, and compound words containing more than one stressed syllable. The monosyllabic set contains a large number of "function words" which are known to have possibly more varying segmental and prosodic realisations [24] which seems to be reflected in the variability of these measurements. The singly and multiply stressed word categories are of particular interest in this work. In the case of words with a single stressed syllable a consistent improvement can be seen over all three corpora. However, for compound words the results are again more variable (admittedly measured over very few samples).

## V. SUBJECTIVE EVALUATION

Since the objective evaluation suggests that the synthesised speech better approximates the original utterances in the corpus when using the NEW1 dictionary, a perceptual experiment was set up to determine whether the results are perceptually significant. For this experiment, two TTS voices were built as before using the Lwazi 2 corpus and RCRL and NEW1 dictionaries. The Lwazi 2 corpus was selected since the recording of small corpora is typical in TTS development for under-resourced languages (see for example [6] and [7]). Unseen sentences (35 in total) were randomly selected from three sources: a few news articles[6] (16 sentences), the Universal Declaration of Human Rights[7] (9 sentences), and Wikipedia[8] (10 sentences). Listeners were presented with synthetic speech pairs and simply asked to select the best sample or "no preference". Participants were free to listen to the speech samples using either headphones or loudspeakers and were asked to provide an indication of their level of experience with speech technology. A total of 20 listeners participated of which 8 were considered "speech experts" and 12 "non-experts". The normalised results for these two partitions and the combined set can be seen in Figure 1. According to McNemar's test statistic using the chi-squared ($\chi^2$) distribution with 1 degree of freedom and Yates' continuity correction, the 95% confidence level is given by $\frac{(|b-c|-0.5)^2}{b+c} \geq 3.841$. The original counts and test statistics are given in Table IV, showing a significant preference for the system based on NEW1.

## VI. CONCLUSION

In this work an evaluation of acoustic modelling of lexical stress was presented, with the following distinct contributions and outcomes:

[6]Articles were manually accessed during September 2016 from: http://www.netwerk24.com/

[7]From: http://www.unicode.org/udhr/d/udhr_afr.html

[8]Sentences were taken from the Afrikaans corpus in [7].

TABLE IV
GENERAL PREFERENCE TEST RESULTS

| | RCRL | NEW1 | No pref. | Total | $\chi^2$ |
|---|---|---|---|---|---|
| combined | 165 | 261 | 274 | 700 | 21.409 |
| non-experts | 95 | 144 | 181 | 420 | 9.842 |
| experts | 70 | 117 | 93 | 280 | 11.563 |

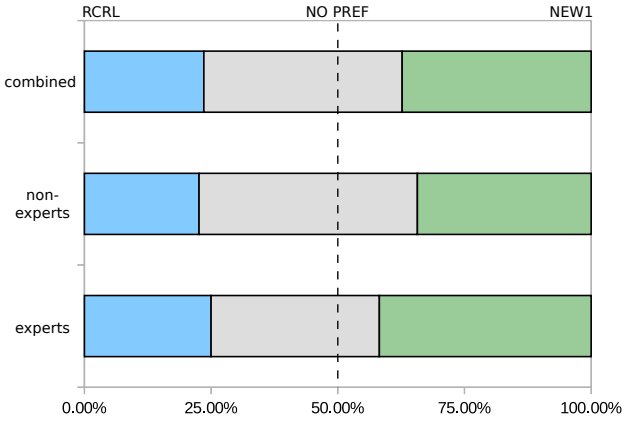


Fig. 1. General preference test results

- The annotation of lexical stress based on an existing pronunciation dictionary and existing descriptions of syllable stress in Afrikaans was motivated and proposed.
- Alternative annotation protocols were objectively evaluated using a combination of segmental and prosodic measures in the context of acoustic modelling for TTS using the Lwazi 2 corpus. The results showed that the newly proposed annotation using a single stress level was most consistent in this context.
- The objective evaluation was repeated on additional corpora and further analysed on the word level. The results confirm that the trends observed on the Lwazi 2 corpus are also evident in these cases which leads to the expectation that the annotated dictionary will also be useful in other contexts.
- A subjective evaluation in the form of a perceptual preference test confirmed that a system built with the proposed annotation is favoured over the baseline RCRL dictionary.

This work has focussed on investigating the impact of the proposed explicit lexical stress annotation on acoustic modelling for TTS. Consequently, the following practical aspects or questions were neglected and should be considered in future work:

- The prediction of lexical stress for out-of-vocabulary (OOV) words was not considered and the impact of prediction error on this task was not considered in the measurements (objective or subjective); all words trained on or synthesised were contained in the dictionary or pronunciation addenda.
- The addition of a lexical stress feature was partly motivated by the potential to explicitly specify it as input in different contexts, the extent to which this can be done in practice with perceptually significant result needs to be investigated.
- A more detailed analysis of the objective measurements may be done to better understand the results in Table V and some indications in Table II that having a secondary stress level may be useful. For these reasons the current annotation should be considered a work-in-progress.
- Given a successful explicit lexical stress feature, it may now be possible to consider further work on prosodic modelling, in the first instance to improve the synthesis of compound words perhaps by generalising patterns from simplex words or by using more powerful modelling techniques (results in Table V may indicate that more work can be done in these sparse contexts), and secondly on the implementation of higher-level prosody (e.g. prosodic prominence) which may depend on the lexical stress pattern [9].

Lastly, the positive perceptual results obtained here on a small speech corpus suggests that more accurate descriptions of lexical pronunciation of other under-resourced languages, especially in South Africa [11], may also be worthwhile.

## VII. ACKNOWLEDGEMENT

This work was partially funded by the Department of Arts and Culture (DAC) of the Government of South Africa. The authors are grateful to Ulrike Janke who assisted with the perceptual evaluation and to all participants involved in listening to speech samples, especially from the National Language Service at the DAC and the Human Language Technologies research group at the Meraka Institute, CSIR.

## REFERENCES

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[2] Z. H. Ling, S. Y. Kang, H. Zen, A. Senior, M. Schuster, X. J. Qian, H. M. Meng, and L. Deng, "Deep Learning for Acoustic Modeling in Parametric Speech Generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, May 2015.

[3] K. Tokuda and H. Zen, "Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 4215–4219.

[4] W. D. Basson and M. H. Davel, "Comparing grapheme-based and phoneme-based speech recognition for Afrikaans," in *Proceedings of the 23rd Annual Symposium of the Pattern Recognition Association of South Africa*, Pretoria, South Africa, 2012, pp. 144–148.

[5] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proceedings of the IEEE Workshop on Speech Synthesis*, 2002, pp. 227–230.

[6] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proceedings of the 5th ISCA Workshop on Speech Synthesis (SSW)*, Pittsburgh, USA, 2004, pp. 223–224.

[7] D. R. van Niekerk, C. van Heerden, M. Davel, N. Kleynhans, O. Kjartansson, M. Jansche, and L. Ha, "Rapid development of TTS corpora for four South African languages," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden, Aug. 2017, pp. 2178–2182.

[8] D. R. van Niekerk, "Syllabification for Afrikaans speech synthesis," in *Proceedings of the Twenty-Seventh Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Stellenbosch, South Africa, Dec. 2016, pp. 31–36.

TABLE V
OBJECTIVE RESULTS OVER WORDS

| Measure | Monosyllabic words (no stress) | | | Single stressed syllable words | | | Multiple stressed syllable words | | |
|---|---|---|---|---|---|---|---|---|---|
| | RCRL | NEW0 | NEW1 | RCRL | NEW0 | NEW1 | RCRL | NEW0 | NEW1 |
| Lwazi 2 (number of words: monosyl:5645; single:3428; multi:144) | | | | | | | | | |
| $MCD$ (dB) | 4.1615 (0.7101) | **4.1533 (0.7011)** | 4.1603 (0.7018) | 4.0729 (0.5159) | 4.0624 (0.5092) | **4.0544 (0.5051)** | 3.9908 (0.3990) | 3.9698 (0.3736) | **3.9531 (0.3801)** |
| $\Delta T$ (%) | **22.1481 (20.9026)** | 22.2801 (20.0662) | 22.3143 (19.7164) | 12.9180 (11.0859) | 12.8612 (11.1666) | **12.6180 (11.1206)** | **8.7367 (5.9213)** | 8.8046 (6.3080) | 8.7535 (6.0449) |
| $RMSE_t$ (s) | 0.0137 (0.0118) | **0.0136 (0.0118)** | 0.0136 (0.0120) | 0.0246 (0.0154) | 0.0241 (0.0148) | **0.0238 (0.0152)** | **0.0388 (0.0196)** | 0.0392 (0.0241) | 0.0388 (0.0217) |
| $RMSE_{f0}$ (st) | **1.2364 (0.9651)** | 1.2512 (0.9773) | 1.2541 (0.9861) | 1.9038 (0.9487) | 1.8952 (0.9346) | **1.8661 (0.9103)** | **2.2519 (0.7699)** | 2.2995 (0.8046) | 2.2685 (0.8255) |
| $Corr_{f0}$ | **0.4125 (0.5549)** | 0.4104 (0.5576) | 0.4085 (0.5583) | 0.4306 (0.4167) | 0.4346 (0.4151) | **0.4567 (0.4041)** | **0.5256 (0.3284)** | 0.4709 (0.3533) | 0.5152 (0.3409) |
| $RMSE_i$ (dB) | 5.3217 (2.3982) | **5.2782 (2.4022)** | 5.3156 (2.4223) | 5.5057 (1.5465) | 5.4676 (1.5121) | **5.4670 (1.5261)** | 5.6457 (1.2245) | **5.4944 (1.0744)** | 5.5289 (1.1242) |
| Lwazi 3 (number of words: monosyl:33201; single:16605; multi:344) | | | | | | | | | |
| $MCD$ (dB) | 4.5975 (0.7029) | **4.5970 (0.7010)** | 4.6006 (0.7026) | 4.4675 (0.5168) | 4.4636 (0.5205) | **4.4579 (0.5167)** | 4.4205 (0.3592) | 4.4057 (0.3521) | **4.3971 (0.3382)** |
| $\Delta T$ (%) | **20.7678 (19.3766)** | 20.9043 (19.2314) | 20.9550 (19.1857) | 11.8860 (10.2223) | 11.9153 (10.2660) | **11.8170 (10.1467)** | 8.0872 (6.1395) | 8.2939 (6.1356) | **8.0642 (6.6172)** |
| $RMSE_t$ (s) | 0.0111 (0.0106) | **0.0111 (0.0105)** | 0.0111 (0.0108) | 0.0197 (0.0147) | 0.0198 (0.0146) | **0.0193 (0.0143)** | 0.0316 (0.0176) | 0.0317 (0.0171) | **0.0310 (0.0178)** |
| $RMSE_{f0}$ (st) | 1.3596 (1.2624) | 1.3613 (1.2724) | **1.3444 (1.2533)** | 2.0867 (1.2933) | 2.0813 (1.3298) | **2.0448 (1.2811)** | **2.3098 (1.0897)** | 2.3149 (1.1082) | 2.3517 (1.1209) |
| $Corr_{f0}$ | 0.4559 (0.5810) | 0.4567 (0.5828) | **0.4614 (0.5797)** | 0.5104 (0.4422) | 0.5139 (0.4360) | **0.5329 (0.4374)** | **0.5808 (0.3338)** | 0.5767 (0.3283) | 0.5747 (0.3304) |
| $RMSE_i$ (dB) | 5.6332 (2.8540) | 5.6278 (2.8541) | **5.6254 (2.8444)** | 5.7336 (1.9850) | 5.7401 (1.9959) | **5.7195 (1.9739)** | **5.7136 (1.4769)** | 5.7884 (1.4600) | 5.7203 (1.4783) |
| Multi-speaker (number of words: monosyl:9108; single:6082; multi:147) | | | | | | | | | |
| $MCD$ (dB) | 5.1578 (0.6378) | 5.1655 (0.6435) | **5.1550 (0.6411)** | 5.1172 (0.4575) | 5.1270 (0.4540) | **5.1137 (0.4521)** | 5.1053 (0.3456) | 5.1116 (0.3431) | **5.1002 (0.3546)** |
| $\Delta T$ (%) | **21.9976 (18.8346)** | 22.4151 (18.8058) | 22.2220 (18.7592) | 13.8326 (11.3141) | 13.9736 (11.4212) | **13.6715 (11.0467)** | 10.2603 (6.8234) | 10.1670 (8.0963) | **9.9815 (8.0482)** |
| $RMSE_t$ (s) | 0.0146 (0.0144) | 0.0149 (0.0136) | **0.0146 (0.0136)** | **0.0252 (0.0159)** | 0.0258 (0.0163) | 0.0253 (0.0161) | **0.0392 (0.0195)** | 0.0423 (0.0473) | 0.0421 (0.0469) |
| $RMSE_{f0}$ (st) | **1.4508 (1.1939)** | 1.5036 (1.2126) | 1.4567 (1.1934) | 2.1140 (1.1542) | 2.2121 (1.1674) | **2.0960 (1.1516)** | **2.6085 (1.0591)** | 2.7427 (1.1223) | 2.6705 (1.0946) |
| $Corr_{f0}$ | **0.3751 (0.5711)** | 0.3506 (0.5698) | 0.3729 (0.5728) | 0.3540 (0.4497) | 0.3191 (0.4492) | **0.3727 (0.4459)** | **0.3913 (0.3262)** | 0.3389 (0.3823) | 0.3539 (0.3646) |
| $RMSE_i$ (dB) | 6.7824 (3.0692) | 6.7780 (3.1035) | **6.7027 (3.0071)** | 6.8926 (2.1826) | 6.8712 (2.1755) | **6.8650 (2.2005)** | 7.7348 (1.7447) | 7.5435 (1.6040) | **7.5295 (1.5903)** |


[9] Y. Xu and C. X. Xu, "Phonetic realization of focus in English declarative intonation," *Journal of Phonetics*, vol. 33, no. 2, pp. 159–197, 2005.

[10] J. C. Roux and A. S. Visagie, "Data-driven Approach to Rapid Prototyping Xhosa Speech Synthesis," in *Proceedings of the Sixth ISCA Workshop on Speech Synthesis (SSW)*, Bonn, Germany, Aug. 2007, pp. 143–147.

[11] K. Calteaux, F. de Wet, C. Moors, D. R. van Niekerk, B. McAlister, A. Sharma Grover, T. Reid, M. Davel, E. Barnard, and C. van Heerden, "Lwazi II Final Report: Increasing the impact of speech technologies in South Africa," Council for Scientific and Industrial Research, Pretoria, South Africa, Tech. Rep. 12045, February 2013.

[12] B. C. Donaldson, *A Grammar of Afrikaans*. Walter de Gruyter, 1993.

[13] E. W. Mouton, "Reëlgebaseerde klemtoontoekenning in 'n grafeem-na-foneemstelsel vir Afrikaans," Master's thesis, North-West University, Potchefstroom, South Africa, 2010.

[14] M. H. Davel and F. De Wet, "Verifying pronunciation dictionaries using conflict analysis," in *Proc. Interspeech*, Makuhari, Japan, Sept. 2010, pp. 1898–1901.

[15] T. A. Hall, "English syllabification as the interaction of markedness constraints," *Studia Linguistica*, vol. 60, no. 1, pp. 1–33, 2006.

[16] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modelling for speech recognition," *Journal of the Acoustic Society of Japan*, vol. 21, no. 2, pp. 79–86, 2000.

[17] T. Toda and K. Tokuda, "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[18] H. Zen, K. Oura, T. Nose, J. Yamagishi, S. Sako, T. Toda, T. Masuko, A. W. Black, and K. Tokuda, "Recent development of the HMM-based speech synthesis system (HTS)," in *Proceedings of th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Sapporo, Japan, 2009, pp. 121–130.

[19] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proceedings of EUROSPEECH*, Aalborg, Denmark, 2001, pp. 2263–2266.

[20] P. Taylor, R. Caley, A. W. Black, and S. King, *Edinburgh speech tools library*, http://www.cstr.ed.ac.uk/projects/speech_tools/, 1999.

[21] P. Boersma, *Praat, a system for doing phonetics by computer*. Amsterdam: Glott International, 2001.

[22] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer Voice Quality of New Languages Calibrated with Mean Mel Cepstral Distortion," in *The First International Workshop on Spoken Language Technologies for Under-resourced languages*, Hanoi, Vietnam, 2008.

[23] N. Titmus, G. I. Schlünz, J. A. Louw, A. Moodley, T. Reid, and K. Calteaux, "Lwazi III Project Final Report: Operational Deployment of Indigenous Text-to-Speech Systems," Meraka Institute, Council for Scientific and Industrial Research, Pretoria, South Africa, Tech. Rep., 2016.

[24] D. Jurafsky, A. Bell, E. Fosler-Lussiery, C. Girand, and W. Raymond, "Reduction of English function words in Switchboard," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 7, Sydney, Australia, 1998, pp. 3111–3114.