

Part-of-Speech Effects on Text-to-Speech Synthesis

Georg I. Schlünz*, Etienne Barnard† and Gerhard B. van Huyssteen‡

*Human Language Technology Competency Area
CSIR Meraka Institute, Pretoria, South Africa
gschlunz@csir.co.za

†Multilingual Speech Technologies Group
North-West University, Vanderbijlpark, South Africa
etienne.barnard@nwu.ac.za

‡Centre for Text Technology (CTeX)
North-West University, Potchefstroom, South Africa
gerhard.vanhuyssteen@nwu.ac.za

Abstract—One of the goals of text-to-speech (TTS) systems is to produce natural-sounding synthesised speech. Towards this end various natural language processing (NLP) tasks are performed to model the prosodic aspects of the TTS voice. One of the fundamental NLP tasks being used is the part-of-speech (POS) tagging of the words in the text. This paper investigates the effects of POS information on the naturalness of a hidden Markov model (HMM) based TTS voice when additional resources are not available to aid in the modelling of prosody. It is found that, when a minimal feature set is used for the HMM context labels, the addition of POS tags does improve the naturalness of the voice. However, the same effect can be accomplished by including segmental counting and positional information instead of the POS tags.

I. INTRODUCTION

The development of text-to-speech (TTS) voices for resource-scarce languages (RSLs) remains a challenge today. RSLs suffer from the problem of little available electronic data, such as texts and recorded speech, and linguistic expertise, such as phonological and morphosyntactic knowledge. The Lwazi project in South Africa [1] is a large-scale endeavour to gather and apply such human language technology resources for all eleven of the official South African languages. One of the current TTS goals of the Lwazi project is to produce more natural voices.

The naturalness of a TTS voice is primarily determined by prosody [2][3]. Prosody includes phrase breaks, sentence-level stress and intonation [4], and possibly word-level stress or tone as well. *Central to the modelling of most of the above effects stands part-of-speech (POS) tagging.* To elaborate:

- Word-level stress is dependent on the POS of the word, for example, in English, nouns often carry stress on different syllables than verbs [5]. This is true for word-level tone as well (which, in addition, requires a morphological analysis for finer grained information, such as tense, on top of the basic POS category [6]).
- Sentence-level stress requires a syntactic structure [4] of which POS information is a building block. Even a simple content-function word rule requires the POS of a word to categorise it.

- Phrase breaks can either be predicted from chunking [4], which in turn requires POS tagging, or directly from the POS tags themselves in a hidden Markov model (HMM) approach to modelling the junctures [7].
- Aspects of intonation, such as the sentence-final pitch of questions, may benefit from identifying, for example, question (“WH”-) words in English through POS tagging.

Solving the POS problem is, therefore, a prudent first step towards meeting the goal of natural TTS voices. But, in the light of the scarceness of resources, the question arises whether it is perhaps possible to circumvent the traditional approaches to prosodic modelling by learning the latter directly from the speech data using POS information. In other words, does the addition of POS features to the context labels of an HMM-based synthesiser improve the naturalness of a TTS voice?

This is the question we aim to answer in this paper. HMM-based voices are trained from English and Afrikaans prosodically rich speech. The voices are compared with and without POS features incorporated into the HMM context labels, analytically and perceptually. For the analytical experiments, measures of prosody to quantify the comparisons are explored. It is then also noted whether the results of the perceptual experiments correlate with their analytical counterparts.

The rest of the paper is structured as follows: in Section II the related work of POS tagging and TTS synthesis is discussed. In Section III the experimental setup and results are recorded. Finally, Section IV draws some conclusions about the results.

II. RELATED WORK

A. Part-of-Speech Tagging

A POS tag is a linguistic category assigned to a word in a sentence based upon its morphological and syntactic—or morphosyntactic—behaviour. Words are grouped into POS categories according to the affixes they take (morphological properties) and/or according to their relationship with neighbouring words (syntactic properties) [8]. Example POS categories common to many languages are *noun*, *verb*, *adjective* and *adverb*. Words are often ambiguous in their POS

categories. The ambiguity is normally resolved by looking at the context of the word in the sentence.

POS tagging is the automatic assignment and disambiguation of POS categories to words in electronic text. It is a prominent topic in NLP which has been well investigated, since it is a fundamental first step to subsequent syntactic, semantic and other NLP procedures, in applications such as TTS, information retrieval and grammar checking.

Approaches to automatic tagging include rule-based and statistical ones. The former use either hand-crafted rules [9][10], which require intricate linguistic knowledge, or rules learned from data [11][12]. The latter are data-driven and use statistical methods, such as Markov models [13] or maximum entropy models [14], to determine the lexical probability (for example, without context, *address* is more likely to be a noun than a verb) and contextual probability (for example, after *to*, *address* is more likely to be a verb). Both approaches to POS tagging are, therefore, very resource-intensive tasks (either in terms of human resources or data resources), and it is a prominent engineering problem in NLP to optimise the use of such resources.

B. Text-to-Speech Synthesis

TTS is the generation of speech signals from text. It comprises the following stages (adapted from [4]):

- 1) *Text segmentation* splits the character stream of the text into more manageable units, namely sentences and tokens (the written forms of the unique words yet to be discovered). The processes are called *sentencisation* and *tokenisation*, respectively.
- 2) *Text decoding* decodes each token into one or more uniquely pronounceable words. Non-standard word tokens such as numbers, dates and abbreviations are classified and expanded into their standard word natural language counterparts in a process called *normalisation*. A special case of *homograph disambiguation* then disambiguates homographs among the token expansions that are not homophones.
- 3) *Text parsing* infers additional lexical, syntactic and morphological structures from the words which are useful for the pronunciation and prosodic modelling stages to follow. The tasks include *POS tagging* (see Section II-A), *chunking* (parsing of non-overlapping phrases) and *morphological analysis* (identification of stems and affixes in words).
- 4) *Pronunciation modelling* models the pronunciation of individual words. It maps the words to their constituent phonemes, either by looking up known words in a *lexicon* or by applying *grapheme-to-phoneme (G2P) rules* to unknown words. *Syllabification* divides the words into syllables. *Word-level stress* or *tone*, depending on the language type, is then assigned to the syllables.
- 5) *Prosodic modelling* predicts the prosody of the whole sentence, namely the *phrasing* (pauses between phrases), *sentence-level stress* (a phenomenon of connected speech: certain words in a phrase are stressed according

to their word-level stress, at the expense of reducing the word-level stress of the other words) and *intonation* (the melody or tune of an entire sentence).

- 6) *Speech synthesis* encodes the above information into speech waveforms. *Hidden Markov Model-based synthesis* is a statistical parametric technique which uses the source-filter paradigm to model the speech acoustics: the source models the glottal waveform (a pulse train for voiced sounds and random noise for unvoiced sounds) and the filter models the formant resonances of the vocal tract. *Excitation* (inter alia fundamental frequency or *F0*), *spectrum* and *duration* parameters are estimated from recorded speech and modelled by context-dependent HMMs. The contexts considered are phonetic, linguistic and prosodic. The excitation and spectrum parameters are used in the excitation generation and synthesis filter module to synthesise the speech waveform [15].

TTS voice quality is deemed acceptable according to two performance criteria: *intelligibility* and *naturalness*. Intelligibility measures how understandable the speech is to a listener, that is to which degree the listener will be able to recount the original words in the text. Typical methods employed to evaluate intelligibility include comprehension and transcription tests. Naturalness measures how much the TTS voice sounds like the voice of a human. Methods of evaluation include perceptual tests where a listener rates a single utterance or compares two utterances relatively.

A multilingual TTS system, called Speect [16], has been developed for the Lwazi project. In the first phase of the project, Speect incorporated the following modules in its natural language processing (NLP) front-end:

- Whitespace-based tokenisation,
- G2P rules,
- Syllabification and
- Punctuation-based phrase break insertion.

The digital signal processing (DSP) back-end was a unit-selection synthesiser. For each language, a small speech corpus was recorded with neutral prosody. The neutral prosody compensated for the few examples that would be present per unique type in the unit-selection database. Using just these few resources, baseline intelligible voices for all the languages could be synthesised [1].

Towards producing more natural voices, Speect is now exploring the HTS engine [17][18] as HMM-based synthesiser for more and, hopefully, better control over the voices (the parameterisation allows for manipulation). Furthermore, the speech corpora are going to be larger (albeit still very small compared to those of majority languages) and prosodically richer.

III. EXPERIMENTS

A. Common Setup

This section describes various aspects common to all the experiments that follow.

1) *Taggers*: The fast, accurate HMM tagger HunPos [19] is incorporated into Speect to obtain the POS information at synthesis time. For English, the tagger is trained on a 40,000-token subset of the Penn Treebank WSJ corpus [20] using a reduced set of 18 tags. The tagger obtains an accuracy of 95.90% on a separate 10,000-token test set from the same corpus. For Afrikaans, 40,000 tokens of the balanced corpus developed in [21] are used as training data. The tagset is reduced to 17 tags. The tagger is 94.64% accurate on the test set of 10,000 tokens, also from the corpus. The reason for the small training corpora is to emulate the impact of a resource-scarce environment on POS tagging accuracy.

2) *Voices*: The data requirements for voice building are recorded speech segmented into utterances (spoken sentences), and transcriptions thereof. For the English voice, data from the CMU_ARCTIC speech synthesis databases (available at http://festvox.org/cmu_arctic/) is used: the speaker is “US bdl”, a United States English-speaking male. The data consist of 1,132 utterances. For the Afrikaans voice, speech data recorded in-house for the Lwazi project is used; the speaker is female. The number of utterances is 1,005. For each voice 100 random utterances from the data are selected for testing; the remainder is used for training. These corpora are also quite small because the recording process is very cumbersome in terms of obtaining enough utterances at a sufficient quality.

An HTS voice is built in a two-stage process: phonetic alignment and HMM training. The phonetic alignment is performed by the Speect NLP front-end and an additional tool (as part of a toolkit released with Speect) that uses forced-alignment based on HTK [22]. The NLP front-end maps the transcriptions to phoneme sequences. The alignment tool allows model initialisation from manually aligned speech data transcribed in a different language or phoneset by mapping to broad phonetic categories. This is highly beneficial as a bootstrap for the alignment of a small corpus of an RSL [23]. The TIMIT corpus [24] is used as bootstrapping data.

The HMM training from the aligned utterances is performed by the demonstration script released with the HTS engine. It uses HTS version 2.1 [17] to build the models for the hts_engine API version 1.02 [18]. The script is only slightly altered to accommodate Speect as a front-end.

The question file of the demonstration script is mostly used as is for the model tying decision tree. Only the POS-related questions are modified to reflect the tagsets chosen in Section III-A1 and, for the Afrikaans voice, the phonetic category questions are altered to reflect the Afrikaans phoneset.

The HTS context labels utilise by default a set of linguistic features defined in [25] and included in the demonstration package. The full context labels comprise, inter alia, features based on the identities of the current and neighbouring phonemes, the number and relative positions of phonemes, syllables, words and phrases, whether syllables are stressed or not, and the POS of words.

The voices are built using two versions of the context labels: one with maximum features and the other with minimum features. The maximum feature set comprises all of the features

provided by the demonstration package, including the segmental counting and positional features, but excluding the stress and intonation-related ones (so that prosody is not modelled explicitly but only implicitly by the POS information). The minimum features only include the phonemic identities. For the experiments, the two versions are then used with and without POS information.

3) *Analytical Test*: In order to measure the prosodic effects of POS information on synthesised speech, it is necessary to understand what the physical manifestations of prosody are. *Duration*, *pitch* (F0) and *intensity* have been shown to be acoustic correlates of prosody [2][26]. These three measures are used to determine the closeness of each synthesised utterance to its natural speech counterpart from the 100-utterance test set:

a) *Duration*: The natural speech utterance is phonetically aligned to determine the duration of each phoneme. For each phoneme, the corresponding duration of the synthesised utterance is subtracted and the absolute value is taken to represent the distance between the natural phoneme and the synthesised phoneme.

b) *Pitch*: The F0 contour of the natural speech utterance is extracted with Praat [27] and divided according to the aligned durations so that each phoneme is assigned its corresponding section of F0 values. For the synthesised utterance, the HTS engine is forced to use the same duration alignments and output the synthesised F0 values. The distance between the natural and the synthesised phoneme is taken as the Mean Squared Error (MSE) of the synthesised F0 values $\hat{\mathbf{f}}_0$ to the natural ones \mathbf{f}_0 :

$$MSE(\hat{\mathbf{f}}_0) = E \left[(\hat{\mathbf{f}}_0 - \mathbf{f}_0)^2 \right] \quad (1)$$

where

$$\begin{aligned} E[\mathbf{x}] &= \sum_i p_i x_i \\ &= \frac{1}{n} \sum_i x_i \quad \text{if } p_i = \frac{1}{n} \end{aligned} \quad (2)$$

Any differences involving undefined F0 values are taken as zeros in the summations.

c) *Intensity*: The intensity contour is extracted in similar fashion to the F0 contour. Praat is used for both the natural speech and synthesised utterances, the latter once again being aligned on the phoneme boundaries of the former. The distance between the phoneme sections of intensity values is also the MSE.

When comparing two TTS voices, for example with and without POS information, the one voice is deemed more natural than the other for a particular utterance if its synthesised version is closer to the natural speech counterpart than the synthesised version of the other voice.

An experiment is thus compiled from the test set by synthesising the 100 test sentences with both voices. Each utterance pair is scored by counting the number of phonemes that are closer to the natural speech for a particular measure. The

utterance with the highest number of phonemes wins, and the corresponding voice is accredited with that test sentence for the measure. The voice accredited with the most test sentences in the end is then more natural. The evaluation takes place on the sentence level, because prosodic effects normally range across several words (as noted in the exposition on prosodic modelling in Sections I and II-B). Finally, McNemar’s test is used to examine the statistical significance of the result.

McNemar’s test is a chi-square test for paired sample data [28]. It is calculated as follows:

$$\chi^2 = \frac{(|B - C| - 0.5)^2}{B + C} \quad (3)$$

χ^2 has a chi-squared distribution with one degree of freedom (if $B + C$ is large enough). To test for significance, χ^2 is compared to the appropriate chi-square table value. A result of probability greater or equal to 0.05 is generally considered to be significant. Lining up this boundary probability with one degree of freedom in the table gives a value of 3.841.

Let n_1 be the number of utterances accredited to the first voice and n_2 to the second voice. McNemar’s test can then be applied by setting $B = n_1$ and $C = n_2$. If $\chi^2 \geq 3.841$ the winning voice is significantly more natural than the other voice. If $\chi^2 < 3.841$ the result is insignificant and the two voices can be said to be similar in their degree of naturalness.

4) *Perceptual Test*: A perceptual test is also set up in an effort to validate the analytical results. Respondents are asked to listen to pairs of utterances from the test set. In a pair the two utterances, A and B , are synthesised from the same sentence, but each by a different TTS voice. The respondents must then choose which utterance out of the pair sounds more natural relatively, or if both sound the same (without listening to the original natural speech). Duration, pitch and intensity are not distinguished; only an aggregate judgement is required.

This “ A versus B ” approach (with McNemar’s test for significance) is preferred above a mean opinion score (with the Wilcoxon signed rank test [29] for significance) that is used, for example, in the Blizzard Challenge [30]. The reason is that it is more robust against respondent subjectivity: different respondents are more likely to judge the same utterance out of a pair as more natural than assign it the same score on a scale of 1 to 5.

McNemar’s test can be used in reverse to calculate how many pairs will be needed to obtain a significant result. Recall that

$$\frac{(|B - C| - 0.5)^2}{B + C} \geq 3.841 \quad (4)$$

is required for statistical significance. This may be rewritten in terms of the total number of pairs N :

$$\frac{(xN - 0.5)^2}{yN} \geq 3.841 \quad x, y \in [0, 1] \quad (5)$$

where $xN = |B - C|$, the ratio of N estimated to be equal to the difference between the discordant pairs. For fixed y , the smaller this difference (or x) is, the bigger N must be for significance. $yN = B + C$, the ratio of N estimated to be

equal to the sum of the discordant pairs, in other words, the number of pairs not judged equal in naturalness. For fixed x , the smaller this sum (or y) is, the smaller N needs to be for significance. A change in x varies N to a greater degree than a change in y does.

For conservative estimates of $x = 0.2$ and $y = 0.8$, $N \geq 82$. Nevertheless, a large safety margin is built into the test by setting $N = 200$. The 200 pairs are divided up among 10 respondents so that each respondent must listen to 20 pairs. The 20 sentences that make up the pairs are randomly selected from the 100-utterance test set, such that every two respondents listen to a unique subset.

For the two languages, the 20 respondents are mother-tongue speakers with an average age of between 30 and 35. Out of the 10 English respondents, 6 are male and 4 female. 7 Afrikaans respondents are male and 3 female. A website facilitates the playback of the audio samples and recording of answers.

B. Experiment 1: POS Effects Using Maximum Features

The first experiment compares the naturalness of two TTS voices of which the HTS context labels use the maximum features. The English voices are dubbed **eng_maxlab_nopos** for the version without POS information, and **eng_maxlab_pos40k** for the version with POS information. Similarly, the Afrikaans voices are named **afr_maxlab_nopos** and **afr_maxlab_pos40k**. The aim is to observe the effect of the POS information in an already feature-rich environment for maximum benefit.

Table I shows the results for English and Afrikaans. The first column lists the measures and the second column the total number of utterances evaluated. Columns 3 and 4 list the number of utterances accredited to each voice and column 5 the number found equal. The last column lists the McNemar χ^2 -scores for significance (which are independent of the equal counts).

TABLE I
NATURALNESS RESULTS WHEN USING MAXIMUM FEATURES

Measure	Utterances				χ^2
	Total	eng_maxlab_nopos	eng_maxlab_pos40k	Equal	
Duration	100	46	49	5	0.066
Pitch	100	52	42	6	0.960
Intensity	100	41	52	7	1.185
Perception	200	72	83	45	0.711
Measure	Utterances				χ^2
	Total	afr_maxlab_nopos	afr_maxlab_pos40k	Equal	
Duration	100	48	47	5	0.003
Pitch	100	48	45	7	0.067
Intensity	100	49	43	8	0.329
Perception	200	72	74	54	0.015

The analytical figures across the two languages show no significant bias towards a particular voice, and the perceptual figures confirm this (there are basically as many votes for the one voice as for the other). Therefore, it may be deduced that the two voices are similar in their degree of naturalness

and that the POS information has no effect when using the maximum features. It may be that the POS effects are “drowned out” by the other features, which inherently carry similar information beneficial towards naturalness.

Finally, it is observed that the equal count among the perceptual figures is proportionally much higher than among the analytical figures. The simple explanation is that it is much more difficult for respondents to hear a distinction between two utterances than what it is to calculate the difference between two discrete values.

C. Experiment 2: POS Effects Using Minimum Features

Since the use of the maximum features is suspected to suppress the POS effects, it is prudent to retest the TTS voices with reduced features in order to lift out the effects. The English voice without POS information is **eng_minlab_nopos** and with **eng_minlab_pos40k**. The corresponding Afrikaans voices are **afr_minlab_nopos** and **afr_minlab_pos40k**. The results are shown in Table II.

TABLE II
NATURALNESS RESULTS WHEN USING MINIMUM FEATURES

Measure	Utterances				χ^2
	Total	eng_minlab_nopos	eng_minlab_pos40k	Equal	
Duration	100	51	44	5	0.445
Pitch	100	26	66	8	16.959
Intensity	100	46	46	8	0.003
Perception	200	76	101	23	3.391

Measure	Utterances				χ^2
	Total	afr_minlab_nopos	afr_minlab_pos40k	Equal	
Duration	100	51	40	9	1.212
Pitch	100	15	79	6	42.896
Intensity	100	35	55	10	4.225
Perception	200	71	95	34	3.327

For both languages, pitch dominates the results by clearly favouring the voices with POS information as more natural. The near significant perceptual figures tend to suggest the same. Of note is the disparate results for intensity: it manifests as a deciding factor only for Afrikaans. It is unclear from this experiment as to what the cause might be; see trends observed in the rest of the experiments.

D. Experiment 3: POS Effects Using a Less Accurate Tagger

From the previous experiment it was seen that, when only minimum features are available, adding POS information improves the pitch component of naturalness. Within the context of an RSL, the question now arises whether the same effect is possible when a less accurate POS tagger, trained on fewer resources, is used. The third experiment thus compares two TTS voices, both with POS information on top of the minimum features, but where the tagger of the one voice has been trained on only 5,000 tokens. For English, the 5,000-token tagger is 90.95% accurate and the corresponding voice is called **eng_minlab_pos05k**. The voice of the normal 40,000-token, 95.90% accurate tagger is called **eng_minlab_pos40k**. For Afrikaans, the tagger trained on 5,000 tokens is 87.95%

accurate and its voice is dubbed **afr_minlab_pos05k**. The voice of the 40,000-token, 94.64% accurate tagger used so far is dubbed **afr_minlab_pos40k**. Table III shows the naturalness results for this experiment.

TABLE III
NATURALNESS RESULTS WHEN USING A LESS ACCURATE TAGGER

Measure	Utterances				χ^2
	Total	eng_minlab_pos05k	eng_minlab_pos40k	Equal	
Duration	100	48	46	6	0.024
Pitch	100	45	49	6	0.130
Intensity	100	39	55	6	2.556
Perception	200	69	94	37	3.683

Measure	Utterances				χ^2
	Total	afr_minlab_pos05k	afr_minlab_pos40k	Equal	
Duration	100	46	46	8	0.003
Pitch	100	48	46	6	0.024
Intensity	100	44	53	3	0.745
Perception	200	57	89	54	6.796

Pitch, the prominent measure in the previous experiment, does not feature here for any of the two languages, nor do any of the other analytical measures. These insignificant differences between the voices with the more and less accurate tagger may support the hypothesis that one can achieve the same prosodic effects with the less accurate tagger (read fewer resources). However, the perceptual figures do show a bias towards the voices using the more accurate tagger, near significantly for English and significantly for Afrikaans. This mismatch between the analytical and perceptual results renders the experiment inconclusive.

E. Experiment 4: Comparing Minimum and Maximum Features

The final experiment revisits the implication of the first, namely that the maximum features might compensate for the effect of adding POS information. The minimum feature voices with POS information, **eng_minlab_pos40k** for English and **afr_minlab_pos40k** for Afrikaans, are set against the maximum feature voices without POS information, **eng_maxlab_nopos** for English and **afr_maxlab_nopos** for Afrikaans. The results are shown in Table IV.

TABLE IV
RESULTS OF THE COMPARISON BETWEEN MINIMUM AND MAXIMUM FEATURES

Measure	Utterances				χ^2
	Total	eng_minlab_pos40k	eng_maxlab_nopos	Equal	
Duration	100	35	59	6	5.875
Pitch	100	44	52	4	0.586
Intensity	100	46	46	8	0.003
Perception	200	69	104	27	6.880

Measure	Utterances				χ^2
	Total	afr_minlab_pos40k	afr_maxlab_nopos	Equal	
Duration	100	42	48	10	0.336
Pitch	100	35	61	4	6.773
Intensity	100	61	34	5	7.392
Perception	200	64	93	43	5.174

The duration and pitch figures of the two languages suggest that the maximum feature voices without POS information are more natural (duration and pitch being significant for each language in turn). The perceptual test results favour these voices significantly as well. Therefore, the extra counting and positional features in the maximum feature set not only compensate for the POS information (they would then have insignificant differences as a result), they improve the naturalness beyond what the POS tags can affect.

The Afrikaans intensity figures contradict this conclusion, but, in the light of the same behaviour observed in Section III-C, it might be a systematic anomaly of **afr_minlab_pos40k**. It can possibly be ascribed to speaker variability (choice) in the speech corpora of the two languages. During the reinspection of the Afrikaans data, alignment errors as a result of unexpected pauses, transcription errors, mispronunciations and G2P errors were also found. Further investigation is required before any conclusions can be drawn.

IV. CONCLUSION

It has been shown that POS information does contribute to the naturalness (specifically in terms of pitch) of a TTS voice when it forms part of a small phoneme identity-based feature set in the HTS labels. However, the same effect, even an improvement, can be accomplished by including segmental counting and positional information instead of the POS tags in the HTS labels—and no extra resources are used. Therefore, it is not necessary to incur the cost of POS tagging when the traditional route of prosodic modelling cannot be followed in the development of a TTS voice. The experiments were limited though to the Germanic languages of English and Afrikaans. It would be prudent to test the effects on the other South African languages, especially the tone-driven Bantu languages.

Notwithstanding the above results, it is problematic that the correlation between the analytical and perceptual methods is not yet clear-cut. This is because the analytical measures did not always behave in a consistent way across the two languages and the four experiments. The problem can be addressed from both sides: either the perceptual tests should be more fine-grained (that is duration, pitch and intensity must be judged separately), or a new analytical framework can be used where, for example, the three measures are combined into a single one. The former is very difficult to achieve since the human ear cannot discern such differences well. The latter is possible by constructing a classifier such as the Gaussian discriminative function presented in [26]. In either case, it warrants a much more thorough study of the acoustic and perceptual factors of prosody.

REFERENCES

[1] The South African Department of Arts and Culture, “Lwazi – a telephone-based, speech-driven information system;” <http://www.meraka.org.za/lwazi/>.

[2] M. Dong, K. Lua, and J. Xu, “Selecting prosody parameters for unit selection based Chinese TTS,” in *Natural Language Processing – IJCNLP 2004*, ser. Lecture Notes in Computer Science, K. Su, J. Tsujii, J. Lee, and O. Y. Kwong, Eds. Springer Berlin / Heidelberg, 2005, vol. 3248, pp. 272–279.

[3] J. Romportl, “Structural data-driven prosody model for TTS synthesis,” in *Proceedings of the Speech Prosody 2006 Conference*, 2006, pp. 549–552.

[4] P. Taylor, *Text-to-Speech Synthesis*, 1st ed. Cambridge University Press, 2009.

[5] P. Roach, *English phonetics and phonology: a practical course*, 4th ed. Cambridge University Press, 2009.

[6] S. Zerbian and E. Barnard, “Word-level prosody in Sotho-Tswana,” *Speech Prosody 2010*, vol. 100861, pp. 1–4, 2010.

[7] P. Taylor and A. W. Black, “Assigning phrase breaks from part-of-speech sequences,” *Computer Speech and Language*, vol. 12, pp. 99–117, 1998.

[8] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd ed. Pearson Education, 2009.

[9] S. Klein and R. Simmons, “A grammatical approach to grammatical coding of English words,” *JACM* 10, pp. 334–347, 1963.

[10] B. Greene and G. M. Rubin, “Automatic grammatical tagging of English,” Providence RI: Department of Linguistics, Brown University, 1971.

[11] D. Hindle, “Acquiring disambiguation rules from text,” in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 1989.

[12] E. Brill, “A simple rule-based part of speech tagger,” in *Proceedings of the third conference on Applied natural language processing*. Association for Computational Linguistics, 1992, pp. 152–155.

[13] T. Brants, “TnT – a statistical part-of-speech tagger,” in *Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000)*, 2000.

[14] A. Ratnaparkhi, “A maximum entropy model for part-of-speech tagging,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-96)*, Philadelphia, PA, 1996.

[15] A. W. Black, H. Zen, and K. Tokuda, “Statistical parametric speech synthesis,” in *Proceedings of ICASSP 2007*, 2007.

[16] J. A. Louw, “Speect: A multilingual text-to-speech system,” in *Proceedings of the 19th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, 2008, pp. 165–168.

[17] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, “The HMM-based speech synthesis system version 2.0,” in *Proceedings of ISCA SSW6*, 2007, pp. 294–299.

[18] K. Tokuda, S. Sako, H. Zen, K. Oura, K. Nakamura, and K. Saino, “The hts_engine API,” <http://hts-engine.sourceforge.net/>.

[19] P. Halácsy, A. Kornai, and C. Oravecz, “HunPos - an open source trigram tagger,” in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, 2007, pp. 209–212.

[20] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, “Building a large annotated corpus of English: The Penn Treebank,” *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1994.

[21] S. Pilon, “Outomatiese Afrikaanse woordsoortetkettering,” Master’s thesis, North-West University, 2005.

[22] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, *The HTK Book (for HTK Version 3.3)*. Cambridge University Engineering Department, 2005.

[23] D. R. van Niekerk and E. Barnard, “Phonetic alignment for speech synthesis in under-resourced languages,” in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK, 2009, pp. 880–883.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *Darpa Timit: Acoustic-phonetic Continuous Speech Corps CD-ROM*. US Dept. of Commerce, National Institute of Standards and Technology, 1993.

[25] K. Tokuda, H. Zen, and A. W. Black, “An HMM-based speech synthesis system applied to English,” in *Proceedings of the 2002 IEEE Speech Synthesis Workshop*, 2002.

[26] A. Waibel, *Prosody and Speech Recognition*, 1st ed. London: Pitman Publishing, 1988.

[27] P. Boersma and D. Weenink, “Praat: doing phonetics by computer,” <http://www.praat.org/>.

[28] S. Boslaugh and P. A. Watters, *Statistics in a nutshell*, 1st ed. O’Reilly Media, Inc., 2008.

[29] R. Lowry, “The Wilcoxon signed-rank test,” <http://faculty.vassar.edu/lowry/ch12a.html>.

[30] S. King and V. Karaiskos, “The blizzard challenge 2010,” in *Proceedings of the Blizzard Challenge 2010 Workshop*, Kansai Science City, Japan, September 2010.