Developing a Corpus to Verify the Performance of a Tone Labelling Algorithm

Mpho Raborife
Human Language Technologies Group
Meraka Institute (CSIR)
Pretoria, South Africa
Telephone: +27 (0)12 8413028
Email: mraborife@csir.co.za

Sabine Zerbian
Department of Linguistics
Potsdam University
Potsdam, Germany
Telephone: +49 (0)331 9772051
mail: sabine.zerbian@uni-potsdam.de

Sigrid Ewert
School of Computer Science
University of the Witwatersrand
Johannesburg, South Africa
Telephone: +27 (0)11 7176180
Email: sigrid.ewert@wits.ac.za

***Abstract*—We report on a study that involved the development of a corpus used to verify the performance of two tone labelling algorithms, with one algorithm being an improvement on the other. These algorithms were developed for speech synthesis purposes with the aim of improving the perceived naturalness as well as the intelligibility of the speech produced by the synthesizer. The corpus used to test the algorithms consisted of 45 Sesotho sentences specifically chosen to represent the contexts that are necessary for the application of the algorithms. These sentences were recorded and transcribed by three independent transcribers each having experience in tonal studies. We employed statistical methods to prove the reliability of our transcriptions, thus allowing us to use the transcriptions to evaluate the algorithms.**

I. Introduction

Bantu languages are tonal languages in that they use pitch to distinguish between meaning as opposed to stress-accent languages such as English. According to [1], pitch is considered a correlate for intonation and tone, both in studies on tone languages such as the southern Bantu languages and in studies on stress-accent languages such as English. Furthermore, in acoustic studies available on Bantu tone, pitch variations have been taken as the sole indicator for tone [1].

Extensive research has been dedicated to tone modeling for East-Asian languages such as Chinese [2]–[5]. Recently, studies have focused on tone modeling for African tone languages such as Yoruba [6]–[8]. For Bantu languages, we only know of one published attempt to model tone, namely for an isiZulu text-to-speech (TTS) system [9]. Not having tone implemented into TTS systems for tonal languages compromises their perceived naturalness as well as their intelligibility. This illustrates the need for research on tone modeling for this group of languages.

To model tone in TTS systems developed for Yoruba, the tone labels on the syllables are assigned appropriate pitch values using methods such as fuzzy logic based rules and decision trees. In Yoruba, tone labels are marked in orthographic writing. In contrast, Southern Bantu languages do not mark tone in orthography, therefore it is difficult to assign appropriate pitch values to syllables as it is not known which of the syllables are high or low-toned.

It is clear that a solution which allows for the equal treatment and advancement of Southern Bantu languages, with respect to other tonal languages in terms of speech systems developed for such languages, is needed. According to [9], such a solution could be an algorithm that predicts which of the syllables of a word are to be marked for a high or a low tone and the appropriate intonation of the word. The authors argue that with such an algorithm, a TTS system would produce more "natural" tones.

The study by [10] focused on how such an algorithm could be implemented using Sesotho as a basis. It was found that such an algorithm would need to use linguistically-defined tonal rules to make the tone mark (label) predictions. That is, the tonal rules are described in the literature of the language in question and the algorithm implements them.

The study by [11] aimed to improve an algorithm developed by [10] that uses three linguistically-defined Sesotho tonal rules to make tone label predictions on the syllables of Sesotho words. The algorithm predicts the tone labels of a syllable based on the underlying tones, as described in a Sesotho dictionary by [12] which marks tone, as well as the tense, mood and aspect of the verb stems. This algorithm restricts its applications to polysyllabic verb stems.

This algorithm is improved by implementing four other linguistically-defined Sesotho tonal rules. Furthermore, its application is extended to other word classes. We will refer to the algorithm developed by [10] as the basic algorithm and its improved version as the extended algorithm.

This article discusses the compilation of a corpus which was used to assess the performance of these algorithms. Since the output produced by both algorithms is compared to recorded speech, we discuss the recording process as well as the transcription process.

The paper is structured as follows: Section II provides the tonal system of Sesotho. Section III provides a detailed description of how the corpus was compiled. Section IV discusses the recording process and we present the method we employed to choose the speaker to be used for our transcriptions in Section V. Section VI presents the transcription process employed on the corpus in order to prepare it for the evaluation of the algorithms. Section VII presents a summary of the research study as well as the major points presented in this paper.

II. Background

Sesotho is a tonal language; it uses pitch variations to convey grammatical or lexical meaning. According to [13], the exact pitch of a tone in a word varies from speaker to speaker; it is the relation of the pitch of one syllable to that of the next that is constant with all speakers. Sesotho has only two tonal levels, namely the high tone (H) and the low tone (L)[1].

Tone in Sesotho may be used to differentiate meaning on a lexical level as shown in Example (1). In this example, the word *bona* has different tonal patterns. It is this difference in tonal patterns that affects the meaning of the word.

(1) bóna - "see"
boná - "they"

Tone in Sesotho may also be used to show grammatical relationships. For instance, first person singular subject markers can only be distinguished from similar third person forms by means of tone [13]. The following example from [13, p.39] illustrates this grammatical relationship:

(2) Ké motho - "It is a person"
Ke motho - "I am a person"

It is also important to note that nasal sounds in Sesotho (*m, n, ng* and *ny*) form syllables on their own when they are not immediately followed by a vowel such as the first *n* in *monna ("man")*. In that instance the nasal sounds can carry tone (*mo*ń*na*). Nasal sounds are sounds that are pronounced with breath escaping mainly through the nose.

Sesotho has a variety of dialects such as the Lesotho dialect and the Free State dialect. However, the tonal rules applied by the basic as well as the extended algorithms are not restricted to a specific Sesotho dialect.

III. Corpus Compilation

Our test data consists of 45 Sesotho sentences (with 565 tone bearing syllables) which were chosen from a large corpus of Sesotho data which includes folk-tales and poetry[2]. The sentences were chosen to include the following occurrences:

- The present principal tense.
- The past tense.
- The perfect tense.
- The imperative mood.
- The future tense.

These tenses are relevant for the application of our algorithm. Proper nouns and/or loan words are not included in the corpus since their tone patterns are not available in the literature [12], [14]. Verb forms showing the Potential Mood (using *-ka*) have been excluded since the tonology of these forms is subject to considerable dialectical variation [15]–[17]. Furthermore, there are no monosyllabic verb stems in our data given that they have complex tonal characteristics [18], [19].

Since we did not have pre-recorded Sesotho sentences for the transcriptions, the 45 sentences were recorded by three native Sesotho speakers. The transcriptions are needed to evaluate the algorithm. In the next section, we discuss the recording process as well as the protocol used for the recording session.

IV. Recording Process

A. Introduction

This phase involved formal introductions between the interviewer and the respondent. The respondent was given the research information sheet which explained the purpose of this research study. The respondent together with the interviewer read the sheet and the interviewer made sure that the respondent understood the contents of the document clearly. The respondent was supervised throughout the entire recording session by a native Sesotho speaker to check for mispronunciations and naturalness.

B. Confirmation of Consent

Once the respondent understood the recording process as well as the purpose of the research study, s/he made a final decision as to whether s/he was willing to participate in the recording session. The interviewer emphasized that participation was completely voluntary and that the data gathered in the recording would remain confidential, but could be used for further research.

The respondent was also informed that pseudonyms were used to store the data. Based on this information, if the respondent chose to continue with the recording session, s/he was given a formal consent form to sign.

The consenting respondent was also required to fill out a language profile questionnaire. This questionnaire allowed us to verify the eligibility of the respondent's participation in our study.

C. Recording Guidelines

In this phase of the research study, the interviewer explained the purpose of the recording and its contribution to the research study. The interviewer also explained and loaded the following instructions onto a computer screen.

- Silently read and make sure you understand the English meaning of the sentence given below the Sesotho sentence.
- Only say the Sesotho sentence, not its English meaning.
- Please speak in a natural voice and tone.
- You can record the sentence again if you are not happy with the first recording.
- Please press enter only after you have finished recording the sentence.

The respondent was then asked to read out the instructions above. The instructions were recorded and played back to the respondent to make sure that the microphone was placed at an appropriate angle. Once the respondent understood the

[1] á: the diacritic on the vowel *a* is used to indicate high tone, low tones are not marked.

[2] The Human Language Technology group at the Meraka Institute has a corpus which has Sesotho sentences taken from poetry and folk-tales as well as the Government Gazette. The sentences were selected from this corpus.

instructions, the interviewer proceeded to the final phase of the session as described in the following section.

D. Recording the Sentences

The research study involved recording native Sesotho speakers (respondents) who were not remunerated for the recordings. Each speaker recorded 45 sentences and eight tonal minimal pairs. Tonal minimal pairs are two words or phrases which differ only in tone and have a distinct meaning. We used them to test if the speaker can distinguish tone. The minimal pairs used in our study are as follows:

a) Minimal pairs in isolation:

- ho báká - "to repent"
  ho baka - "to cause"
- ho téná - "to put on"
  ho tena - "to disgust"

b) Minimal pair verb phrases:

- o já dijó - "you eat food"
  ó já dijó - "he eats food"
- le já dijó - "you (plural) eat food"
  lé já dijó - "he (class 5) eats food"
- bá a sébá - "they are doing mischief"
  bá a seba - "they are gossiping"

c) Short sentences differing in tone:

- Ke ngwaná wá háo - "I am your child"
  Ké ngwaná wá háo - "S/he is your child"
- O mobé - "You are ugly"
  Ó mobé - "'S/he is ugly"
- Ke motho - "I am a person"
  Ké motho - "It is a person"

These minimal pairs were used to select the respondent with the most pronounced tones as discussed in Section V. This session took approximately three and a half hours for all speakers.

The tonal minimal pairs were presented to the respondent before the 45 sentences. Tone labels were not marked on the minimal pairs, nor on the 45 Sesotho sentences. However, the English meaning was given beneath the original sentence/phrase in order to guide the respondent as to which tone pattern was to be pronounced for that phrase/word.

For the purposes of our research study, only one respondent was needed for transcriptions. In the next section, we provide a reason for using only one respondent as well as for how the respondent was chosen from the three recorded respondents.

V. Speaker Selection

To evaluate the tone labelling algorithm, we compared its output (the 45 selected sentences described in Section III, including the tone marks predicted by the algorithm) to the speech recorded by a native Sesotho speaker. This speech needed to be encoded in a way that allows for this comparison, that is, it needed to have tone marks (either high or low). Since our algorithm output the tone marks on the syllables, to label the tone marks on the recorded speech, we transcribed the speech to orthography which included the tone marks on the syllables.

It was not necessary to transcribe the recordings made by all three recorded respondents, because the tonal rules implemented by our algorithm are not restricted to speakers of a specific Sesotho dialect. We used only one respondent's recording for the transcriptions. The respondent we chose for transcriptions had to have produced the most "clearest" tones. By "clearest", we mean the most pronounced and thus better perceived tones for the transcribers (labelers). We needed clearly pronounced tones for the transcription process since the labelers used mostly perception in addition to acoustic means to transcribe the speech as discussed in Section VI.

To choose the speaker with the "clearest" tones in their speech from the three respondents, two processes of elimination were employed.

- The above mentioned eight tonal minimal pairs were used to determine the speaker's ability to distinguish tone.
- The clarity of the tones the speaker produced in their speech was verified perceptually and acoustically.

We named our respondents TK, MM and NM. In this process of elimination, respondent NM could not perceptually identify the tonal distinction in the following minimal pair:

Le já dijó - "you (plural) eat food"
Lé já dijó - "s/he (class 5) eats food"

This minimal pair shows the changing of tone affecting the subject marker *le*. In the first sentence, *le* refers to a 2^{nd} person plural whereas in the second sentence it refers to a 3^{rd} person singular. We thus excluded respondent NM's recording from the transcriptions.

Respondent TK produced clear tones and spoke in a natural voice compared to respondent MM, who spoke in an unnatural voice and exaggerated the tone and words. In some instances, speaker MM produced each word without taking the context into consideration. This is problematic since the tonal rules are restricted within certain phonological boundaries and some of the rules in specific morphological contexts. Also, a word in isolation might be tonally ambiguous and one would need the context to pronounce the appropriate tone pattern.

Respondent TK's recording was transcribed by three independent labelers, each having some experience in tonal studies. The transcriptions are in orthography and include tone labels. In the next section, we discuss how the labelers transcribed the speech by TK.

VI. Transcription of the Sentences

A. Transcription Process

The transcription of prosody is widely used in linguistic research as well as in the development and testing of speech systems [20]. It is used in understanding linguistic variables and their relation to the interpretation of human speech. In our study, three independent expert labelers transcribed tone as perceived from TK's speech. We used these transcriptions to evaluate our algorithm. The transcriptions represent spoken language in written form including tone labels.

Perceptual tests are the most widely used methods for evaluating TTS systems. In these tests, native speakers of the language in question are required to grade the synthesized speech with respect to naturalness. This is because synthesized speech produced by TTS systems ideally needs to sound as natural as human speech.

We used transcriptions to evaluate our algorithm since they represent Sesotho speech as pronounced by a native Sesotho speaker. Given that our algorithm is developed for the purpose of tone modeling for a Sesotho TTS system, it is crucial in our study that the tone labels predicted by our algorithm match those in the transcriptions. We did not employ any statistical methods to predict the tone labels as such methods are not based on perception and thus cannot give us an indication of how well our algorithm can fare on a perception test. However, we employed statistical methods to establish the reliability of the transcriptions as discussed in Section VI-B. The transcriptions are in orthography and include tone labels as exemplified below.

(3) Noha é fálla metsí - "The snake is emigrating from its waters"

Initially, we aimed to use 12 labelers to transcribe TK's recording. These labelers were to be divided into two groups, six being experts in tonal studies and the other six having no experience in tonal studies. In each of the two groups, three labelers were to be native Sesotho speakers and the other three were to be speakers of non-tonal languages. However, in the expert group, only three labelers were found, whereas in the non-expert group, six labelers were found. Table I summarizes the initial characteristics of the labelers.

Group	Native	Non-native
Expert	1	2
Non-expert	3	3

TABLE I
LABELER CHARACTERISTICS

We found that the labelers in the non-expert group had difficulty transcribing the tone labels. In this group, even native speakers had difficulty transcribing the labels. It is not straightforward to transcribe tone and one usually needs some experience with the transcription system. Also, native speakers of tonal languages are usually not aware of their language's tonal system as this is not taught at schools. Since there were more labelers in the non-expert group, they would have outweighed the expert labelers in the majority votes used to choose the tone labels of the syllables in cases of disagreements between the labelers as discussed later in this section. Thus, we chose to use the three expert labelers to transcribe the tone labels. Although it would have been preferable to have more expert transcribers, it was not easily achievable since there are not many experts in this area of study.

The transcribers have different backgrounds with respect to tonal studies, with two of them being linguists and the other a speech engineer. Furthermore, two of the transcribers are speakers of Germanic languages (linguist (A) and speech engineer (B)) and the other is a native Sesotho speaker (linguist (C)). Their background is as follows: The labelers who are non-native Sesotho speakers have developed a framework that introduces an algorithm which derives word-level tone assignments for the Sotho-Tswana languages [14]. It is this framework on which the study by [10] is based. The labeler who is a native Sesotho speaker has implemented this framework by developing the basic algorithm [10].

To evaluate the improvement made by our algorithm on the basic algorithm, we used the labels transcribed by our individual labelers. The labelers were not given a training manual, thus they relied on perception and acoustic means to label the tone marks. The individual transcriptions were compared across the three labelers, finding an agreement of 55.9% (316 out of 565 tone-bearing syllables). In the case of disagreements between the labelers, the final tone label was decided by a majority vote, that is, it was based on the tone label that two labelers agreed on.

The agreement between each pair of transcribers out of 565 tone-bearing syllables is shown in Table II.

Transcriber	Matched Syllables	Agreement Percentage
A and B	375	66.4%
A and C	374	66.5%
B and C	400	70.8%

TABLE II
AGREEMENT BETWEEN PAIRS OF TRANSCRIBERS

The labelers used Praat to label the tone marks on the syllables whose tones they could not audibly perceive. Praat is a speech analysis software that has a pitch tracking algorithm which estimates pitch from a speech waveform [21]. It produces a contour that estimates pitch, which the labelers observed when marking the tone labels. In perceptual analysis, the transcribers relied on hearing, interpreting and understanding the tone patterns in the speech. However, perception is subjective and varies from one transcriber to the other. This served as a disadvantage in our study as discussed in the following section.

According to [22], the two most widely used pitch tracking algorithms are Praat and Yin. The authors compared these algorithms in extracting pitch for isiZulu and found that Praat performs slightly better than Yin. According to [23], for the Nguni languages, Praat is more accurate and noise-resistant than Yin, which has no post-processing on the input speech. Thus, our transcribers used Praat for the acoustic analysis aspect on the test data. An example of an extracted pitch contour (in blue; if you have no colour, it is the faint line between 75Hz and 500Hz across the spectrum) by Praat is shown in Fig. 1.

Even though the Praat pitch tracking algorithm has been found to be accurate, it does not fare well with unvoiced

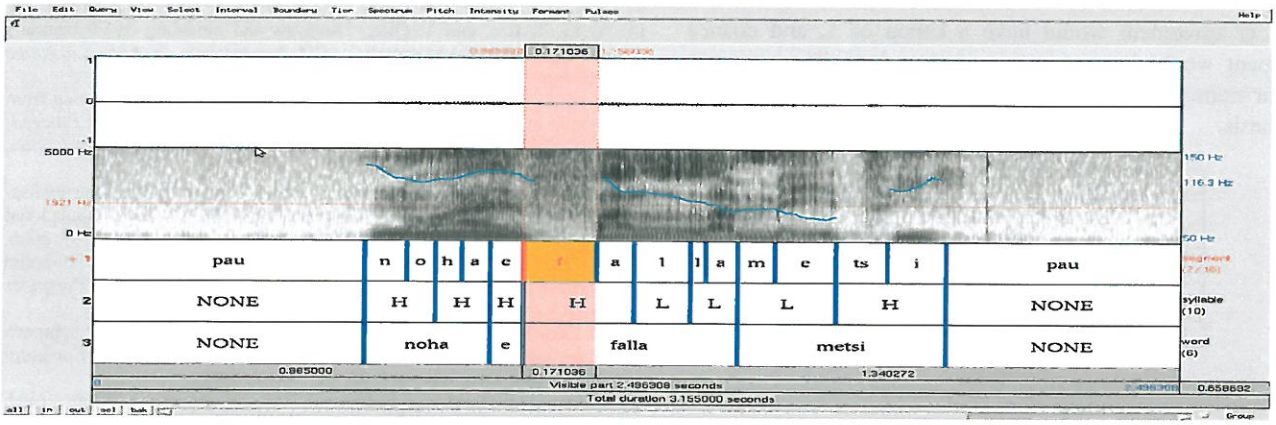

Fig. 1. Extracted signal from a Sesotho speech recording

sounds as shown in Fig. 1 (the unvoiced sound *f* is highlighted and has no visible pitch track) [24]. In speech we have voiced and unvoiced sounds. When producing voiced sounds the vocal folds vibrate and when unvoiced sounds are produced, the vocal folds are at rest. According to [24], unvoiced segments usually do not have a defined pitch contour, for example, the unvoiced sound *f*. Unvoiced segments were ignored when assigning tone labels to the syllables. A syllable always contains at least one voiced segment.

In the study by [10], the labelers reached unanimous agreement on the tone labels in 72.3% of the cases. The recordings used in [10] were for speech synthesis purposes, thus the speaker was asked to produce a rather monotonous speech melody. The higher agreement in that study can be attributed to this speech melody. The sentences chosen in our study were fairly complex and spoken in a natural, fast voice making it difficult to track pitch changes.

B. Inter-Transcriber Reliability

As discussed in Section VI-A, three labelers transcribed the 45 recorded sentences independently. They reached unanimous agreement on the tone labels in 55.9% of the cases. A reason for such a low agreement between the transcribers might lie in the fact that perception is subjective and that the tone transcribers did not have a training manual. Two of our labelers have non-tonal languages as mother tongue. Studies have shown that speakers of non-tonal languages are less sensitive to the perception of tone, but are not necessarily tone "deaf" [25], [26]. In those studies, the speakers of the non-tonal languages had no experience in tonal studies. In our study, all the labelers were sensitive to tone. Furthermore, they had different backgrounds and experiences regarding tone in Bantu languages.

Our study relied heavily on the transcriptions since they formed a central part in evaluating our algorithm. Thus, it was important to statistically verify the transcriptions since there was a fairly low agreement between our transcribers. To determine the reliability of the transcriptions, we used the kappa (κ) coefficient which is a statistical measure used for analyzing the reliability of agreement between a number of labelers (raters) during observer reliability studies [27]–[29]. This statistical measure was more reliable than a simple percentage agreement calculation since it took into account the agreement occurring by chance between the labelers.

Fleiss' kappa is a well-known multirater kappa which assumes that raters are forced to assign a certain number of cases to each category, hence, it is referred to as a fixed-marginal kappa [28]. For instance, in our study, it assumes that the labelers are restricted in the number of syllables (cases) they can assign to each of the tone labels (categories). However, this is not the case as the labelers are free to assign any number of the syllables to each of the two tone labels. We thus make use of a free-marginal kappa which assumes that raters are not restricted to assign a specific number of cases to each of the categories, as proposed by [28].

[29] defines the free-marginal kappa, κ, as follows:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

$\bar{P}$ is the observed percentage agreement and $\bar{P}_e$ is the percentage of agreement expected by chance alone.

According to [27], the factor $1 - \bar{P}_e$ gives us the measure of agreement that is attainable above what would be predicted by the labelers by chance and $\bar{P} - \bar{P}_e$ is the degree of agreement actually achieved above chance. If $\kappa = 1$, then there is complete agreement between the labelers. If $\kappa \leq 0$, then there is no agreement among the labelers other than what would be expected by chance, as if the raters had simply guessed every rating.

We used an online kappa calculator which allows the user to enter all the necessary variables and then calculates the κ-value [30]. The free-marginal κ-value between three labelers, assigning 565 syllables to two categories as calculated by [30], is as follows: $\kappa = 0.41$.

The positive value of kappa indicates that the agreement is better than what would have been expected by chance. To interpret this value we use Table III as described by [31]. According to the classification by [31], the labelers in our study seem to be in moderate agreement on the tone labels they have transcribed, since $\kappa = 0.41$. Recall that

a perfect agreement would have a kappa of 1, and chance agreement would have a kappa of 0 or less. Thus we can use our transcriptions to evaluate our algorithm and test our hypothesis.

κ	Interpretation
≤ 0	Poor agreement
$0.0 - 0.20$	Slight agreement
$0.21 - 0.40$	Fair agreement
$0.41 - 0.60$	Moderate agreement
$0.61 - 0.80$	Substantial agreement
$0.81 - 1.00$	Almost perfect agreement

TABLE III
INTERPRETATION OF κ-VALUES

VII. Conclusion

The development of a tone label prediction algorithm requires effective means to test its accuracy. This paper has described our approach to collecting data that allows us to test the accuracy of such algorithms. Our corpus was relatively small due to the manual preparation that was required, such as the tone label transcriptions. Thus, the efficient collection of a larger corpus remains a challenge for future work.

In our study, we used only one speaker for the transcription process. This does not affect our evaluation since the tonal rules implemented by the algorithms are not dialectical. Furthermore, we used minimal pairs to verify the tonal awareness of the speaker. The clarity of the tone produced by the speaker was also perceptually and acoustically verified. Thus, the speech used for the transcriptions was valid for the purposes of our study.

During our transcription process, we found that native Sesotho speakers have difficulty classifying tones. Thus, in future, more research should be dedicated to adopting acoustic measures such as the fundamental frequency to perform such a classification.

Acknowledgements

The authors would like to thank the reviewers for their valuable comments which led to the improvement of the paper.

This material is based upon work supported financially by the National research Foundation. Any opinion, findings and conclusion or recommendations expressed in this material are those of the authors and therefore the NRF does not accept liability in regard thereto.

References

[1] S. Zerbian and E. Barnard, "Phonetics of intonation in South African Bantu languages," *Southern African Linguistics and Applied Language Studies*, vol. 26, no. 2, pp. 235–254, 2008.

[2] R. Wang, Q. Liu, and D. Tang, "A new Chinese text-to-speech system with high naturalness," in *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, USA, 1996, pp. 1441–1444.

[3] T. Lee, G. Kochanski, C. Shih, and Y. Li, "Modeling tones in continuous Cantonese speech," in *Proceedings of the 7th International Conference on Spoken Language Processing*, Colorado, USA, 2002, pp. 2401–2404.

[4] Y. Li, T. Lee, and Y. Qian, "Analysis and modeling of F0 contours for Cantonese text-to-speech," *ACM Transactions on Asian Language Information Processing*, vol. 3, no. 3, pp. 169–180, 2004.

[5] J. Ni, S. Sakai, T. Shimizu, and S. Nakamura, "Prosody modeling from tone to intonation in Chinese using a functional F0 model," in *Proceedings of the 2nd International Symposium on Universal Communication*, Osaka, Japan, 2008, pp. 397–404.

[6] O. Odejobi, A. Beaumonta, and S. Wonga, "Intonation contour realisation for Standard Yoruba text-to-speech synthesis: a fuzzy computational approach," *Computer Speech and Language*, vol. 20, pp. 563–588, 2006.

[7] A. Dtunji, S. Wonga, and A. Beaumonta, "A fuzzy decision tree-based duration model for standard Yoruba text-to-speech synthesis," *Computer Speech and Language*, vol. 21, pp. 325–349, 2007.

[8] O. Odejobi, S. Wonga, and A. Beaumonta, "A modular holistic approach to prosody modelling for standard Yoruba speech synthesis," *Computer Speech and Language*, vol. 22, pp. 39–68, 2008.

[9] J. Louw, M. Davel, and E. Barnard, "A general-purpose isiZulu speech synthesiser," *South African Journal of African Languages*, vol. 25, no. 2, pp. 92–100, 2005.

[10] M. Raborife, "The implementation of Sesotho tonal rules in a text-to-speech system," Honours Research Report, School of Computer Science, University of the Witwatersrand, 2009.

[11] ——, "Tone labelling algorithm for Sesotho," Masters Dissertation, School of Computer Science, University of the Witwatersrand, 2011.

[12] J. Du Plessis, J. Gildenhuys, and J. Moiloa, *Tweetalige woordeboek Afrikaans-Suid-Sotho / Bukantswe ya maleme-pedi Sesotho-Seafrikanse*, 1st ed. Cape Town: Via Afrika Beperk, 1974.

[13] C. Doke and S. Mofokeng, *Textbook of Southern Sotho grammar*, 1st ed. London: Longmans Green and Co, 1957.

[14] S. Zerbian and E. Barnard, "Word-level prosody in Sotho-Tswana," in *Proceedings of Speech Prosody*, Chicago, USA, 2010.

[15] D. Creissels, A. Chebane, and H. Nkhwa, *Tonal morphology of the Setswana verb*, 1st ed. Lincom, Europe: Munich, 1997.

[16] D. Cole and D. Mokaila, *A course in Tswana*, 1st ed. Washington, USA: Georgetown University, 1962.

[17] D. Lombard, "Aspekte van toon in Noord-Sotho," PhD Thesis, University of South Africa, 1976.

[18] B. Khoali, "A Sesotho tonal grammar," PhD Thesis, University of Illinois, 1991.

[19] S. Zerbian, "Segmental and suprasegmental properties of monosyllables in Sotho/Tswana," in *International Conference "Monosyllables - from phonology to typology"*, University of Bremen, Germany, 2009, pp. 111–115.

[20] A. Syrdal and J. McGory, "Inter-transcriber reliability of ToBI prosodic labeling," in *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, 2000.

[21] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, pp. 341–345, 2001.

[22] N. Govender, E. Barnard, and M. Davel, "Developing intonation corpora for isiXhosa and isiZulu," in *Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa*, Cape Town, South Africa, 2005, pp. 147–151.

[23] N. Govender, "Intonation modelling for the Nguni languages," Master Thesis, University of Pretoria, 2006.

[24] N. Govender, E. Barnard, and M. Davel, "Pitch Modelling for the Nguni Languages," *South African Computer Journal*, vol. 38, pp. 28–39, 2007.

[25] P. Halle, Y. Chang, and C. Best, "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners," *Journal of Phonetics*, vol. 32, pp. 395–421, 2004.

[26] J. T. Gandour and R. A. Harshman, "Crosslanguage differences in tone perception: a multidimensional scaling investigation," *Language and Speech*, vol. 21, pp. 1–33, 1978.

[27] J. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 5, pp. 378–382, 1971.

[28] R. L. Brennan and D. J. Prediger, "Coefficient kappa: some uses, misuses, and alternatives," *Educational and Psychological Measurements*, vol. 41, no. 1, pp. 687–699, 1981.

[29] J. J. Randolph, "Free-marginal multirater kappa: an alternative to Fleiss' fixed-marginal multirater kappa," in *Joensuu University Learning and Instruction Symposium*, Joensuu, Finland, 2005.

[30] J. Randolph, "Online kappa calculator," 2008. [Online]. Available: http://justus.randolph.name/kappa

[31] J. Landis and G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.