Automatic Stylization, Coding and Modelling of Intonation in Text-to-Speech for Under-Resourced Languages

Johannes A Louw
Human Language Technologies Research Group
Meraka Institute, CSIR
Pretoria, South Africa
jalouw@csir.co.za

Avashlin Moodley
Human Language Technologies Research Group
Meraka Institute, CSIR
Pretoria, South Africa
amoodley1@csir.co.za

***Abstract*—In this paper an automatic method to implicitly model intonation for statistical parametric speech synthesis (SPSS) is presented. The approach is ideally suited to single speaker speech databases as used in text-to-speech (TTS), due to the models being speaker-specific. Fundamental frequency curves are automatically stylized based on the speaker-specific acoustics in the recorded database, requiring no models rooted in linguistic theory, and therefore being well suited to intonation modelling in under-resourced languages. The stylized curves are then coded into abstract pitch labels, which are used as features in the training of the statistical parametric acoustic models. A conditional random field (CRF) model is trained in order to predict the abstract pitch labels from the text for synthesis. The CRF model can be used to predict the abstract pitch labels on the syllable, word and phrase tiers. Objective and subjective results on synthetic voices built from English and isiXhosa speech databases are shown.**

***Index Terms*—speech synthesis, text-to-speech, prosody, intonation, automatic stylization**

I. Introduction

The goal of a text-to-speech (TTS) system is to synthesize natural sounding synthetic speech, with appropriate emotion, emphasis and prosody for the synthesized text. These speech qualities are broadly measured, subjectively and objectively, in terms of their naturalness and intelligibility in order to ascertain the performance of the TTS system.

Prosody contributes to the naturalness and intelligibility of speech by emphasizing groupings of semantic content through a combination of rhythm, stress and intonation on a suprasegmental level.

Intonation is defined in [1] as "the use of *suprasegmental* phonetic features to convey 'postlexical' or *sentence-level* pragmatic meanings in a *linguistically structured* way". The intonation function augments an utterance by adding emotion and attitude and structuring a sentence into phrases. Intonation is often used as a synonym for prosody, and therefore includes phrasing and prominence [2], but as used in this work it is concerned with the use of pitch for communication, or the *speech melody*. Intonation is both a physiological and cognitive function, it is physiological due to physical limitations (minimum and maximum frequency of the individual's vocal cords) and cognitive to add meaning.

The main acoustic realization of intonation is pitch, which from a signal point of view is measured (or calculated) as the fundamental frequency (f_0) curve of the speech signal. Although pitch and f_0 are sometimes used interchangeably, there is a difference: pitch is the frequency of the sound that is perceived, while f_0 can be defined as the frequency of the vocal cords producing the perceived sound [2]. The f_0 curve can be considered to be a combination of a *macroprosodic* component due to the conscious choice of the speaker to add meaning, and a *microprosodic* component due to the segmental units in the utterance being spoken [3].

The synthesized f_0 or intonation curve needs to be generated in some or other fashion with only the target utterance text available to the TTS system. Therefore, intonation models are *generally* derived from large corpora of text that have been labelled with discrete intonation symbols which can then be mapped to an f_0 curve (in some languages, for example Yorùbá [4], tones are marked explicitly on the orthography, thereby negating the need of deriving the discrete intonation symbols from the text). The labelling of the discrete intonation symbols and modelling a mapping thereof to an f_0 curve requires expert linguistic knowledge, which is rarely available for under-resourced languages.

The aim of this work is to improve the naturalness of the synthesized intonation by automatically stylizing the broader macroprosodic f_0 component, coding the stylized curves into discrete symbols, and modelling the coded symbols from a single speaker TTS corpus. The coded symbols are then used *implicitly* in the modelling of the intonation in statistical parametric speech synthesis (SPSS). No manually annotated labels are required in this proposed method, making it suitable to improve the naturalness of TTS systems for under-resourced languages. Another advantage of the proposed method is that the intonation model is specific to the speaker, and not a broader language specific model as is the general case.

The organization of the paper is as follows: Section II details previous work done on intonation modelling, whilst Section III

describes the proposed method. In Section IV the experiments and results of the proposed method are given, and lastly a discussion and conclusion is presented in Section V.

II. Background and Related Work

The goal of this work is to find an appropriate model for intonation that will fit into the SPSS framework for under-resourced languages. Such a model should have the following characteristics:

1) Language independent,
2) Automatic stylization, and
3) Abstract symbolic pitch codes.

The automatic stylization falls partly into the language independent category and is partly due to the single speaker databases (generally) used for building synthetic TTS voices. An f_0 stylization that is specific to the speaker should be able to better synthesize intonation for that specific speaker.

The need for the abstract symbolic pitch codes is to be able to model the acoustics of the speaker in the SPSS framework with discrete features for intonation.

There are many models, theories and explanations of intonation, from a linguistic and acoustic point of view. Some models do not even have any grounding in linguistic theory or biological plausibility [2] and are purely pragmatic engineering solutions. In the next sections we give a general classification of the different categories of intonation models as well as some examples of models that have been popular in TTS systems.

A. Classifications

Intonation models can be broadly classified into the following categories [5], [2]:

1) *Phonological vs. Phonetic*,
2) *Tones vs. Shapes*, and
3) *Single-layered vs. Superpositional*.

1) Phonological vs. Phonetic: A *phonological* intonation model is descriptive and discrete, using an inventory of abstract phonological categories representing linguistic functions [6].

Phonetic intonation models are motivated from acoustic (f_0) data.

The models attempt to describe f_0 movements and often link these back in some or other way to the linguistic level. According to [7], phonetic intonation is universal or language independent, while language specific meaning is given by intonational phonology.

2) Tones vs. Shapes: Intonation is modeled as either tone or pitch levels, or pitch shapes and dynamics. This corresponds to the "description" classification in [5].

3) Single-layered vs. Superpositional: In single-layered intonation models intonation events are modeled as a linear sequence while in superpositional models the events may be superposed on top of for example the phrase level intonation. In other words, in superpositional models intonation events can be modelled as a "modulation" of the phrase level or higher intonation. This corresponds to the "arangement" classification in [5].

B. Models

An example of a phonological intonation model is ToBI (**To**nes and **B**reak **I**ndices) [8], which has its roots in autosegmental-metrical phonology [9]. ToBI specifies an inventory of tones: one set is used to mark accented syllables, while another set is used to mark phrase boundaries. Each tone marks a different type of accent or boundary. ToBI is an example of a phonological, tone, single-layered model.

ToBI was developed for Standard American English, but there is a German version called GToBI [10]. AuToBI [11], which is a tool for the automatic analysis of ToBI tone labels for Standard American English, was developed due to the annotator reliability of ToBI labels being unsatisfactory. AuToBI is not suitable for under-resourced languages because it is language specific and adapting it would require linguistic resources which might not be available.

The Tilt [12] model, which is a phonetic, shapes, single-layered model, describes pitch accents and boundary tones via rising and falling quadratic functions that are derived from acoustic data. Straight-line interpolations are used for stretches of speech between intonational events. The Tilt model is language independent, but not well suited to SPSS because there is no intermediate symbolic pitch level which can be used as a feature for intonation to the acoustic models. Tilt is better suited to the modelling of the f_0 curve directly.

Another example of a phonetic model is the Fujisaki model [13], which is classified as a shapes and superpositional intonation model. It aims to create an accurate parametric description of the f_0 curve. The model is superpositional in that it has separate components for phrase and accent, where accent is modulated onto the phrase intonation. The Fujisaki model has the same problems as the Tilt model in fitting into the SPSS framework.

The MOMEL (**Mo**delling **mel**ody) and INTSINT (**Int**ernational **T**ranscription **S**ystem for **Int**onation) [14] model can be viewed as a hybrid phonetic/phonological model. MOMEL stylizes the f_0 curve with a series of quadratic splines, which after some processing, can then be assigned to discrete symbolic tone labels with the INTSINT algorithm.

MOMEL/INTSINT has been used previously [6] in the automatic modelling of intonation in TTS in a unit selection framework, but there is no clear anchor between the phonetically derived INTSINT abstract pitch labels and a phonological utterance tier. This introduces timing issues into the synthesized intonation, resulting in poor prediction of the abstract pitch labels.

The PENTA (**P**arallel **En**coding and **T**arget **A**pproximation) [15] model assigns an f_0 target to each syllable in an utterance based on given tone labels. The f_0 curve is generated by connecting targets with an interpolation function which can be realized in different forms. The PENTA model is a phonetic, shapes and single-layered model in the classification as given in Section II-A. The PENTA model has the same shortcomings in an SPSS framework as the Fujisaki and Tilt models due to it being a shapes type model.

III. Proposed Method

The method proposed in this work is broadly based on SLAM (Automatic **S**tylization and **La**belling of Speech **M**elody) [16]. While SLAM was developed for the automatic labelling of intonation for research into speech prosody in communication, this work builds on it in order to model intonation in an SPSS framework. In [17] a similar approach as this work was followed in that a simplified set of class labels of SLAM was used to model intonation in an SPSS framework, although the work in this paper was developed independently from [17] and does not build on it. The main differences between this work and the work of [17] is in the prediction of class labels as explained in III-D.

The following sections give the details of the proposed method.

A. Preprocessing

First the f_0 curve of each recorded utterance in the speech database is extracted using the SWIPE [18] pitch estimator, with a period of 10 ms. The extracted values are then converted to the logarithmic domain. Next, the mean (μ) and standard deviation (σ) of all the valid f_0 values (where f_0 is defined over voiced regions) in the whole speech database are calculated, in order to calculate the statistical z-score (f_{0_z}). The reason for working in the logarithmic domain is that the f_0 distribution approaches the normal curve in the logarithmic domain [19], while it tends to be skewed in the linear domain.

In [16] the f_0 values are expressed as semi-tones ($f_{0_{st}}$) with respect to the overall mean f_0 of the speaker, with the reason being that SLAM was developed for multi-speaker speech prosody research.

Finally the extracted f_0 curves are smoothed with the MOMEL [14] algorithm, which fits quadratic splines over the f_0 curve. This smoothing does two things: it minimizes the microprosodic effects (which can include a sudden raising or lowering of f_0 at a voiced/unvoiced boundary), and it simplifies the stylization due to the interpolation of unvoiced regions. Unvoiced regions in f_0 are interpolated with MOMEL with quadratic splines.

B. Stylization

In SLAM an f_0 curve of arbitrary duration can be stylized. The stylization algorithm assigns abstract symbolic pitch values (discussed in Section III-C) to three points in the *smoothed* curve, namely the initial f_0 point, the final f_0 point, and the main saliency of the f_0 curve. These points are depicted in Figure 1. In the proposed method the stylization is simplified (discussed in Section III-D).

Two different simplified stylizations were examined:

- **Style A**: The initial f_0 point, and the dynamics, or direction of movement to the final f_0 point (this scheme is also depicted in Figure 1).
- **Style B**: Only the dynamics between the initial and final f_0 points are used in the stylization.

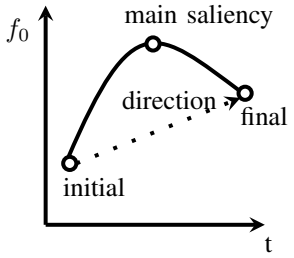

Fig. 1. SLAM vs. proposed f_0 stylization.

C. Coding

In SLAM the abstract symbolic class label assigned to a stylized f_0 curve consists of a concatenation of three sub-labels which are assigned to the initial, final, and the main saliency points of the f_0 curve.

First, the semi-tone frequency axis is quantized into five intervals as given in Table I. Next, the f_0 points are converted to the semi-tone scale, $f_{0_{st}}$ (as mentioned in Section III-A). Finally the $f_{0_{st}}$ points are assigned a class label according to the interval the point falls in from Table I.

If the main saliency point is less than two semi-tones from either the initial or final points then it is omitted from the class label and the class label consists of only the two sub-label components of the initial and final point.

TABLE I
Quantized frequency intervals and sub-labels for SLAM and the proposed method.

Label	**SLAM** *Range in semi-tones*	**Proposed** *Range in σ*
H	$f_{0_{st}} \geqslant 6$	
h	$2 \leqslant f_{0_{st}} < 6$	$f_{0_z} \geqslant 1$
m	$-2 \leqslant f_{0_{st}} < 2$	$-1 \leqslant f_{0_z} < 1$
l	$-6 \leqslant f_{0_{st}} < -2$	$f_{0_z} < -1$
L	$f_{0_{st}} < -6$	

The time axis is also divided into three equal intervals from the start to the end of the f_0 curve. If the main saliency is present (as per above) then it is assigned a positional label (on the time axis): 1 if in first third of the f_0 curve, 2 if in middle of the f_0 curve, and 3 if in last third of the f_0 curve.

As per illustration, assume the initial, final and main saliency points of Figure 1 have semi-tone values of -3, 3 and 7 respectively, and also that the main saliency is in the middle of the curve. This would then be coded as l, h and H as per Table I. The main saliency is more than two semi-tones away from either the initial and final point and therefore does make part of the label, and it's time position is coded as 2 due to being in the middle of the curve. The complete class label would then be coded as $lhH2$.

In the coding scheme of the proposed method the frequency axis is quantized into *three* intervals in terms of the standard deviation of the speaker, also given in Table I. Next, the f_0 points are normalized to the z-score, f_{0_z} (as mentioned in Section III-A). Now, only the first and last f_{0_z} points are

assigned a sub-label according to the interval the point falls in from Table I. Finally, the class label assigned to the stylization for the different styles are given in Table II.

Style A reduces the final point code to the dynamics between the code points, up (u), same (s), or down (d), while Style B just codes the dynamics.

For the stylization in Figure 1 with the same values as used in the SLAM example, the SLAM stylized coded class, $lhH2$, would become lu in Style A of the proposed method and u in Style B of the proposed method.

TABLE II
MAPPING FROM QUANTIZED FREQUENCY INTERVAL SUB-LABELS FOR INITIAL AND FINAL POINTS TO STYLE A AND STYLE B CLASS LABELS OF PROPOSED METHOD.

Initial and final point codes	Style A	Style B
hh	hs	s
hm	hd	d
hl	hd	d
mh	mu	u
mm	ms	s
ml	md	d
lh	lu	u
lm	lu	u
ll	ls	s

There are 400 possible class labels in SLAM, while Style A of the proposed method has 7 and Style B has 3.

D. Modelling

The goal of the model is to assign a class label from Table II to a speech unit based on the synthesis target text.

In general, any machine learning technique can be used and many approaches have been proposed, such as decision trees [6] and HMMs [20] or rule based systems [21]. The speech unit has normally been a syllable, since words may be too long and the intonation may change faster than the stylization can cater for, and phonemes are too short.

Conditional random fields (CRF) were chosen as the machine learning technique to model the classes on syllable units within the context of a *prosodic phrase*, as CRFs are a sequence modeling framework that are well suited for modeling sequential data that is encountered in natural language processing (NLP) problems [22].

Initial results on predicting the intonation class labels on the syllables were poor due to not having enough context of the syllable word and phrase intonation structure. [17] also reported relatively poor results on syllable units, using bidirectional Long Short Term Memory Neural Networks (LSTM) [23] to model the class labels on the syllable contexts.

Subsequently a cascading modelling approach was attempted and proved to predict the class labels with better accuracy (results given in Section IV). In the cascading modelling approach there are three CRF models for predicting the intonation class labels. An intonation class label is first predicted on a phrase tier with all the phrase context. Next, a class label is predicted on a word tier, this time including the phrase tier context and the *predicted* class label of the phrase together with the word tier context. Finally a syllable class label is predicted, using all the context and *predicted* class labels of the phrase and word tiers.

The features used as context for the different utterance tiers are given in Table III and are similar to that of [24].

TABLE III
UTTERANCE TIER SPECIFIC FEATURES USED FOR CRF MODEL TRAINING FOR PREDICTION OF SYLLABLE INTONATION CLASS LABELS.

Utterance tier	Feature description
Utterance	# phrases in utterance # words in utterance
Phrase	Intonation class of phrase # syllables in phrase # words in phrase Phrase position in utterance
Word	Guessed part-of-speech (GPOS) Is the word followed by punctuation A single quote in this or previous word? # syllables in word Word position in phrase Word position in utterance Phrase break tag after word Type of punctuation after word Intonation class of word
Syllable	Syllable nucleus position Segments after syllable nucleus Syllable position in word Syllable position in phrase Syllable identity

IV. EXPERIMENTS AND RESULTS

Two single speaker corpora, with recordings as well as text annotations, were used in order to compare SLAM to the proposed method styles A and B. One American English male speaker, named *rms*, with a duration of 01:06:07.02 (*hours:minutes:seconds*) and 1131 utterances (from the `CMU ARCTIC` [25] database). The other corpus was an isiXhosa female speaker, named *zoleka*, a duration of 01:25:52.36 and 600 utterances.

A. Automatic Stylization, Coding and Modelling

The training phase is a sequential procedure which can be explained as follows:

- **Forced Alignment**: The text and audio data was put through a forced-alignment process. This process uses the Hidden Markov Model Toolkit (HTK) [26] to produce phonetically aligned utterances.
- **Frontend Processing**: The aligned utterances were processed with the NLP frontend of Speect [27] in order to create the utterance structures and assign GPOS[1] labels.
- **Prosodic Phrasing**: A phrasing and phrase break model, used to break the utterance into prosodic phrases, is trained on the text and audio data (described in [29]).
- **Feature Extraction**: Features (as given in Table III) are extracted from the utterances to train the intonation class prediction models.

[1]In this work a *guessed part-of-speech* (GPOS) tagger was used for POS tagging, similar to [28]

- **Automatic Stylization and Coding**: The phrase-, word- and syllable tier elements are stylized and coded as explained in Sections III-B and III-C.

The three tier specific CRF models (phrase, word and syllable) are then trained on the extracted features for the stylized and coded intonation class labels. The prediction results of the three different stylization methods on a held out test set are given in Table IV. The held out test set was a randomly selected subset of 10% of each corpus which were not used in any acoustic modelling (that being duration, pitch or spectra).

TABLE IV
F_1 SCORES OF THE PREDICTED CLASS LABELS OF THE THREE DIFFERENT STYLIZATION METHODS.

Method	**Phrase** F_1	**Word** F_1	**Syllable** F_1
SLAM	0.044	0.126	0.153
Style A	0.279	0.473	0.653
Style B	0.544	0.798	0.788

In [17] a prediction accuracy of 25.2% was reported on a SLAM simplification similar to Style A, on a single speaker United States English Female corpus with a size of 641 000 syllables ($\pm$ 24 hours of speech at an average syllable duration of 140 milliseconds [30])

B. TTS Voice

Three TTS voices were built for each of the data sets. One baseline voice, one voice including the Style A and one including the Style B intonation class labels.

From Table IV it can be seen that the prediction of SLAM class labels are poor. This might be ascribed to the fact that the stylization technique found 39 intonation classes and for most of the classes there are very few examples in the data. For this reason no TTS voices were built using SLAM intonation class labels.

The voices were built using the technique described in [31], except that the phrase-, word- and syllable tier intonation class labels are now included in the linguistic features of [24].

1) Objective Evaluation: Table V gives the root mean squared error (RMSE) f_0 values for the different stylization methods on held out test sets.

TABLE V
RMSE OF f_0 (HZ) ON A HELD OUT TEST SET FOR THE DIFFERENT STYLIZATION METHODS.

Method	*rms*	*zoleka*
Baseline	10.13	36.98
Style A	10.43	43.79
Style B	10.75	42.36

2) Subjective Evaluation: Two subjective evaluations were done on the English (*rms*) data, and one on the isiXhosa (*zoleka*) data. The first evaluation for English compared the baseline voice to the *Style B* method. Twelve respondents were asked to choose a preference between ten sets of two synthesized samples.

The results were significant ($p < 0.05$) with a 59% preference for the Style B method. Then Style A was compare to B, again twelve respondents and ten sets of two synthesized samples. The results were not significant ($p < 0.05$) with a 57% preference for the Style B method.

For isiXhosa the test was the same but there were only three respondents, for baseline vs. *Style B* there was a 56% preference for Style B, whilst for Style A vs. Style B there was a 61% preference for Style B. No significance testing was done due to the low number of respondents.

V. DISCUSSION AND CONCLUSION

In this paper a new automatic stylization, coding and modelling technique for modelling intonation implicitly in SPSS TTS was proposed.

The proposed technique shows a marked improvement on modelling and predicting symbolic intonation class labels from text, when compared to a similar technique in [17], where the corpus used was an order of size larger than the one used in this work. It is believed that this improvement stems from the use of three phonological utterance tiers (phrase, word, syllable) in the prediction of the syllable level intonation class labels, whereas [17] used just the syllable tier.

The technique was evaluated on an English and isiXhosa single speaker TTS corpus and initial results prove promising. Even though the objective scores in Table V do not show an improvement, there is a definite improvement in the "focus" of the intonation on a phrase level.

The intonation class labels are also prominent in the f_0 decision trees used in the SPSS modelling. The technique is particularly well suited to under-resourced languages in TTS due to the automatic methods described requiring no special intonation annotations.

Future work include an in depth analysis of the modelling on tone languages such as isiXhosa, stylization and modelling improvements, and an attempt at deducing the f_0 curve from the intonation class labels.

REFERENCES

[1] D. R. Ladd, *Intonational Phonology*, ser. Cambridge Studies in Linguistics. Cambridge, UK: Cambridge University Press, 1996, no. 79.

[2] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[3] E. Campione, E. Flachaire, D. Hirst, and J. Véronis, "Stylisation and symbolic coding of F0: A quantitative model," in *ESCA Tutorial and Research Worskhop on Intonation: Theory, Models and Applications*, Athens, Greece, September 1997, pp. 71–74.

[4] D. R. van Niekerk, "Tone realisation for speech synthesis of Yorùbá." Ph.D. dissertation, North West University, 2014.

[5] U. D. Reichel, "Linking bottom-up intonation stylization to discourse structure," *Computer Speech & Language*, vol. 28, no. 6, pp. 1340–1365, November 2014.

[6] J. A. Louw and E. Barnard, "Automatic intonation modeling with INTSINT," in *Proceedings of the Fifteenth Annual Symposium of the Pattern Recognition Association of South Africa*, Grabouw, South Africa, November 2004, pp. 107–111.

[7] C. Gussenhoven, "Intonation and interpretation: Phonetics and phonology," in *Proceedings of the First International Conference on Speech Prosody*, Aix-en-Provence, France, 2002, pp. 47–57.

[8] K. Silverman, M. E. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: a standard for labeling English prosody," in *Proceedings of the Second International Conference on Spoken Language Processing (ICSLP)*, Alberta, Canada, October 1992, pp. 867–870.

[9] M. Liberman, "The Intonational System of English," Ph.D. dissertation, MIT, 1975.

[10] S. Baumann, M. Grice, and R. Benzmüller, "GToBI-a phonological system for the transcription of German intonation." in *Proceedings of Prosody*, 2000, pp. 21–28.

[11] A. Rosenberg, "AuToBI- A tool for Automatic ToBI annotation," in *Proceedings of Interspeech*, Japan, September 2010, pp. 146–149.

[12] P. A. Taylor, "A Phonetic Model of English Intonation," Ph.D. dissertation, University of Edinburgh, 1992.

[13] H. Fujisaki and K. Hirose, "Modelling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation," in *In Proceedings of 13th International Congress of Linguists*, 1982, pp. 57–70.

[14] D. Hirst, A. Di Cristo, and R. Espesser, "Levels of representation and levels of analysis for the description of intonation systems," *Prosody: Theory and experiment*, pp. 51–87, 2000.

[15] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Communication*, vol. 46, no. 3, pp. 220–251, 2005.

[16] N. Obin, J. Beliao, C. Veaux, and A. Lacheret, "SLAM: Automatic Stylization and Labelling of Speech Melody," in *Speech Prosody*, Ireland, May 2014, pp. 246–250.

[17] R. Dall and X. Gonzalvo, "JNDSLAM: A SLAM extension for Speech Synthesis," in *Proceedings of Speech Prosody 8*, Boston, USA, May 2016, pp. 1024–1028.

[18] A. Camacho, "SWIPE: A sawtooth waveform inspired pitch estimator for speech and music," Ph.D. dissertation, University of Florida, 2007.

[19] L. Menn and S. Boyce, "Fundamental frequency and discourse structure," *Language and Speech*, vol. 25, no. 4, pp. 341–383, 1982.

[20] T. Nagano, S. Mori, and M. Nishimura, "A stochastic approach to phoneme and accent estimation," in *Proceedings of Interspeech*, Lisbon, Portugal, September 2005, pp. 3293–3296.

[21] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Computer Speech & Language*, vol. 10, no. 3, pp. 155–185, 1996.

[22] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning, ICML*, vol. 1, 2001, pp. 282–289.

[23] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth Annual Conference of the International Speech Communication Association*, Dublin, Ireland, 2014, pp. 338–342.

[24] H. Zen, K. Tokuda, T. Masuko, T. Kobayasih, and T. Kitamura, "A hidden semi-markov model-based speech synthesis system," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 5, pp. 825–834, 2007.

[25] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases for speech synthesis research," Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA, Tech. Rep. CMU-LTI-03-177 http://festvox.org/cmuarctic/, 2003.

[26] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The HTK book," *Cambridge University Engineering Department*, vol. 3, p. 175, 2002.

[27] J. A. Louw, "Speect: a multilingual text-to-speech system," in *Proceedings of the Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa*, Cape Town, South Africa, November 2008, pp. 165–168.

[28] A. Parlikar and A. W. Black, "Data-driven phrasing for speech synthesis in low-resource languages," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Kyoto, Japan: IEEE, 2012, pp. 4013–4016.

[29] J. A. Louw and A. Moodley, "Speaker Specific Phrase Break Modeling with Conditional Random Fields for Text-to-Speech," in *The 27th Annual Symposium of the Pattern Recognition Association of South Africa*, Stellenbosch, South Africa, December 2016.

[30] E. Banzina, L. C. Dilley, and L. E. Hewitt, "The Role of Secondary-Stressed and Unstressed–Unreduced Syllables in Word Recognition: Acoustic and Perceptual Studies with Russian Learners of English," *Journal of Psycholinguistic Research*, vol. 45, no. 4, pp. 813–831, 2016.

[31] J. A. Louw, G. I. Schlünz, W. Van der Walt, F. De Wet, and L. Pretorius, "The Speect text-to-speech system entry for the Blizzard Challenge 2013," in *Blizzard Challenge Workshop 2013*, Barcelona, Spain, September 2013.