ACCENT GROUP MODELING FOR IMPROVED PROSODY IN STATISTICAL PARAMETERIC SPEECH SYNTHESIS

Gopala Krishna Anumanchipalli[†‡] Luís C. Oliveira[‡] Alan W Black[†]

[†]Language Technologies Institute, Carnegie Mellon University, USA
[‡] L^{2}F Spoken Language Systems Lab, INESC-ID / IST Lisboa, Portugal
{gopalakr,awb}@cs.cmu.edu, lco@l2f.inesc-id.pt

ABSTRACT

This paper presents an 'Accent Group' based intonation model for statistical parametric speech synthesis. We propose an approach to automatically model phonetic realizations of fundamental frequency(F0) contours as a sequence of intonational events anchored to a group of syllables (an Accent Group). We train an accent grouping model specific to that of the speaker, using a stochastic context free grammar and contextual decision trees on the syllables. This model is used to 'parse' an unseen text into its constituent accent groups over each of which appropriate intonation is predicted. The performance of the model is shown objectively and subjectively on a variety of prosodically diverse tasks- read speech, news broadcast and audio books.

Index Terms— Intonation Modeling, Prosody, Phonology, Statistical Parametric Speech Synthesis, Foot, Accent Group

1. INTRODUCTION

Intonation (fundamental frequency, F0) is a key expressive component of speech that a speaker employs to convey his intent in delivering a sentence. It encodes a lot more information in the form of structure and type into an utterance than conveyed by the words. The scope of this information may well be beyond words, as broad phonetic phenomena like emphasis [1], or at the frame level, as microprosody, rendering some naturalness to speech [2]. In Text-to-Speech synthesis, text is the only input information from which appropriate intonation has to be predicted.

Initial approaches to intonation generation were primarily rule-based [3][4][5], where phonetic and phonological findings were programmed on computers to generate speech with the desired properties. These methods were overtaken as data driven approaches (e.g., Unit Selection [6]) made it easier to copy-paste pieces of natural F0 contours from a speech database of the desired style [7]. However, the need for small and flexible voices that can fit on mobile devices led way to the next generation of statistical parametric speech synthesizers (SPSS) [8, 9]. In these approaches, average statistics are stored in contextual decision trees, from which predictions are made about unseen text. Today, while spectral quality of synthetic speech is quite acceptable, the prosodic quality is still very poor and is perhaps the weakest component in state-of-the-art speech synthesizers.

Synthetic speech receives the criticism of sounding unnatural and void of affect, because the relationship between the low level intonation contour and the high level input i.e, words is still not well modelled [10]. While in speech science (phonetics and phonology), the F0 contour is discussed at broad levels of syllables, phrases and beyond [11], in practice, all statistical TTS systems analyze and synthesize contours at the frame or at best sub-phonetic levels, generating in the order of about one F0 value for every 5-10 millisecond interval of speech. It has been shown in prior work that this segmental approach to F0 generation is sub-optimal since linguistic features do not have such low resolution to discriminate F0 values at the level of a frame, thereby generating implausible F0 contours, assigning same values to consecutive frames of speech. This artefact of statistical models leads to a perceived 'processed' quality of speech that doesn't retain the dynamic range or functional value of natural speech. There are several broad directions from which these issues are being addressed.

From a speech production perspective, essentially rooted in the Fujisaki model [12] several attempts employ additive strategies for intonation, modeling the F0 contour as a sum of component contours at different (often phonological) levels like the phrase and syllable [13] [14] [15]. These approches preserve the variance in F0 models by essentially distributing it across different levels.

From a statistical modeling standpoint, to address the issue of 'averaging out' of synthetic speech, Tokuda et al., use maximum likelihood parameter generation [16] to improve the local dynamics of synthetic speech. Toda et. al., [17] suggest imposing the variance of natural speech on synthetic speech to improve its perceptual quality. Yu et al., [18] propose splitting the feature set between stronger and weaker context features and building separate models that are optimized for different functions.

Despite all these efforts, synthesizing appropriate intonation has eluded statistical speech synthesizers. This can perhaps be attributed to the disconnect between the theory and practice of intonation. Statistical intonation models use only rudimentary knowledge of intonation theories in them. Also, these theories remain qualitative and descriptive, hardly providing any predictive knowledge about prosody [19], that can be exploited for SPSS. This work attempts to lessen this gap by employing a phonologically sound representational level for modeling F0.

One key aspect in the design of intonation models that effects the quality of the linguistic→prosodic mapping is the representational level at which the contour is modelled. Openmary [20] employs word level pitch target estimation and interpolation strategy for F0. HTS [21] predicts F0 at the HMM state level and does a maximum likelihood based interpolation. Clustergen [8] models and predicts F0 value at the frame level. There is not a general agreement on the right level to model intonation contour for SPSS. We attempt to address precisely this in this work — What is the right level to model intonation for SPSS?

We propose the phonologically motivated "Accent Group" as the modeling unit for intonation. Since accent placement is non-trivial [22], we develop strategies to automatically derive and pre-

dict accent groups from speech data. We use the TILT phonetic scheme [23] to model the F0 contour itself, since any arbitrary excursion on the contour can be efficiently modelled as a TILT vector and the scheme also conforms with established phonological schemes like ToBI [24].

2. SPEECH DATABASES AND BASELINES

In this work, we use three different speech databases, one in each genre of read isolated speech (ARCTIC, SLT [25]), Radio News (BURNC, F2B[26]) and Audiobook (The adventures of Tom Sawyer, TATS [27]). These cover the range of variety in prosodically interesting tasks for SPSS. The baseline systems we use are the Clustergen frame-based SPSS System [8]. In all systems tested, the same set of core features are used. These include the base feature set in Festival and additional features devised on the Stanford dependency parser. There are a few model-specific features specific to each modeling unit considered.

3. INTONATION MODELING IN SPSS

Most SPSS systems employ the Festival speech synthesis architecture [28], which realizes an utterance as a heterogeneous relation graph of phonological constituents [29]. Fig 1 illustrates the prosodic structure used in Festival. An utterance is modelled as sequences of phrases, words, syllables, phonemes, phoneme states and frames.

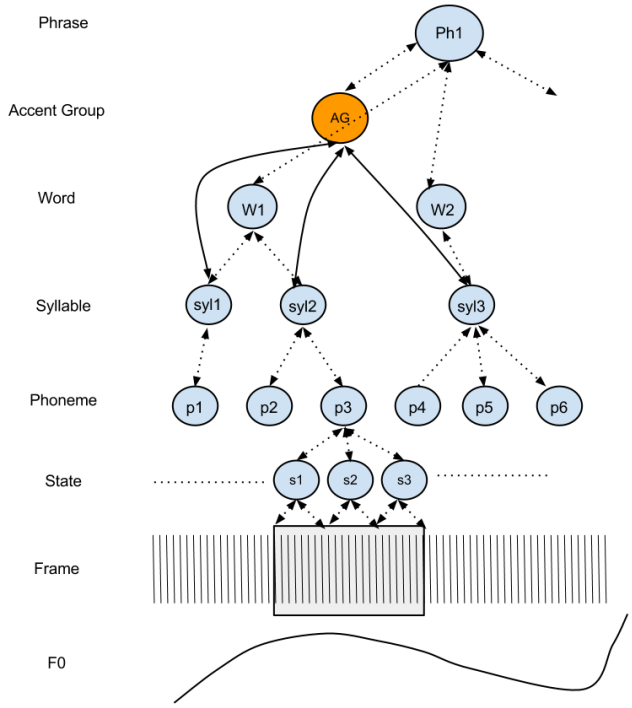


Fig. 1. Illustration of Festival prosodic structure, highlighted is the proposed Accent Group relation.

In Clustergen [8] SPSS system, during training, features regarding each of these levels are extracted to predict the associated spectral/F0 value for each frame. The base features used include those of identity, position, category etc., of each phonological level that a frame corresponds to. These include lexical, syntactic and prosodic features. To capture the quasi-static nature of speech phenomena, the features of respective neighbouring classes are also included. The default feature set uses 61 features. Based on these features, the F0 values are clustered as a CART decision tree. A trained intonation model has questions about these features at the intermediate nodes and leaf nodes contains statistics like mean and variance of the F0 values of frame instances falling under that path of the decision tree. The quesitons selected in the decreasing order of entropy gain against a held out set.

At test time, an appropriate utterance structure is built for an input text sentence and all the associated features are initialized. These features are then used to traverse the built CART models to estimate the parameters to synthesize for each frame. It can be easily seen from the prosodic structure that there is a one-to-many relation between the feature vectors and F0 value. This explains the the lost variance in the finally trained models and consequent prediction of implausible intonation contours at test time.

Our goal in this work is to model each intonational event as itself, without modelling parts and pieces of it, as currently done. Towards realizing this, we introduce a new level within the festival prosodic structure called as the "Accent Group". Each *Accent Group* has one or more syllables as its child nodes and has the *Phrase* as its parent node. The *Accent Group* level is explicitly not linked to the word level since accents could span syllables across words or a word itself can have multiple accents on it [30]. Given an Accent Group, the associated F0 contour is modelled as a TILT vector, which quantitatively describes each event as a 4-valued tuple, comprising the peak position, total length, duration and a shape parameter that can continuously represent any arbitrary rise-fall shape on the contour.

A brief description of the definition of Accent Group, as we use in this work, along with associated training and synthesis procedured is given in the following section.

4. THE ACCENT GROUP IN SPSS

Intonational Phonology views the F0 contour as a sequence of intonational events that can be related to associated syllables. It gives qualitative descriptions about the nature of the event as a 'rise', 'fall', 'dip' etc. in relation to the underlying syllable(s). Each intonational event, often referred to as an accent, could be spread across one or more syllables. The syllables associated with one accent are referred to as its accent group. Further, autosegmental metrical phonology prescribes schemes to organize a syllable sequence in terms of weak and strong syllables to hierchically form intonational phrases of metrical feet. However, when dealing with real speech, most of these prescriptions do not hold. Hence, though we appeal to the idea of grouping syllables, we do not use any definition of what an accent group should be — except that it should have only one accent on it. We use a data-driven approach to automatically determine the accent grouping as appropriate to that particular speaker and speaking style used in the training speech data.

4.1. AUTOMATIC ACCENT GROUP EXTRACTION FROM F0

In order to 'chunk' the syllables of each sentence in the training data as a sequence of accent groups, we employ a resynthesis error minimization algorithm, linear in the number of its syllables. Using TILT as the representation scheme, a decision is made for each syllable whether or not to include it into an accent group. It is included, if

and only if doing so reduces the error (or is within an accepted error threshold ϵ) of the resynthesized F0 contour with respect to the original F0 contour, as compared to modeling it out of the accent group. The exact procedure followed is given as Algorithm 1

Algorithm 1: Algorithm for automatic Accent Group Extraction Method

```
for all phrases do
    accent_group initialized
    for all syllables do
        add syllable to accent_group
        syl_accent = tilt_analyze (log(f0)) over syllable
        syl_err = log(f0) - tilt_resynth(syl_accent)
        accgrp_accent = tilt_analyze (f0) over accent_group
        accgrp_err = log(f0) - tilt_resynth(accgrp_accent)
        if ( accgrp_err ≥ prev_accgrp_err + syl_err + ϵ) then
            accent_group= accent_group - { current syllable}
            /* accent group ended on previous syllable */
            output prev_accgrp_accent
            accent_group = current syllable
            prev_accgrp_err = syl_err
            prev_accgrp_accent = syl_accent
        else
            prev_accgrp_err = accgrp_err
            prev_accgrp_accent = accgrp_accent
        end if
    end for
    if accent_group ≠ ϕ then
        /* accent_group must end at phrase boundary */
        output prev_foot_accent
        accent_group = ϕ
        prev_accgrp_err = 0
    end if
end for
```

ϵ is the acceptable error threshold within which a syllable will be included within the accent group. For the databases experimented in this work Table 1 presents the number of accent groups against number of syllables and words. ϵ was set at 0.01, which is very conservative for log(f0) error. Note that the method retains most syllables and ends up having more than one accent per word on average. The threshold can however be raised so that increasingly more syllables are grouped and resynthesized contours can get excessively smooth, in the limit, modeling the entire phrase as a smooth contour as an "accent".

Table 1. Comparison of the derived accent groups for each task

Task	#words	#syllables	#Accent groups
SLT	9603	12694	10833
F2B	8134	13209	9719
TATS	60444	79225	62762

4.2. SPEAKER-SPECIFIC ACCENT GROUP MODELLING

Given the acoustically derived accent groups for the training data, we model the speaker's grouping as a stochastic context free grammar (SCFG) [31]. The problem of accent group prediction is analogous to prosodic break prediction, where at each word boundary, a decision is made whether or not to have a phrase boundary. In the current scenario, accent groups are analogous to the phrases and syllables are equivalent to words. We employ an approach similar to the one built for such a phrasing model [32]. In order to have a unique set of terminals over which to train an SCFG, the syllables are tagged with six broad boolean descriptors — if the syllable is phrase final, initial, word final or initial, lexically stressed and has a predicted accent on it. Such a scheme uses about 30 combinations of tags in the data presented. Higher number of tags would lead to an increase in the number of tags to process, for which there may not be sufficient data to train an SCFG. To illustrate, a sentence having 4 syllables with 2 accent groups of 1 and 3 syllables each may be represented as —

```
(( syl_1_1_1_1_1_0 ) ( syl_1_1_1_1_0_0 syl_1_1_1_0_0_0
syl_1_0_0_1_0_0 ))
```

Such parses are created using the automatic accent group extraction method and given as the input to the SCFG. Once trained, the grammar can produce parse structures for unseen sequences of syllables in test. While useful, these parses are not very accurate since they encode limited information. However, we use the grammar along with higher level linguistic features on the syllable level to model the accent boundary decision after each syllable. In addition to the conventional syntactic and positional features, we have used dependency parses since we'd like to evaluate the effect of dependency roles and related features in prediction of F0. In all there are about 83 questions from which decision trees are trained for Accent boundary detection in unseen text. In all the three databases, we have about 70% accuracy in Break/Non-Break prediction at all syllable boundaries, compared to the reference sequences.

4.3. F0 MODELING OVER THE ACCENT GROUP

Given the Accent Group boundaries, the F0 contour is analyzed as a 4-valued TILT tuple over each accent group. These are clustered against the feature set specific for the Accent group model, which include features related to the main syllable of the accent group, which we consider as the first lexically stressed syllable of a content word, the features related to the first syllable, last syllable and word level features for these syllables. In all, 63 features were used for the clustering at this stage. The TILT parameter for duration is currently not included in this phase as it is derived from the early phase of phoneme prediction (though we are aware a closer integration of duration prediction could be advantageous). This leaves the TILT amplitude, peak and tilt shape as the vector to be predicted. Mean subtraction and variance normalization is done on these features so as not to bias the models optimized towards one of these values.

5. EXPERIMENTAL RESULTS

The discussed intonation models are applied in TTS and predictions are made about unseen text data. Figure 5 compares the proposed accent group model against the baseline frame based model and the reference F0 contour. It can be seen the variance and peak alignment with reference are much better in the 'Accent Group' intonation model.

While perceptual judgments by human listeners are the main evaluation technique for evaluating intonation [27], it is also important to look at the objective performance of intonation models, at least to highlight how bad usual optimization criteria are. The conventional metrics are Mean error (`err`) and Correlation(`corr`) of predicted F0 contours with respect to the reference contours from the subject. The reference durations are maintained even in the synthetic contour to enable a point-to-point comparison of the reference

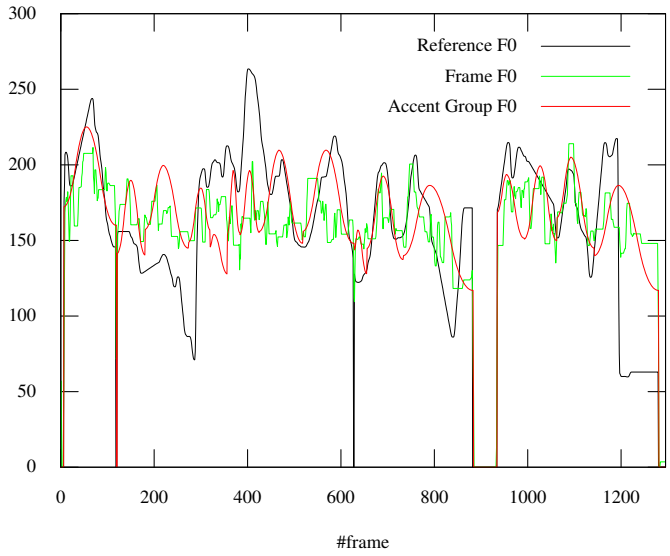


Fig. 2. An example of synthetic F0 contours using the Clustergen default frame model and the proposed Accent group model. The reference is also shown to compare against.

and test intonation contours. Table 2 presents the metrics on the three tasks. The last row 'Accent Group Oracle', is the model where the true accent grouping of the speaker is employed instead of predicted grouping.

Table 2. Objective comparisons proposed vs. default models

	SLT		F2B		TATS	
Unit	err	corr	err	corr	err	corr
Frame	10.97	0.62	37.22	0.38	29.95	0.079
Syllable	12.15	0.47	37.05	0.23	25.28	0.066
Word	12.65	0.46	36.30	0.33	25.80	0.0810
Accent Group	13.13	0.43	35.79	0.33	25.96	0.064
Accent Group Oracle	11.49	0.51	35.50	0.34	24.91	0.092

The primary conclusions from this table are (i) read speech databases have predictable intonation values that statistical models seem to model well. (ii) As the prosodic complexity increases, the default statistical models fail to capture the prosodic variance (iii) As increasingly more data is made available, models employing higher order phonological units tend to converge to similar predictions and (iv) Accent grouping is indeed a hidden part of intonation, when the true accent grouping is provided, F0 estimates are more close to natural in all tasks— better than any other phonological unit.

As RMSE and correlation are not ideal metrics for evaluating perceptual goodness of synthetic intonation [33], we carried out subjective ABX listening tests on pairs of the above models. We have chosen the audio book task for this purpose. We have synthesized a random 45 sentences from the test set. This set was synthesized by each of the candidate intonation models, all other TTS components remaining the same. The listening tests were carried out via crowdsourcing on the Amazon Mechanical Turk, where listeners were asked to select the stimulus they prefer to hear. They can also choose a 'both sound similar' option. Each pair of stimuli was rated by 10 different listeners, making the following preferences reliable.

Fig. 3. Subjective Result: Listener Preference for TTS with Word Vs. Accent Groups as the F0 modeling unit

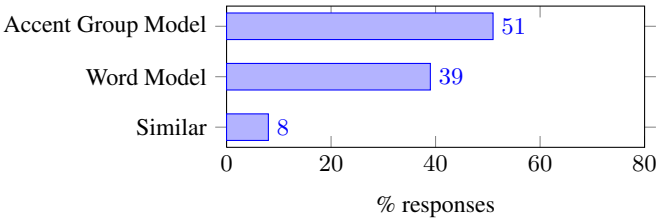


Fig. 4. Subjective Result: Listener Preference for TTS with Syllable Vs. Accent Groups as the F0 modeling unit

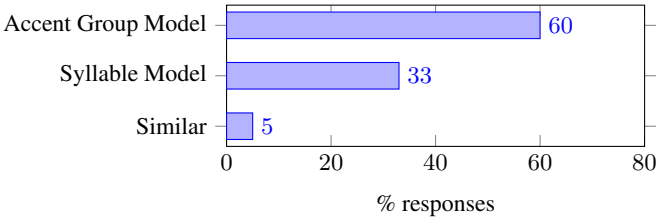
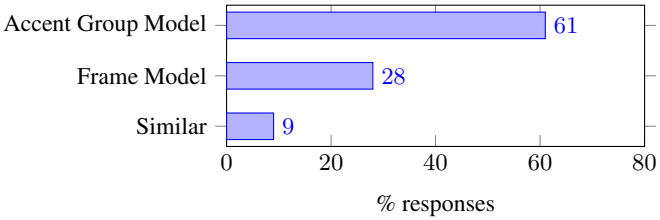


Fig. 5. Subjective Result: Listener Preference for TTS with Frame Vs. Accent Groups as the F0 modeling unit



The user preferences clearly suggest the superiority of the proposed Accent Group model against the reference baseline. They also show that Accent Group intonation model is better than other phonological levels, which is a very welcome observation, since it may mean that the proposed model is language universal. (e.g., for agglutinative languages like Turkish or German where word level intonation models are grossly fallible.)

6. CONCLUSIONS

This work proposes an intonational model for SPSS based on 'Accent Group' as the modeling unit. We have presented algorithms to train such a model from speech data and use it for prediction of appropriate intonation contours from text. We have demonstrated the superior perfomance of the proposed model both objectively and subjectively against the frame-level models currently in use in F0 modeling. The evaluations are shown on three different speaking styles.

7. REFERENCES

[1] D. Bolinger, *Intonation and its Uses*, Stanford University Press, 1989.

[2] J.P.H. van Santen and J. Hirschberg, "Segmental effects on timing and height of pitch contours," in *ICSLP*, Yokohama, 1994, vol. 2, pp. 719–722.

[3] Ignatius G. Mattingly, "Synthesis by Rule of Prosodic Features," *Language & Speech*, vol. 9, pp. 1–13, 1966.

[4] S.J. Young and F. Fallside, "Synthesis by rule of prosodic features in word concatenation synthesis," *International Journal of Man-Machine Studies*, vol. 12, no. 3, pp. 241 – 258, 1980.

[5] M. Anderson, J. Pierrehumbert, and M. Liberman, "Synthesis by rule of English intonation patterns," in *Proceedings of ICASSP 84*, 1984, pp. 2.8.1–2.8.4.

[6] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP96*, Atlanta, GA, 1996, vol. 1, pp. 373–376.

[7] A. Raux and A. Black, "A unit selection approach to F0 modeling and its application to emphasis," in *ASRU2003*, St Thomas, USVI, 2003.

[8] Alan W Black, "Clustergen: A statistical parametric synthesizer using trajectory modeling," in *Interspeech 2006*, Pittsburgh, PA, 2006.

[9] Heiga Zen, Keiichi Tokuda, and Alan W Black, "Review: Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, November 2009.

[10] Paul Taylor, *Text-to-Speech Synthesis*, Cambridge University Press, 2009.

[11] D.R. Ladd, *Intonational Phonology*, Cambridge Studies in Linguistics. Cambridge University Press, 1996.

[12] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The Production of Speech*, P MacNeilage, Ed., pp. 39–55. Springer-verlag, 1983.

[13] J.P.H. van Santen, Alexander Kain, Esther Klabbers, and Taniya Mishra, "Synthesis of prosody using multi-level unit sequences," *Speech Communication*, vol. 46, no. 3-4, pp. 365 – 375, 2005.

[14] Gopala Krishna Anumanchipalli, Luis C. Oliveira, and Alan W Black, "A Statistical Phrase/Accent Model for Intonation Modeling," in *Interspeech 2011*, Florence, Italy, 2011.

[15] Yi-Jian Wu and Frank Soong, "Modeling pitch trajectory by hierachical hmm with minimum generation error training," in *ICASSP 2012*, Kyoto, Japan, 2012.

[16] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *ICASSP*, 2000, vol. 3, pp. 1315–1318.

[17] Tomoki Toda and Keiichi Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE - Trans. Inf. Syst.*, vol. E90-D, pp. 816–824, May 2007.

[18] Kai Yu, Heiga Zen, Francois Mairesse, and Steve Young, "Context adaptive training with factorized decision trees for hmm-based speech synthesis," *Speech Communication*, 2011.

[19] Yi Xu, "Speech prosody: A methodological review," *Journal of Speech Sciences*, vol. 1, no. 1, 2012.

[20] M. Schröder and J. Trouvain, "The german text-to-speech synthesis system mary: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[21] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, and K. Tokuda, "The hmm-based speech synthesis system (hts) version 2.0," *Proc. of Sixth ISCA Workshop on Speech Synthesis*, pp. 294–299, 2007.

[22] Shimei Pan and Julia Hirschberg, "Modeling local context for pitch accent prediction," in *Proceedings of the ACL*, 2000.

[23] P Taylor, "Analysis and synthesis of intonation using the tilt model," *Journal of the Acoustical Society of America*, vol. 107 3, pp. 1697–1714, 2000.

[24] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: a standard for labelling English prosody.," in *Proceedings of ICSLP92*, 1992, vol. 2, pp. 867–870.

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

[26] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Tech. Rep. ECS-95-001, Electrical, Computer and Systems Engineering Department, Boston University, Boston, MA, 1995.

[27] Simon King, "The Blizzard Challenge 2012," in *Blizzard Challenge 2012*, Portland, Oregon, 2012.

[28] A. W. Black and P. Taylor, "The Festival Speech Synthesis System: system documentation," Tech. Rep. HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Scotland, UK, January 1997, Available at http://www.cstr.ed.ac.uk/projects/festival/.

[29] P. Taylor, A. Black, and R. Caley, "Hetrogeneous relation graphs as a mechanism for representing linguistic information," *Speech Communications*, vol. 33, pp. 153–174, 2001.

[30] Esther Klabbers and J.P.H. van Santen, "Clustering of foot-based pitch contours in expressive speech synthesis," in *ISCA Speech Synthesis Workshop V*, Pittsburgh, PA, 2006.

[31] F. Pereira and Y. Schabes, "Inside-outside reestimation from partially bracket corpora," in *Proceedings of the 30th conference of the Association for Computational Linguistics*, Newark, Delaware, 1992, pp. 128–135.

[32] Alok Parlikar and Alan W Black, "Data-driven phrasing for speech synthesis in low-resource languages," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 2012.

[33] R. Clark and K. Dusterhoff, "Objective methods for evaluating synthetic intonation," in *Proc. Eurospeech 1999*, 1999.