Foot-based Intonation for Text-to-Speech Synthesis using Neural Networks

Mahsa Sadat Elyasi Langarani, Jan van Santen

Center for Spoken Language Understanding, Oregon Health & Science University, Portland, OR, USA

{elyasila, vansantj}@ohsu.edu

Abstract

We propose a method ("FONN") for F0 contour generation for text-to-speech synthesis. Training speech is automatically segmented into left-headed feet, annotated with syllable start/end times, foot position in the sentence, and the number of syllables in the foot. During training, we fit a superpositional intonation model comprising accent curves associated with feet and phrase curves. We propose to use a neural network for model parameter estimation. We tested the method against the HMM-based Speech Synthesis System (HTS) as well as against a template based variant of FONN ("DRIFT") by imposing contours generated by the methods onto natural speech and obtaining quality ratings. Test sets varied in degree of coverage by training data. Contours generated by DRIFT and FONN were strongly preferred over HTS-generated contours, especially for poorly-covered test items, with DRIFT slightly preferred over FONN. We conclude that the new methods hold promise for high-quality F0 contour generation while making efficient use of training data.

Index Terms:Prosody, Intonation modeling, Text-to-Speech synthesis, Artificial Neural Networks

1. Introduction

Generating fundamental frequency (F_0) in text-to-speech synthesis (TTS) takes many forms, from rule-based methods in older systems where F_0 is generated by rule and then imposed on concatenated sequences of stored acoustic units [1], to HMM based synthesis in which F_0 is generated frame-wise in parallel with spectral frame generation and is, again, imposed on the spectral frames [2], to unit selection systems with enough recordings that stored F_0 can be used *as-is* [3].

A fundamental issue is whether frame-based methods are able to capture a key property of F_0 movement, which is that they have — except where perturbed or interrupted by obstruents — a *smooth suprasegmental shape* with typically *no more than two inflection points*. A recent study explicitly addressing this issue [4] considered various phonological units in a statistical parametric speech synthesis framework. "Accent group" was defined as a sequence of syllables containing an accented syllable and not necessarily as a (left-headed) foot, which requires that the first syllable is accented (e.g., [5, 6, 7]). Anumanchipalli [4] showed that the best-performing phonological unit is the accent group. This result suggests that we may need to consider units that are larger than the syllable and that, importantly, do not need to coincide with word boundaries.

Earlier, we proposed a rule based F_0 generation method that guarantees that contours will have a smooth suprasegmental shape [8]. Unlike Anumanchipalli et al. [4], we use the foot as phonological unit [9]. It is based on the General Superpositional Model (GSM) [9], which posits that the F_0 curve for a single-phrase utterance can be written as the (generalized) sum of a phrase curve and one accent curve for each foot.

Recently, we proposed a *data-driven* foot-based intonation generator method ("DRIFT") for English language [10] based on the same underlying model as [8]. DRIFT employs a model-based F_0 generation method that guarantees that contours will have a smooth suprasegmental shape [8, 11]. In contrast to Anumanchipalli et al. [4, 12], the phonological unit used in DRIFT is the foot. In [13], we proposed a new intonation adaptation method using the DRIFT to transform the perceived identity of a TTS system to that of a target speaker with a small amount of training data.

In the present study, we propose a foot-based neural network intonation generator ("FONN") for English language that maps foot-based features to accent parameters using a simple ANN. We hypothesize that this method has the advantages of foot-based methods over frame based methods that were demonstrated by DRIFT, but has even lower training data requirements than DRIFT. We note that several Artificial Neural Network (ANN) based parametric methods have been proposed to predict intonation from different phonological units: phonemes [14], syllables [15], phoneme sequences [16], and syllable sequences [17, 18]. The SFC model [17] simulates intonation by superpositionally combining multiple elementary contours that are functionally defined at several levels: word, group, phrase, and clauses. Bailly [17] used a feed-forward neural network (FFNN) with two layers per syllable in the syllable set. The fact that SFC represents prototypical contours summarized from training data means that it avoids direct modeling of any articulatory constraints. This limits its ability for modification. Reddy [18] proposed a two-stage FFNN to predict intonation pattern of a sequence of syllables. The first stage has three layers with 35 dimensional feature vector (per syllable) as input, and three F_0 values — for start, middle, and end of syllable based on tilt model [19] — as the output. The structure of the second stage is the same as that of the first stage with one difference: the input vector of second stage is obtained by concatenating the input and output of the first stage.

We will compare F_0 contours generated by FONN with HTS-generated contours [20] and with DRIFT in a subjective listening experiment with stimuli created by imposing contours generated by the three methods onto natural speech. In this test, we also explore the role of sparsity, by comparing test items whose constituent phoneme sequences, stress patterns, and phrasal structures are well- vs. poorly covered by the training data. This exploration is based on the conjecture that FONN and DRIFT less sensitive to sparsity than HTS. Since DRIFT uses templates associated with *individual* curves in the training data, while FONN computes curves based on multiple observed

This material is based upon work supported by the National Science Foundation under Grant No. 0964468.

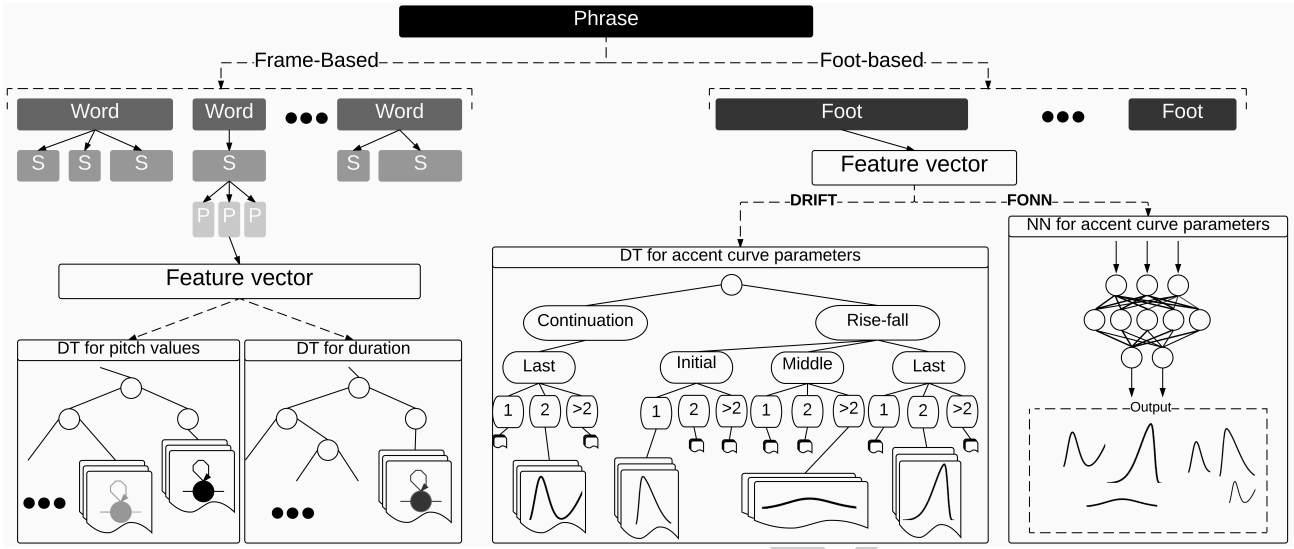

Figure 1: Overview of foot-based and frame-based schemes

curves in the training data, we expect DRIFT to have a relative advantage over FONN in well-covered test data because such data would provide ample stored templates that closely match the test context in terms of the selection features, but the reverse in poorly-covered test data.

2. Model-driven, frame-based F0 generator

Multi-space probability distribution HMM (MSD-HMM) [21] is a special case of using HMM to model observed F_0 values. MSD-HMM includes discrete and continuous mixture HMMs to model F_0. We used the HTS toolkit (version 2.2) [20] to perform HMM-based TTS synthesis. HTS uses the Festival speech synthesis architecture to extract a sequence of contextual and phonological features at several levels, such as, for a given utterance, the phrase, word, syllable, phoneme, and frame levels. As a result, there are many combinations of contextual features to consider when obtaining models. HTS employs decision-tree (DT) based context clustering for handling a large number of feature combinations. The left panel in Figure 1 shows independent DT-based context clustering for F_0 and duration, respectively.

3. DRIFT: Data-driven, foot-based F0 generator

3.1. Intonation model

In a previous study [8], we proposed a new method to decompose a continuous F_0 contour — interpolated in unvoiced regions — into component curves in accordance with the GSM: a phrase curve and a sum of one or more accent curves (one accent per foot).

In this method, the phrase curve consists of two log-linear curves, between the phrase start and the start of the phrase-final foot, and between the latter and the end point of the last voiced segment of the phrase, respectively. We use a combination of the skewed normal distribution and a sigmoid function to model three different types of accent curves. First, the skewed normal distribution is employed to model rise-fall accents that occur in non-phrase-final positions as well as, in statements, in utterance-final positions. Second, a sigmoid function is used to model the rise at the end of a yes/no question utterance. And, third, the sum of the skewed normal distribution and the sigmoid function is used to model continuation accents at the end of a non-utterance-final phrase. (for details, see [8]).

3.2. Analysis

In order to segment training utterances (training and test set selection explain in subsections 5.1 and 5.2) into foot sequences, this method uses three contextual features: accent labels, syllable labels, and phrase boundaries, to automatically create foot boundaries. In contrast with HTS, which uses a large number of contextual features, we only extract five contextual features per foot:

$$Set = \begin{cases} PT\text{: phrase type (statement, continuation)} \\ FPos\text{: foot position in phrase (initial, final, other)} \\ SNum\text{: number of syllables in foot (1, 2, >2)} \\ OD\text{: onset duration of stressed accented syllable} \\ RD\text{: rime duration of stressed accented syllable} \end{cases}$$

A *curve inventory* is created as follows. For each training utterance, we extract F_0 and then fit the intonation model described in subsection 3.1 to compute the phrase curve and the accent curves. We store the vector comprising the estimated accent curve parameters and the values of OD and RD in the inventory. The inventory contains twelve *sub-inventories* defined in terms of the Set features $PT, FPos$, and $SNum$ (middle panel of Figure 1). Because the data were not tagged for y/n (or any) questions, we did not include a y/n question sub-inventory.

3.3. Synthesis

In this method, an input sentence is segmented into phrases, each phrase is segmented into a foot sequence, and for each foot the Set features are extracted from text data. The three first features are extracted from text data, and the value OD and RD are predicted using force alignment applied on original utterances [22]. A suitable accent sub-inventory is chosen for that foot by traversing the proposed DT using the first three features: PT, $FPos$, and $SNum$ (middle panel of Figure 1). We calculate the euclidean distance between the OD, and RD of the current foot and the stored accent curves in the chosen sub-inventory. The five candidate accent curves with the low-

	Poor				Random				Well			
	t-test		Randomization		t-test		Randomization		t-test		Randomization	
	t(49)	p-value	mean	SD	t(49)	p-value	mean	SD	t(49)	p-value	mean	SD
HTS vs. **DRIFT**	7.9034	2.6854*e*-10	1.3277	1.2454	5.9978	2.3584*e*-7	1.1718	1.0709	4.9139	1.0389*e*-5	0.6584	1.4475
HTS vs. **FONN**	6.7803	1.4528*e*-8	0.4120	1.2189	5.7140	6.4363*e*-7	0.2137	1.1669	2.0512	0.0456	0.5868	0.9291
DRIFT vs. **FONN**	-0.6974	0.4888	-0.8512	0.8353	-2.2792	0.0270	-0.1916	0.9297	-2.3892	0.0208	-0.1571	1.0863

Table 1: Results of one-sample t-tests [t-value(df), p-value], and mean and standard deviation (SD) of the randomization-based t-statistic distribution for three pairwise comparisons in three test sets that vary in how well they are covered by the training data

est distance in that sub-inventory are retrieved. To minimize the differences between successive accent curve heights in a phrase, we apply a Viterbi search to the sequence of candidate accent curves; the observation matrix consists of the normalized duration distances and the transition matrix consists of the normalized accent curve height differences.

4. FONN: Foot-based F0 Generator using Neural Networks

4.1. Intonation model

We employ the same intonation model as the DRIFT method (Section 3.1). In this model, the phrase curve consists of two linear curves. We use the combination of the skewed normal distribution and the sigmoid function to model the three different types of accent curves.

4.2. Analysis

For each utterance of *trainSET* (described in Section 5.1, 5.2), we extract F_0 and then fit the intonation model described in subsection 3.1 to compute the phrase curve and the accent curves. We store two vectors as input and target vector for each foot. The input vector comprised of the features from feature Set. We normalize the OD and RD by foot duration. The target vector comprises of the parameters of the estimated accent curve. Before storing the target vector, we normalize the parameters.

We use the input and target vector to train an Artificial Neural Network (ANN) [23]. The ANN consists of two layers as shown in the right panel of Figure 1. The input and output dimensions are 10 and 7, respectively. The hidden layer size is 200. The activation function in the hidden layer is sigmoid and the activation function in the output layer is linear.

4.3. Synthesis

Like the DRIFT method (Section 3.3), an input sentence is segmented into phrases, each phrase is segmented into a foot sequences, and for each foot the Set features are extracted from text data. These feature vectors sequentially are given to the trained ANN to predict accent curves parameters. We use the predicted parameters to create accent curves for each foot.

5. Experiments

5.1. Databases

We use a US English female speaker of the CMU arctic database (SLT) [24]. This corpus contains 1132 utterances, which are recorded at 16bit 32KHz, in one channel. The database is automatically labelled by CMU Sphinx using FestVox labeling scripts. No hand corrections are made.

5.2. Set coverage

In data driven approaches, data sparsity is a pervasive challenge [25]. To investigate the effects of sparsity, we employ an algorithm [10] to select four subsets of the data: *trainSet*, containing the training data; *wellSET*, containing test data that are well covered by *trainSET*; *poorSET*, containing test data that are poorly covered by *trainSET*; and *randomSET*, a random selection from the test data. The algorithm iterates to find a *wellSET* and *poorSET* that are maximally different in terms of their coverage by *trainSET*.

5.3. Evaluation

For subjective evaluation of the intonation generation performance of the three approaches, we design a test that measures naturalness. We use Amazon Mechanical Turk [26], with participants who have approval ratings of at least 95% and were located in the United States.

We prepare three separate tests to compare each pairs of three approaches combination (HTS Vs. DRIFT, HTS Vs. FONN, and DRIFT Vs. FONN). For each pairs, We use a comparison test to evaluate the naturalness of the F_0 contours synthesized by the two approaches. In this test, listeners hear two stimuli with the same content back-to-back and then are asked which they prefer using a five-point scale consisting of -2 (definitely First one), -1 (probability First one), 0 (unsure), +1 (probability Second one), +2 (definitely Second one). We randomly switch the order of the two stimuli. The experiment includes 50 utterance pairs for each of the three test sets (total 150 pairs). We employed 150 listeners, that each listener only can chose one test set (i.e., *poorSET, randomSET,* and *wellSET*) to judge. Three trivial-to-judge utterance pairs are added to the experiment to filter out unreliable listeners.

We evaluate the two approaches by imposing the F_0 contours generated by the two approaches onto recorded natural speech, thereby ensuring that the comparison strictly focused on the quality of the F_0 contours and is not affected by other aspects of the synthesis process [27]. To ensure that the F_0 contours are properly aligned with the phonetic segment boundaries of the natural utterance, the contours are time warped so that the predicted phonetic segment boundaries correspond to the segment boundaries of the natural utterance. Note that the predicted phonetic segment boundaries are the same for the two approaches. To compute the segment boundaries of the natural utterance, we used the HTS state duration and phoneme duration. Finally, we use PSOLA to impose the synthetic contour onto the natural recording[1].

Figure 2 shows the results of the pair-wise comparisons between the naturalness of the F_0 contours synthesized by the two configuration pairs (HTS-DRIFT, HTS-FONN, and DRIFT-FONN). In general, perceptual results indicated superior performance of DRIFT and FONN over HTS. DRIFT over-performed FONN in random and well coverage cases. For significance testing, we first compute a score for each utterance using Equation 1, and then, separately for each test set, apply a one-sample t-test (results are summarized in Table 1). In Equation 1, j, n, m, and C_{ji} stand for j^{th} utterance of current test set, number of listeners, number of utterance of current test set, and the rating of the i^{th} listener for the j^{th} utterance, respectively, and ‖

[1]The synthetic waves are available under following repository: http://cslu.ohsu.edu/~elyasila/wav_SP16/

indicates the absolute value.

$$score_j = \frac{\sum_{i=1}^{n}(C_{ji}|C_{ji}|)}{\sum_{j=1}^{m}(\sum_{i=1}^{n}(|C_{ji}|))} \quad , C_{ji} \in \{-2,-1,0,1,2\} \quad (1)$$

Conventional t-test results for the first and second comparisons (Table 1, first and second rows) show that the scores for DRIFT and FONN are significantly better than those for HTS for all test sets. The third comparison (Table 1, last row) indicates that the scores for DRIFT and FONN differ significantly from each other for two test sets (random and well), but are the same for poor test set. The superiority of FONN over HTS, but not that of DRIFT over HTS, is reduced in the *well*SET.

For showing the robustness of the t-test results, we also perform a randomization test for each comparison in each test set. We randomly change the signs of all ratings, compute the scores for each utterance, and then calculate the t statistic. We repeat these steps 2000 times. The means and standard deviations of the resulting distributions are reported in Table 1, confirming the conventionally obtained significance levels. For example, the t-value of the first comparison (HTS-DRIFT) for *poorSET* is far from the chance (e.g., 7.90 deviates by 5.3 standard deviations from the randomization mean of 1.33, for a normal t(49)-distribution with mean 1.33 and SD of 1.24, this yields a chance of 5.79*e*-8).

In another experiment, we perform a test in which we compare the systems based on the impact of coverage. We first compute a difference score for each utterance, defined by the difference between the scores for the two approaches, and subsequently perform a two-sample t-test comparing these difference scores between the *poorSET* and *wellSET* data. Only for HTS-FONN comparison, statistically significant results were found($t(49) = -3.5675$, $p = 2.8036e-4$, one-tailed; these results were again confirmed using a randomization tests). This result shows a powerful significant trend for the impact of coverage to be stronger for the HTS approach than for FONN. Figure 2 (gray curve, right y-axis) also shows the results of comparing the two systems in terms of the impact of coverage.

6. Conclusion

We proposed a neural network foot-based intonational approach (FONN) for F_0 generation, with these key characteristics. First, like DRIFT, usage of a superpositional model in which selected accent curves are added to a phrase curve [8]. Second, also like DRIFT, usage of a structured inventory of fitted accent curves. Third, unlike DRIFT which uses accent curve parameter *templates*, usage of a trainable parametric method to *compute* accent curve parameters. Both FONN and DRIFT methods result in F_0 curves that are guaranteed to have the desired smooth suprasegmental shapes, and are well-suited to handle sparse training data as well. Perceptual results indicated superior performance of FONN and DRIFT compared to a frame-based approach (HTS). Using a test data selection algorithm, we were able to evaluate the impact of sparsity, with results that tentatively confirmed, as we predicted, the ability of the FONN to handle sparse training data better than HTS. However, there was a trend for DRIFT to outperform FONN except in *poor*SET. We surmise that, for speech synthesis, template based approaches such as DRIFT that create accent curves that inherently preserve natural detail are to be preferred over approaches that compute accent curves. It remains to be seen, however, whether FONN may nevertheless outperform template based approaches in exceptionally sparse data conditions where several slots in the template tree are missing.

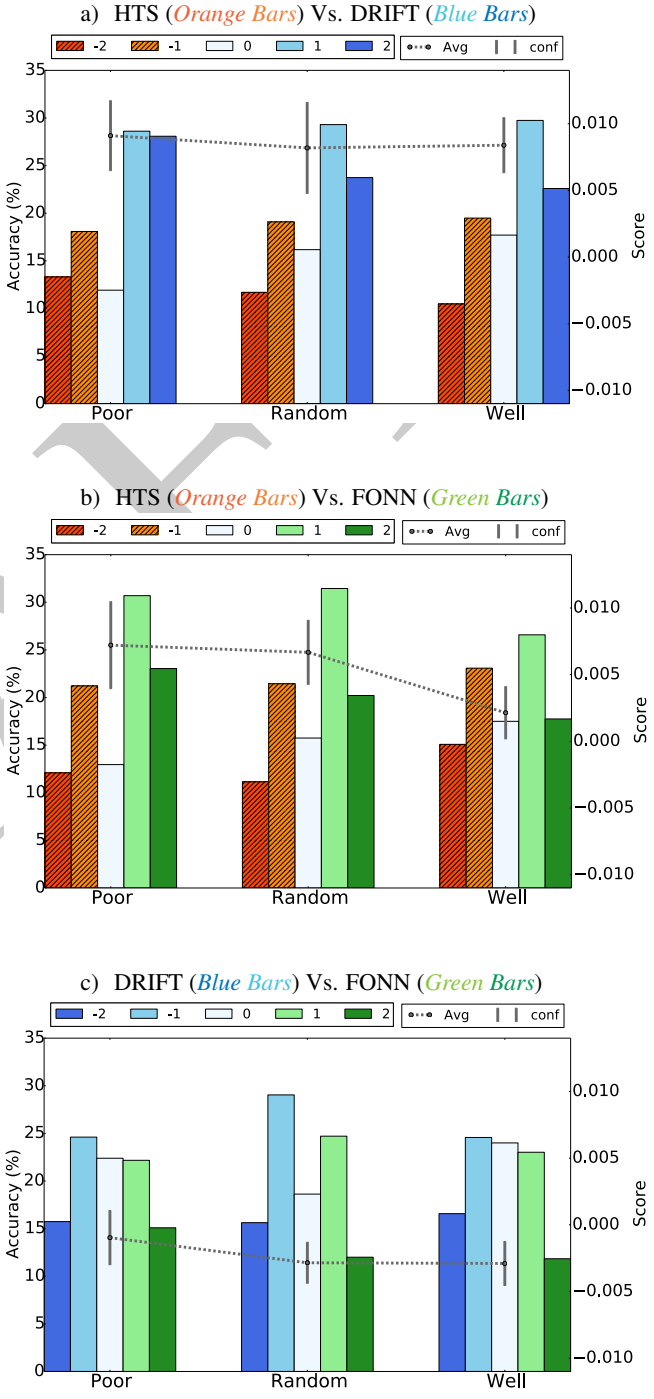

Figure 2: Each group-bars (poor, random, and well) represent the histogram (in percentage (left y-axis)) of the related preference point: The five-point scale consisting of -2 (definitely first version), -1 (probability first), 0 (unsure), +1 (probability second), +2 (definitely second). Dotted line and confidence intervals correspond to the values (right y-axis) computed via Equation 1.

7. References

[1] R. Sproat, *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Boston, MA: Kluwer, 1997.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," 1999.

[3] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The ibm expressive text-to-speech synthesis system for american english," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1099–1108, 2006.

[4] G. K. Anumanchipalli, "Intra-lingual and cross-lingual prosody modelling," Ph.D. dissertation, Google Inc, 2013.

[5] J. P. van Santen, A. Kain, E. Klabbers, and T. Mishra, "Synthesis of prosody using multi-level unit sequences," *Speech Communication*, vol. 46, no. 3, pp. 365–375, 2005.

[6] J. P. van Santen, E. Klabbers, and T. Mishra, "Toward measurement of pitch alignment," *Italian Journal of Linguistics*, vol. 18, no. 1, p. 161, 2006.

[7] E. Morley, E. Klabbers, J. P. van Santen, A. Kain, and S. H. Mohammadi, "Synthetic f0 can effectively convey speaker id in delexicalized speech." in *INTERSPEECH*, 2012.

[8] M. S. Elyasi Langarani, E. Klabbers, and J. P. van Santen, "A novel pitch decomposition method for the generalized linear alignment model," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2584–2588.

[9] J. P. Van Santen and B. Möbius, "A quantitative model of fo generation and alignment," in *Intonation*. Springer, 2000, pp. 269–288.

[10] M. S. Elyasi Langarani, J. P. van Santen, S. H. Mohammadi, and A. Kain, "Data-driven foot-based intonation generator for text-to-speech synthesis." in *INTERSPEECH*, 2015.

[11] M. S. Elyasi Langarani and J. P. van Santen, "Modeling fundamental frequency dynamics in hypokinetic dysarthria," in *Spoken Language Technology (SLT), 2014 IEEE International Workshop on*. IEEE, 2014.

[12] G. Krishna Anumanchipalli, L. C. Oliveira, and A. W. Black, "Accent group modeling for improved prosody in statistical parameteric speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6890–6894.

[13] M. S. Elyasi Langarani and J. P. van Santen, "Speaker intonation adaptation for transforming text-to-speech synthesis speaker identity," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE International Workshop on*. IEEE, 2015.

[14] M. S. Scordilis and J. N. Gowdy, "Neural network based generation of fundamental frequency contours," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. IEEE, 1989, pp. 219–222.

[15] C. Traber, "F0 generation with a data base of natural f0 patterns and with a neural network," in *The ESCA Workshop on Speech Synthesis*, 1991.

[16] M. Vainio *et al.*, *Artificial neural network based prosody models for Finnish text-to-speech synthesis*. University of Helsinki, 2001.

[17] G. Bailly and B. Holm, "Sfc: a trainable prosodic model," *Speech Communication*, vol. 46, no. 3, pp. 348–364, 2005.

[18] V. R. Reddy and K. S. Rao, "Two-stage intonation modeling using feedforward neural networks for syllable based text-to-speech synthesis," *Computer Speech & Language*, vol. 27, no. 5, pp. 1105–1126, 2013.

[19] P. Taylor, "The rise/fall/connection model of intonation," *Speech Communication*, vol. 15, no. 1, pp. 169–186, 1994.

[20] H. Zen, K. Oura, T. Nose, J. Yamagishi, S. Sako, T. Toda, T. Masuko, A. W. Black, and K. Tokuda, "Recent development of the hmm-based speech synthesis system (hts)," 2009.

[21] T. Masuko, K. Tokuda, N. Miyazaki, and T. Kobayashi, "Pitch pattern generation using multispace probability distribution hmm," *Systems and Computers in Japan*, vol. 33, no. 6, pp. 62–72, 2002.

[22] A. Black, P. Taylor, R. Caley, R. Clark, K. Richmond, S. King, V. Strom, and H. Zen, "The festival speech synthesis system, version 1.4. 2," *Unpublished document available via http://www. cstr. ed. ac. uk/projects/festival. html*, 2001.

[23] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 19–23.

[24] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[25] J. P. van Santen and A. L. Buchsbaum, "Methods for optimal text selection," in *Fifth European Conference on Speech Communication and Technology*, 1997, pp. 553–556.

[26] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk — a new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, January 2011.

[27] S. H. Mohammadi and A. Kain, "Transmutative voice conversion," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6920–6924.