Data-driven Foot-based Intonation Generator for Text-to-Speech Synthesis

Mahsa Sadat Elyasi Langarani, Jan van Santen, Seyed Hamidreza Mohammadi, and Alexander Kain

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

{elyasila, vansantj, mohammah, kaina}@ohsu.edu

Abstract

We propose a method for generating F0 contours for text-to-speech synthesis. Training speech is automatically annotated in terms of feet, with features indicating start and end times of syllables, foot position, and foot length. During training, we fit a foot-based superpositional intonation model comprising accent curves and phrase curves. During synthesis, the method searches for stored, fitted accent curves associated with feet that optimally match to-be-synthesized feet in the feature space, while minimizing differences between successive accent curve heights. We tested the proposed method against the HMM-based Speech Synthesis System (HTS) by imposing contours generated by these two methods onto natural speech, and obtaining quality ratings. Test sets varied in how well they were covered by the training data. Contours generated by the proposed method were preferred over HTS-generated contours, especially for poorly-covered test items. To test the new method's usefulness for processing marked-up text input, we compared its ability to convey contrastive stress with that of natural speech recordings, and found no difference. We conclude that the new method holds promise for generating comparatively high-quality F0 contours, especially when training data are sparse and when mark-up is required.

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

1. Introduction

The control of fundamental frequency (F_0) in speech synthesis can take many forms, ranging from rule-based methods in generally older systems whereby F_0 contours are 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 where the data base is sufficiently rich that stored F_0 can be used *as-is* [3].

A fundamental issue is whether frame-based methods are able to capture the property of F_0 movement—except where perturbed or interrupted by obstruents—to have a *smooth polysyllabic shape*. For example, in English, standard H*L pitch accents involve a smooth rise in course of the accented syllable followed by a descent until the next accented syllable or phrase boundary [4, 5, 6, 7]. A recent study explicitly addressing this issue [8] considered various phonological units in a statistical parametric speech synthesis framework, including the frame, syllable, word, accent group, phrase, and sentence. "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., [9, 10, 11]). Anumanchipalli [8] 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, in addition, do not need to coincide with word boundaries.

A second fundamental issue is how well an F_0 generation method performs when input text is marked up to create prosodic constellations that are not present in the training data. For example, suppose that one instructs the system, via markup, to convey strong contrastive stress, can the system create compelling-sounding contrastive stress when the training data do not contain any instances of contrastive stress?

We propose an F_0 generation method that guarantees that contours will have a smooth polysyllabic shape. In contrast to Anumanchipalli et al. [8, 12], the phonological unit we use is the foot. The method can be used in combination with any synthesis method that allows for F_0 to be imposed on the spectral frame sequence via appropriate signal modification methods; it thus excludes only certain unit selection systems, namely those in which either one wants to avoid signal modification at any cost or no such modification is needed because of the all-encompassing coverage provided by a vast training corpus. The method is a hybrid that combines concepts of classic superpositional F_0 modeling [7, 13] with data driven methodology. It is based on the General Superpositional Model (GSM) [7], which posits that the F_0 curve for a single-phrase utterance can be written as the (generalized) sum of a phrase curve and any number of accent curves, one for each foot. The intonation model assumes that accent curves can be described by certain parametric curves [14, 15], and uses an efficient optimization method for decomposing the pitch contour into the phrase curve and the accent curves.

We will compare F_0 contours generated by the new method with HTS-generated contours [16] in a subjective listening experiment with stimuli created by imposing contours generated by the two methods onto natural speech. In this test, we also explore the role of sparsity, by comparing tests 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 the new method is less sensitive to sparsity than HTS. In a second experiment, we determine the ability of the proposed method to convey contrastive stress. This serves to demonstrate the ability of the method to generate F_0 contours from marked-up input text.

2. Model-driven frame-based intonation generator

2.1. Intonation model

Multi-space probability distribution (MSD) HMM [17] is a special case of using HMM to model observed F_0 values.

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

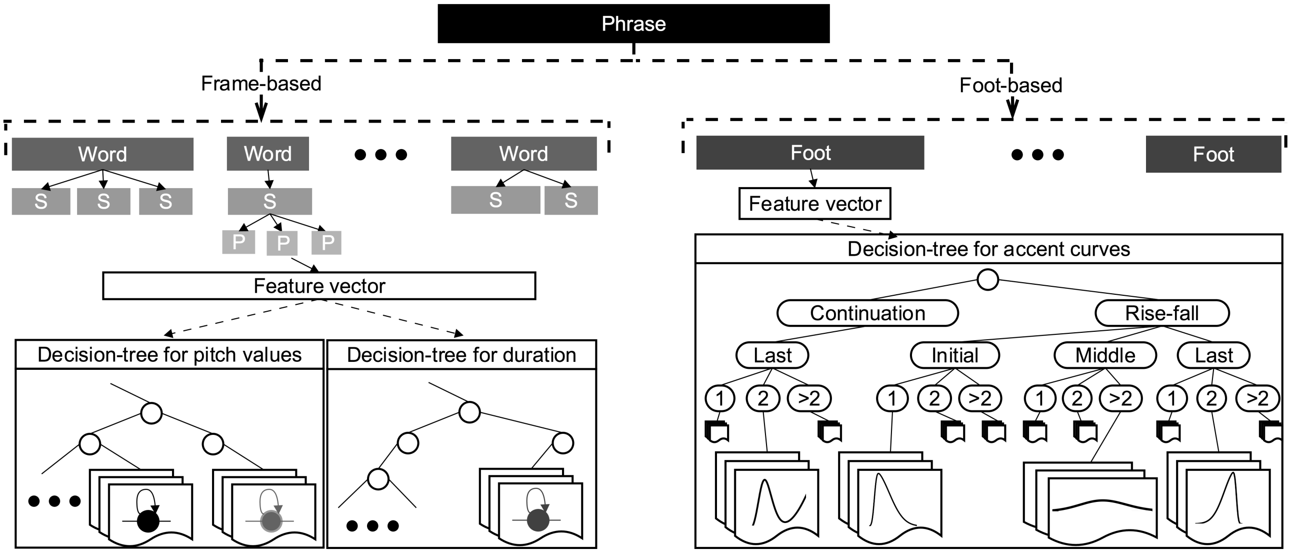

Figure 1: Overview of foot-based and frame-based schemes

MSD-HMM includes discrete and continuous mixture HMMs to model F_0. The state output probability is defined by an MSD, which is a joint distribution of discrete F_0 values and voicing labels [18].

2.2. Analysis

We used the HTS toolkit (version 2.2) [16] to perform HMM-based TTS synthesis. HTS uses the Festival speech synthesis architecture to extract a sequence of contextual and phonological features at severals 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.

2.3. Synthesis

Synthesis comprised these steps: A to-be-synthesized sentence is converted into a contextual label sequence; the utterance HMM is constructed by concatenating the context-dependent state HMMs given the label sequence; state durations of the utterance HMM are determined [19]; a sequence of pitch values (one value per frame), including a voiced/unvoiced label, is generated given the utterance HMM and the state durations. The left panel in Figure 1 shows independent DT-based context clustering for F_0 and duration, respectively.

3. Data-driven foot-based intonation generator (DRIFT)

3.1. Intonation model

In a previous study [14], 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 ($P(t)$ in Equation 1) and a sum of one or more accent curves ($A(t)$ in Equation 1).

$$F_0(t) = P(t) + A(t) \tag{1}$$

In this method, the phrase curve consists of two log-linear curves, between the phrase start and the start of the phrase-final foot (generally associated with the nuclear pitch accent), 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 (f in Equation 2). Second, a sigmoid function is used to model the rise at the end of a yes/no question utterance (g in Equation 2). And, third, the sum of the skewed normal distribution (f) and the sigmoid function (g) is used to model continuation accents at the end of a non-utterance-final phrase (h in Equation 2). The number of accent curves (n) is equal to the number of feet in a phrase (equation 2). The a and b parameters are binary and are used to compactly express the three accent types as

$$A(t) = \sum_{i=1}^{n} (b_i(a_i f(t) + (1-a_i)g(t)) + (1-b_i)h(t)). \tag{2}$$

For example, a yes/no question sentence with two feet (rise-fall ($\%LH^{*}L$) and yes/no question ($L^{*}H\%$) accent types) is represented by $a_1 = 1, b_1 = 1$ and $a_2 = 0, b_2 = 1$, respectively. In Equation 3 and 4, C and D stand for the amplitudes of the accent curves. The two parameter sets $\{\omega, \xi, \alpha\}$ and $\{\beta, \gamma\}$ indicate {scale, location, skewness} of the skewed normal distribution, and {slope, location} of the sigmoid function. These parameters together with the three parameters of the phrase curve are optimized using Sequential Least Squares Programming (for details, see [14]).

$$f(t) = C\frac{2}{\omega}\phi(\frac{t-\xi}{\omega})\Phi(\alpha(\frac{t-\xi}{\omega})) \tag{3}$$

$$g(t) = D\,\frac{1}{1+e^{-\beta(t-\gamma)}} \tag{4}$$

3.2. Analysis

In order to segment training utterances (training and test set selection are explained in subsections 4.1 and 4.2) into foot sequences, our 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:

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

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 then extract for each foot the root weighted mean square error between the modeled accent curve and the raw accent curve, defined as the raw F_0 curve minus the fitted phrase curve. The weight is computed as the multiplication of the voicing flag and the normalized signal energy. If the error is under a certain threshold (in this study 5.0), we store the vector comprising of the estimated accent curve parameters plus the value of $SASD$ and the phoneme identity of $SASN$ for the current foot in the inventory; otherwise, this vector is discarded. The inventory contains twelve *sub-inventories* defined in terms of the Set features PT, $FPos$, and $SNum$ (right 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.

To determine how distinct these sub-inventories are, we performed a classification experiment. We employed an RBF kernel based SVM [20] to classify each pair of sub-inventories by using these features: all accent curve parameters plus $SASD$. The $F1$ averages over all inter sub-inventories were: Continuation-Last(0.4917, $range = [0.3500, 0.6250]$), and Rise-Fall-{First(0.8228, $range = [0.7500, 0.8839]$), Middle(0.8595, $range = [0.7631, 0.9387]$), Last(0.6325, $range = [0.5647, 0.6779]$)}. This shows that we can merge the *continuation* sub-inventories. It also shows, for rise-fall cases, that accent curves vary systematically as a function of the Set features, which thereby validates our feature scheme.

3.3. Synthesis

In our 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. The four first features are extracted from text data, and the value of $SASD$ are predicted using HTS. A suitable accent sub-inventory is chosen for that foot by traversing the proposed DT using the first three features: PT, $FPos$, and $SNum$ (right panel of Figure 1). We, additionally, attempt to match the $SASN$ of the current foot and that of the stored accent curves; if no such match is found, this feature is ignored. Next, we calculate the manhattan distance between the $SASD$ of the current foot and the stored accent curves in the chosen sub-inventory. The five candidate accent curves with the lowest 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. Experiments

4.1. Databases

We use a US English female speaker of the CMU arctic database (SLT) [21]. 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.

Algorithm 1 Automatic selection of test data

```
for 2000 iteration do
    $A \leftarrow$ Choose randomly half of data for training set
    for A do
        $Frq \leftarrow$ Extract the frequency of all features
    end for
    $B \leftarrow$ data - A
    for B do
        $C \leftarrow$ Replace (features of B's, number from Frq)
        $C \leftarrow$ Sort (C, ascending order)
        $F1 \leftarrow$ median of the C
        $F2 \leftarrow$ number of zero in the C
        $DisWell \leftarrow$ Euclidean((1,0),(F1,F2))
        $DisPoor \leftarrow$ Euclidean((0,1),(F1,F2))
    end for
    $Well \leftarrow$ Choose the lowest 50 from DisWell
    $Poor \leftarrow$ Choose the lowest 50 from DisPoor
end for
$trainSET \leftarrow$ choose more frequent samples from As
$wellSET \leftarrow$ choose 50 more frequent samples from Wells
$poorSET \leftarrow$ choose 50 more frequent samples from Poors
$randomSET \leftarrow$ choose randomly 50 samples from remaining data
```

4.2. Set coverage

In data driven approaches, data sparsity is a pervasive challenge [22]. To investigate the effects of sparsity, we employ an algorithm (Algorithm 1) 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*. Units used to compute coverage are as follows. They include the diphone, which is commonly used as a feature for set coverage [23] because it does not have sparsity of triphone and context independency of phonemes. They also include prosodic context, via syllable (lexical) stress and word accent labels. Thus, each sentence is represented as a sequence of $diphone/stress/accent$ features, which is then fed into the algorithm.

4.3. Evaluation

For subjective evaluation of the intonation generation performance of the two approaches, we design two tests: the first test measures naturalness and the second test measures the ability to convey contrastive stress. We use Amazon Mechanical Turk [24], with participants who have approval ratings of at least 95% and were located in the United States.

4.3.1. Naturalness test comparing HTS and DRIFT

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) [25]. We randomly switch the order of the two stimuli. The experiment is administered to 50 listeners, with each listener judging 50 utterance pairs for each test set (i.e., *poorSET, randomSET,* and *wellSET*). 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. 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 for the three test sets. For significance testing, we first compute a score for each utterance using Equation 5, and then, separately for each test set, apply a one-sample t-test. In Equation 5, 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 || indicates the absolute value.

$$score_j = \frac{\sum_{i=1}^{n}(C_{ji}|C_{ji}|)}{\sum_{j=1}^{m}(\sum_{i=1}^{n}(|C_{ji}|))} \qquad , C_{ji} \in \{-2,-1,0,1,2\} \quad (5)$$

Conventional *t*-test results show that the scores of the two methods different significantly from each other for each test set: *poorSET* ($t(50) = 8.6482$, $p = 1.9832e$-11), *randomSET* ($t(50) = 6.5282$, $p = 3.5707e$-8), and *wellSET* ($t(50) = 4.4943$, $p = 4.4943e$-5). We also perform a randomization test for the same difference by (2000 times) randomly changing the signs of all ratings, computing the scores for each utterance, and calculating the *t* statistic. The means and standard deviations of the resulting distributions are: *poorSET* $(0.7403,\ 1.7791)$, *randomSET* $(-0.2341,\ 1.2940)$, and *wellSET* $(0.2241,\ 1.3972)$, thus confirming they conventionally obtain significance levels.

Figure 2 (green curve) also shows the results of comparing the two systems in terms of the impact of coverage. For this, we perform a test in which we, first, compute for each utterance a difference score, defined by the difference between the scores of the two approaches, and subsequently perform a two-sample t-test comparing these difference scores between the *poorSET* and *wellSET* data. The results show a statistically significant trend — but not a powerful one — for the impact of coverage to be indeed stronger for the HTS approach than for DRIFT ($t(50) = 1.7198$, $p = 0.044$, one-tailed).

4.3.2. Testing ability to synthesize text marked up for contrastive stress

To evaluate the ability of DRIFT to handle marked-up input, we design a contrastive emphasis test. First we select 22 sentences from the test data that contain a pair of noun-adjective words for which contrastive stress is meaningful [26]. Then for each of these sentences, we generate two utterances such that in each utterance one of the two words is emphasized. For example, for the sentence “This is a red house”, with capitals indicating stress, we consider “This is a RED house” and “This is a red HOUSE”. For generating the pitch curves, we use the algorithm in subsection 3.3, and then implement a simple rule whereby we increase and decrease amplitudes of the accent curves associated with the emphasized and non-emphasized words by multiplication with factors of 3 and 0.5, respectively.

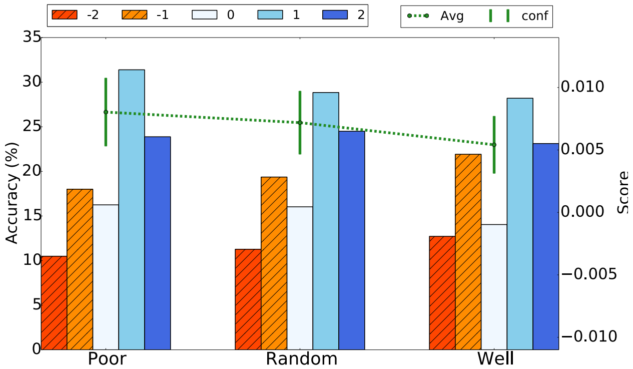

Figure 2: Preference score: The five-point scale consisting of -2 (definitely HTS), -1 (probability HTS), 0 (unsure), +1 (probability DRIFT), +2 (definitely DRIFT)

In the perceptual test, each listener is asked to imagine the following situation: “Two people, John and Mary, are having a dialogue; unfortunately, John is not a good listener so that Mary has to repeat what she just said, emphasizing the word that John — apparently — got wrong. Your task is to figure out which word John got wrong.” The experiment is administered to 50 listeners with each listener judging 44 (22*2) sentences. The percentage of emphasized words conveyed correctly is 84.85%. We also apply the same test for a recorded natural voice (female native American English speaker) for the 44 sentences, and obtain a nearly-identical accuracy of 85.15%. We conclude that the DRIFT’s ability to convey contrastive stress is comparable to that of natural speech.

5. Conclusion

We proposed a data driven foot-based intonational approach (DRIFT) for F_0 generation, with these key characteristics. First, its use of a structured inventory of fitted accent curves, using an accent curve model proposed in [14]. Second, usage of a data driven (fitted) accent curve selection process, in which curves are selected based on (1) the distance in a low-dimensional feature space between a foot in the to-be-synthesized sentence and the feet associated with the accent curves in the inventory and (2) height differences between successive accent curves. Third, usage of a superpositional model in which selected accent curves are added to a phrase curve. In combination, this results in F_0 curves that are guaranteed to have the desired smooth polysyllabic shapes, and are well-suited to handle sparse training data as well. Perceptual results indicated superior performance of DRIFT compared to a frame-based model (HTS). Using a test data selection algorithm, we were able to evaluate the impact of sparsity, with results that tentatively confirmed the ability of the DRIFT to handle sparse training data better than HTS. Finally, we showed that the DRIFT, via markup, can generate compelling contrastive stress.

Future work will focus on creating a parallel inventory of fitted phrase curves and on further improvements of the per-foot distance measure, such as one based on phoneme-class accented syllable structure. Also this foot-based intonation generator can be used for converting pitch contours using neural networks [27].

[1] The synthetic waves are available under following repository: http://www.freesound.org/people/mahsaelyasi/

6. 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] J. Vaissière, "10 perception of intonation," *The handbook of speech perception*, p. 236, 2008.

[5] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

[6] P. Lieberman, "Intonation, perception, and language." *MIT Research Monograph*, 1967.

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

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

[9] 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.

[10] 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.

[11] 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.

[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] H. Fujisaki, "A model for synthesis of pitch contours of connected speech," *Annual Report, Engineering Research Institute, University of Tokyo*, vol. 28, pp. 53–60, 1969.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in hmm-based speech synthesis," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4238–4241.

[19] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for hmm-based speech synthesis." in *ICSLP*, vol. 98, 1998, pp. 29–31.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

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

[22] 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.

[23] T. Lambert, N. Braunschweiler, and S. Buchholz, "How (not) to select your voice corpus: random selection vs. phonologically balanced." in *SSW*, 2007, pp. 264–269.

[24] 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.

[25] 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.

[26] J. Hirschberg, "Pitch accent in context predicting intonational prominence from text," *Artificial Intelligence*, vol. 63, no. 1, pp. 305–340, 1993.

[27] 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.