Superpositional Modeling of Fundamental Frequency Contours for HMM-based Speech Synthesis

Keikichi Hirose¹, Hiroya Hashimoto², Daisuke Saito³, and Nobuaki Minematsu²

¹Professor Emeritus,  
²Department of Electrical Engineering and Information Systems,  
³Department of Information and Communication Engineering,  
The University of Tokyo  
{hirose, hiroya, dak_saito, mine}@gavo.t.u-tokyo.ac.jp

Abstract

Statistical parametric speech synthesis technologies, such as HMM-based and DNN-based ones, gain special attention from researchers because of their ability in generating speech in various voice qualities and styles. In these methods, all acoustic parameters (except durational ones) are handled in a frame-by-frame manner, which is not appropriate for prosodic features. Although relation of adjacent frames is viewed, it is not enough. Prosodic features are related to words, phrases, sentences, and even paragraphs, and should be viewed in a wider time span. One possible way to handle the features well in speech synthesis process is to model fundamental frequency ($F_0$) movements and to apply its constraints. Among several models of $F_0$ contours, the generation process model of $F_0$ contours is ideal for the purpose, since it can well represent hierarchical structure of prosody as superposition of phrase and accent components keeping a clear relationship with linguistic information. A method is developed which decomposes $F_0$ contours into three layers based on the model, and handles them as different streams in the HMM-based speech synthesis process. Advantage of the method is confirmed through objective and subjective evaluations. Issues of flexible control of prosody are also addressed.

Index Terms: generation process model, HMM-based speech synthesis, superpositional modeling, multi-stream, flexible control

1. Introduction

Synthetic speech close to human utterances is now available through concatenation-based speech synthesis. However, the method requires a large amount of speech corpus of the speaker and style to be synthesized. Ultimate goal of speech synthesis will be enabling to generate speech in any voice quality and style, which a user requires. This goal will be difficult to be realized only by concatenation-based speech synthesis since, to realize a new voice quality with a new style, it is necessary to prepare such speech corpus from the beginning. A scheme is necessary to control voice quality and speech style ideally without such speech corpus, or at least from a small amount.

From this viewpoint, statistical parametric speech synthesis technologies, viz. HMM-based speech synthesis [1] and recently speech synthesis with deep learning [2], gain a special interest from researchers, since it can generate synthetic speech with a rather high quality from a smaller-sized speech corpus, and can realize a flexible control in voice qualities and speech styles through statistical adaptation techniques. During speech synthesis process both segmental and prosodic features of speech are processed together in a frame-by-frame manner, which is appropriate for training acoustic models using a large amount of speech corpus. However, we should note that the frame-by-frame processing includes an inherent problem in handling prosodic features. Prosodic features are related to words, phrases, sentences, and even paragraphs, and should be viewed in a wide time span. Relations between frames are taken into account as $\Delta F_0$ and $\Delta^2 F_0$ features and/or by handling several frames at one process, but they are not enough. Generated speech often has over-smoothed fundamental frequency ($F_0$) contours with occasional $F_0$ undulations not observable in human speech. Moreover, relation of the generated $F_0$ contours with linguistic (and para-/non-linguistic) information conveyed by them is unclear, making further processing, such as to add emphasis, to change speaking styles, etc., not straightforward.

One possible way to cope with the situation will be to assume an $F_0$ contour model, and to introduce model constraint during the speech generation process. In the current paper, after viewing several models for $F_0$ contours, the generation process model ($F_0$ model, [3]) developed by Fujisaki and his co-workers is introduced to HMM-based speech synthesis. In order to fully take the benefit of the $F_0$ model, $F_0$ contours are decomposed into three layers, phrase, accent, and residual ones during the training and synthesis processes.

By handling $F_0$ contours in the framework of $F_0$ model, a “flexible” control of prosodic features comes possible. A corpus-based method has been developed to predict differences in the $F_0$ model commands between two versions of utterances of the same linguistic content [4]. Applying the predicted differences to the baseline version of speech, the new version of speech can be realized. A large speech corpus is not necessary to train the $F_0$ model command differences. The validity of the method has been shown through prosodic focus placing [5], speaking style conversion, and voice conversion [6].

The rest of the paper is organized as follows: After comparing several prosody models for speech synthesis with discussions on the required properties in section 2, the method of handling $F_0$ contours as three layers in HMM-based speech synthesis is introduced in section 3, followed by experimental results in section 4. In section 5, the issue of approximating generated $F_0$ contours by the $F_0$ model is viewed for realizing “flexible” control. Section 6 concludes the paper.

2. Modeling of $F_0$ contours

$F_0$ contours of sentences show quasi-continuous curves decoupled at unvoiced periods (and pauses). Although no $F_0$ is observable at unvoiced periods, a sentence $F_0$ contour is well interpreted as a fully continuous curve by interpolating unvoiced periods. It is generally admitted that an $F_0$ contour
consists of global and local movements, which may be related to phrases/clauses and tones/accents/stresses, respectively. Also, prosody has a hierarchical structure; from shorter time span covering a syllable/word to longer time span covering a phrase/sentence/paragraph. Models should well relate the prosodic structure with \( F_0 \) movements.

The well-known ToBI system [7] represents hierarchical structure of prosody as tone and break index tiers. However, it is a labeling scheme, and does not aim at parametric representations of \( F_0 \) contours. Several models are developed for the purpose, including Tilt [8] and PENTA [9]. However, most of them only try to trace observed \( F_0 \) movements and fail to decompose \( F_0 \) contours into their constituents with clear physical meanings. PENTA includes the concept of pitch target, which is useful to relate \( F_0 \) model parameters with linguistic information, such as word accents and syntactic structures. It assumes tilted targets, which maybe suitable to tonal languages but not for non-tonal languages. Phrase or longer time-span movements of \( F_0 \) are not clear in the model. There are several attempts to model \( F_0 \) contours as superposition of components representing gradual movements of longer time spans and sharp movements of shorter time spans [10, 11]. However, in many cases, an \( F_0 \) contour is decomposed simply as a smoothed \( F_0 \) contour and residuals. For instance, MOMEL uses spline functions for smoothing, and phrase-level and words/syllable-level \( F_0 \) movements are not well decomposed.

The \( F_0 \) model has two important features; super-positional and command-response ones. It describes \( F_0 \) contours in a logarithmic scale as the superposition of phrase and accent components, represented as responses of impulse-like and step-wise commands, respectively [3]. The model has a clear advantage in that both components are represented as responses to discrete commands, which have clear relations with linguistic information of the utterance. The response functions are those of second-order linear systems, which are common physical constraints when a system is controlled by inertia and damping.

Figure 1: An example for observed \( F_0 \)'s (in red dots) and their \( F_0 \) model approximation (in blue solid line). \( F_0 \) model parameters (accent and phrase commands) are also shown. ("terebigeemuya pasokoNde geemuoshite asobu": ((We) played games with TV gamers and/or personal computers.)

Figure 1 shows an example of \( F_0 \) contour approximation by the \( F_0 \) model. Although the approximated \( F_0 \) contour is close to that of the utterance, some discrepancies are observable. The \( F_0 \) model only takes phrase and accent components into account, and does not count micro-prosodic \( F_0 \) movements. Also, minor \( F_0 \) undulations without clear correspondences to linguistic information are ignored. Furthermore, \( F_0 \) contours may not strictly follow the (critically-damped) second-order linear systems, causing minor deviations from the model. Although these minor \( F_0 \) movements may not necessarily be included in the speech synthesis process from the speech quality viewpoint, it is worthwhile to develop a scheme to include them into the process. Here, we should note that pitch extraction processes may not always correct. Since the "erroneous" \( F_0 \)'s are not counted in the model, they are "unwillingly" counted as \( F_0 \) residuals. A scheme is necessary to exclude them from \( F_0 \) residuals, though the issue is not addressed in the current paper.

One major drawback of the \( F_0 \) model is that extraction of model parameters from observed \( F_0 \) contours requires a recursive process. Therefore, assignment of initial values is crucial for the performance. Although several methods have already been developed to automatically estimate initial values, their performance is not satisfactory [12, 13]. This is because they first smooth and interpolate \( F_0 \) contours and then take derivatives to obtain the initial values without taking linguistic information of the utterance into account. The process is not robust for pitch extraction errors, and produce erroneous commands not corresponding to the linguistic information of the utterances. To solve this situation, we recently have developed a method, which extracts phrase and accent components viewing the \( F_0 \) contours as mora-based high-low patterns [14]. The method takes features of Japanese prosody into account, and follows to the manual process of extracting model parameters by an expert.

3. \( F_0 \) model and HMM-based speech synthesis

A corpus-based method was already developed synthesizing \( F_0 \) contours in the framework of \( F_0 \) model, and was combined with HMM-based speech synthesis. Speech synthesis in reading and dialogue styles with various emotions was realized [15]. However, the method simply substitutes \( F_0 \) contours generated by HMM-based speech synthesis to those generated by the model. Although, a better quality of synthetic speech is obtainable, independent control of segmental and prosodic features violates maximum likelihood criterion of HMM-based speech synthesis.

Introducing the \( F_0 \) model constraint directly in HMM-based speech synthesis is not easy, since the \( F_0 \) model commands cannot be well represented in a frame-by-frame manner. An effort was reported to represent the \( F_0 \) model in a statistical framework to cope with the problem, but was not combined with HMM-based speech synthesis [16]. We developed a simple way; to approximate \( F_0 \) contours of speech with the \( F_0 \) model, and to use these \( F_0 \)'s for HMM training [14].

As mentioned already, one of major advantages of the \( F_0 \) model is that it can well decompose an observed \( F_0 \) contour into phrase and accent components. Phrase components represent gradual \( F_0 \) declination corresponding to phrases, while accent ones represent local \( F_0 \) humps corresponding to word accents. Since they are related to linguistic information of the utterance differently, a better control of prosody is expected by handling them separately. We already have realized this idea by predicting \( F_0 \) model commands; first phrase commands and then accent commands taking the predicted phrase commands into consideration [15]. Y. C. Huang et al developed a similar method for Chinese [17]. C. C. Hsia et al applied another hierarchical modeling of prosodic units to generate global \( F_0 \) movements and combined them with frame-by-frame \( F_0 \)'s generated by HMM-based speech synthesis [18]. They introduced syllable level \( F_0 \) layer, which is considered to be suitable for Chinese (but not for non-tonal languages). The modeling is based on approximating global \( F_0 \) movements with Legendre polynomials, which cannot
represent phrase component well. These methods generate global F0 movements outside HMM-based speech synthesis processes. From these considerations, we have developed a method to decompose F0 contours into three layers by the F0 model, and to handle each of them as different stream in the training and synthesis processes of HMM-based speech synthesis. A preliminary attempt has already been made to decompose F0 contours into F0 model components, and generate each component individually by the HMM-based speech synthesis framework [19]. However, in their work, phoneme durations were fixed to those of target utterances, and quality of synthetic speech was not assessed. Furthermore, detailed analyses on the results, such as on the context clustering of each component, are not provided.

In our method, an observed F0 value of each frame is represented as:

\[
\ln F_{\text{obs}}(t) = \ln F_{\text{phrase}}(t) + \ln F_{\text{accent}}(t) + \ln F_{\text{residual}}(t), \quad (1)
\]

where \( F_{\text{obs}}(t) \), \( F_{\text{phrase}}(t) \), \( F_{\text{accent}}(t) \), and \( F_{\text{residual}}(t) \) denote observed F0, phrase component F0, accent component F0, and F0 residual at frame \( t \), respectively. “ln” indicates to take (natural) logarithmic values. Since these three components are related to linguistic information of utterances differently, they are better to be handled as different streams in the HMM-based speech synthesis. Contexts are expected to be clustered differently for each component. We also tested a scheme to represent phrase and accent component F0’s of each frame as a two-dimensional vector, since these two components are closely related.

One issue of HMM-based speech synthesis is how to handle voiceless phoneme periods, where F0 values are unavailable. Although multi-space probability distribution HMM (MSD-HMM) is commonly used [20], it is pointed out that MSD-HMM has a limitation in representing F0 movements around voiced/voiceless boundaries. When using the F0 model, since continuous F0’s are obtainable, continuous F0 HMM [21] comes an attractive alternative.

4. Experiments

Speech synthesis experiments are conducted using ATR continuous speech corpus of 503 sentences uttered by male speaker MHT [22]. Out of 503 sentences, 450 sentences are used for HMM training, keeping 53 sentences for evaluation. Speech signals are sampled at 16 kHz sampling rate, and STRAIGHT analysis [23] is used to extract the spectral envelope, which is converted to mel-cepstral coefficients (0th to 40th) using a recursion formula, with 5-ms frame shift. F0 and 5 band-aperiodicity (0–1 kHz, 1–2 kHz, 2–4 kHz, 4–6 kHz, 6–8 kHz) are also extracted. These features together with their \( \Delta \) and \( \Delta^2 \) consist HMM feature vector. Five-state left-to-right hidden semi-Markov model with single Gaussian distribution for each state, provided in HTS-2.1 [24], is used. A Gaussian distribution is represented by a diagonal covariance matrix. Decision tree-based context clustering is conducted with MDL stop criterion.

The following four versions of speech are synthesized.

1) Original: HMM-based speech synthesis trained using extracted F0’s of training corpus.

2) F0 model: F0 is handled as two streams consisting of F0 model-based F0 and F0 residual.

3) Multi-stream: F0 is handled as three streams consisting of phrase component, accent component and F0 residual.

4) Vector: Phrase and accent components are represented as a two-dimensional vector. F0 residual is handled as a separate stream.

While MSD-HMM is used for version 1), continuous F0 HMM is used for other versions with an extra-stream of voiced/unvoiced labels. F0 residual are assumed to be 0 for voiceless frames.

Figure 2 shows F0 contours of versions 1) -3) as compared to the F0 contour of the target utterance. As a reference, F0 contour without F0 residuals is also shown for “Multi-stream.” The accent component around “muhoo” is generated well by the proposed method. The sharp dip around /h/ is due to F0 residuals (see panel (e) without F0 residuals). This dip is considered to be due to erroneous F0’s of the training data. As for this specific example, it does not affect the speech quality so much, since /h/ is synthesized as voiceless.

Generated F0 contours by the four methods are evaluated through \( F_0 \text{RMSE} \), which is defined by the following equation:

\[
F_0 \text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (\ln F_{\text{argmax}}(t) - \ln F_{\text{original}}(t))^2}, \quad (2)
\]

where \( F_{\text{argmax}}(t) \) and \( F_{\text{original}}(t) \) are F0 of target utterance and generated F0 by HMM-based speech synthesis at frame \( t \), respectively. Before calculating \( F_0 \text{RMSE} \), generated F0 contours are time-aligned to target F0 contours by DP matching. Summation is conducted only for frames judged as voiced in both F0 streams. \( F_0 \text{RMSE} \)’s averaged over 53 test sentences are 0.2102, 0.2185, 0.1590, 0.1573 for “Original,” “F0 model,” “Multi-stream,” and “Vector,” respectively. Around 25% reductions are obtained for “Multi-stream” and “Vector” as compared to “Original.” \( F_0 \text{RMSE} \) of “F0 model” is slightly larger than “Original.” Although a further analysis is necessary, this is considered to be due to time mismatch between F0 model-based F0’s and F0 residuals, since in our former experiment ignoring F0 residuals [14] the result was better than “Original.” The mismatch is reduced by handling phrase and accent parts of F0 in “Multi-stream/Vector.” A version of representing phrase, accent, and residual as a vector may further reduce \( F_0 \text{RMSE} \).

A listening test of synthetic speech was conducted for “Original,” “F0 model,” and “Multi-stream” versions involving 11 native speakers of Japanese. Ten sentences are selected randomly out of 53 test sentences, and naturalness of their synthetic speech by the three versions was evaluated through the five-scale scoring: 5: natural, 4: moderately natural, 3: neutral, 2: rather unnatural, and 1: unnatural. The averaged scores with 95 % confidence intervals are summarized in Table 1. The best score is obtained for “Multi-stream.” The score is worst for “F0 model.” It may be due to mismatch between F0 model-based F0’s and F0 residuals, as explained already.

Benefit of handling F0 contours as the three streams is clear from the result of context clustering. Questions regarding to longer time spans such as breath group length and sentence length are selected for phrase component F0’s, while questions regarding to (accent types of) accent phrases are selected for accent component F0’s. Questions on phoneme identities are
selected for $F_0$ residuals. These results coincide with our expectation.

![Figure 2: $F_0$ contours generated by the four versions of HMM-based speech synthesis, as compared to $F_0$ contour of the target utterance: (a) target, (b) Original, (c) $F_0$ model, and (d) Multi-stream. Panel (e) is $F_0$ contour by Multi-stream without $F_0$ residuals. (“teNimuhoo oorakana monoda”: (He) is such a flawless, natural and generous (person).) (Images not shown.)](image)

![Figure 3: An example for generated $F_0$'s (in red dots) by $F_0$ model-based HMMs, and their $F_0$ model approximation (in blue solid line). $F_0$ model parameters (accent and phrase commands) are also shown. (“ameno tameka yachooga muragatte kiseio ageteita”: (Presumably due to raining, wild birds gathered with strange crying sounds.)](image)

Table 1. Scores of listening test with 95% confidence intervals for the four versions of synthesized speech.

<table>
<thead>
<tr>
<th>Method</th>
<th>Score with 95% conf. int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>3.209±0.206</td>
</tr>
<tr>
<td>$F_0$ model</td>
<td>2.309±0.199</td>
</tr>
<tr>
<td>Multi-stream</td>
<td>3.254±0.226</td>
</tr>
</tbody>
</table>

5. Discussion

By the proposed method, generated $F_0$ contours are represented as the sum of three contours; two contours generated from HMM’s trained using phrase and accent components of the $F_0$ model, and one contour generated from HMM’s trained using $F_0$ residuals. Extraction of $F_0$ model commands is considered to be easy for the former two contours, leading to a flexible and systematic control of prosody as mentioned already. Figure 3 shows an example of $F_0$ model command extraction for an $F_0$ contour generated by the proposed method. ($F_0$ residuals are not included in the $F_0$ contour for better visibility.) It is clear from the figure that the generated contour could be well represented by the $F_0$ model.

Two versions are tested for the proposed method; one is to represent phrase and accent component $F_0$ streams as separate ones and the other is to represent them as a vector sequence. The phrase and accent components have tight relations, and therefore the latter version may have a benefit. However, no clear difference is observed between these two versions. In our former experiments to predict $F_0$ model commands from input sentences, phrase commands are first predicted, and then accent commands are predicted taking into account the predicted phrase commands [15]. This process need to be realized in the HMM-based speech synthesis.

6. Conclusion

A method is developed to represent $F_0$ contours as three layers based on the $F_0$ model, and to use them for HMM-based speech synthesis. Continuous $F_0$ HMM is adopted instead of MSD-HMM. Listening test of synthetic speech indicates that the method can generate better quality than the original HMM-based speech synthesis, which handles $F_0$'s as they are, though the improvement is not significant. As for the objective evaluation, a clear reduction in $F_0$RMSE is realized by the proposed method.

One of major advantages of adding $F_0$ model constraint during HMM-based speech synthesis is that resulting $F_0$ contours are easily analyzed in the $F_0$ model framework, and, therefore, a clear relationship is obtainable between the $F_0$ contours and linguistic information. This enables additional manipulation of $F_0$ contours [4-6].

This work is partly supported by Grant-in-Aid for Scientific Research (B) #24300068, JSPS, and the Major Program for the National Social Science Fund of China (13&ZD189).
7. References


