Speech prosody in phonetics and technology
Keikichi Hirose
National Institute of Informatics/University of Tokyo, Japan,
hirose@gavo.t.u-tokyo.ac.jp

Abstract
As features unique to spoken language, speech prosody plays an important role in human communication. Although the acoustic features of speech are viewed most frequently in a frame-by-frame manner, this is not always appropriate for prosodic features, since they are tightly related to higher level linguistic information, such as syntactic and discourse structures, and spread to wide time spans, such as syllables, words, and phrases. In order to handle the situation, models for prosody have been developed. Among many models, the generation process model of fundamental frequency contours is attractive, since it can relate well to the linguistic information of utterances. The model was successfully applied to hidden Markov model (HMM) based speech synthesis and a listening test to determine the (perceptual) categorical boundaries of Japanese accent types.

Keywords: Speech prosody, Fundamental frequency, Generation process model, Speech synthesis, Japanese word accent

1. Introduction
Prosody of speech plays an important role in speech communication. It is well known that words with the same phonetic constitution are distinguished through prosody, such as word accents in the Japanese and syllable tone types in the Chinese language. Although phonetic constitution plays a major role in the transmission of meanings of utterances, the role of prosody becomes important for higher levels of linguistic information, such as syntactic structures and discourse structures. During conversations, humans often express various intentions and emotions. Such para-/non-linguistic information is conveyed mostly through prosody. While most information transmitted by segmental features of speech can be included in the transcribed text, information transmitted by prosodic features cannot. Therefore, prosody is a very important issue in the study of speech phonetics and technology. However, studies of prosody have been conducted rather non-systematically.

One of the major difficulties in handling prosody is that prosodic features cover wide time spans, such as words, phrases, sentences, and even paragraphs. Current speech technologies, in particular statistical ones, such as those based on hidden Markov models (HMMs) and neural networks, rely mainly on frame-by-frame processing without a sufficient view of the relation between frames in wide time spans. To handle this situation, models of prosody have become indispensable. Among the various models that attempt to represent the global movements of fundamental frequencies ($F_0$'s), the generation process model is widely known among researchers, because of its super-positional and command response features [1, 2]. Prosody has a hierarchical structure covering features from those of shorter time spans, such as syllable and words, to those of longer time spans, such as phrases and sentences. This structure is well modeled by super-posing local undulations corresponding to word accents onto gradual decays corresponding to phrases. Therefore, the model has become a strong tool in speech science and technology research. The model may increase the naturalness of synthetic speech and allow systematic control of $F_0$'s in speech synthesis, thus offering a good tool for phonetic research, generating speech samples for various perceptual experiments.

In this paper, after briefly surveying the models of $F_0$'s, the generation process model is introduced. Some of our contributions using the model for HMM-based speech synthesis are then presented [3, 4]. To verify the usability of the model for perceptual experiments, we conducted several studies, including those to determine the perceptual categorical boundaries of Japanese word accent types [5, 6]. Based on the results, systems teaching Japanese word accent type pronunciation were developed [5, 7].

2. Prosody modeling

2.1. Modeling fundamental frequency contours
$F_0$ contours of sentences appear as piecewise curves decoupled at unvoiced periods. Although no $F_0$ is observable at unvoiced periods, a sentence $F_0$ contour is well interpreted as a fully continuous curve with unvoiced periods interpolated. It is in general agreed that an $F_0$ contour consists of global and local movements, which may be related to phrasing and accentuation, respectively. In addition, prosody has a hierarchical structure, from a shorter
time span covering a syllable/word to a longer time span covering a phrase/sentence/paragraph. Models should relate well with the prosodic structure with $F_0$ movements.

The well-known ToBI system [8] counts the 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 were developed for the purpose, including Tilt [9] and PENTA [10]. However, most of these attempt only to trace observed $F_0$ movements and fail to decompose $F_0$ contours into their constituents retaining clear physical meanings. Several attempts have been made to model $F_0$ contours as the super-position of components representing the gradual movements of longer time spans and the sharp movements of shorter time spans [11, 12]. However, in many cases, an $F_0$ contour is decomposed simply as a smoothed $F_0$ contour and residuals. For instance, MOMEL uses spline curves for smoothing. In this model, phrase-level and word/syllable-level $F_0$ movements are not well decomposed. Focusing on the multi-scale feature of wavelet transform, continuous wavelet transform is used to represent the hierarchical structure of prosody [13]. However, the relation with the decomposed components and linguistic information of the utterance is still not sufficiently clear.

### 2.2. Generation process model

The generation process model of $F_0$ contours ($F_0$ model), frequently called Fujisaki’s model, has two important features for $F_0$ modeling: super-positional and command-response. It describes $F_0$ contours in a logarithmic scale as the super-position of phrase and accent components, represented as responses to impulse-like and step-wise commands, respectively [1, 2]. The model has a clear advantage in that both components are represented as responses to discrete commands, which have clear relations with the linguistic information of the utterance. The response functions are those of critically-damped second-order linear systems, which are common physical constraints.

![Figure 1](image.png)

**Figure 1:** Example of observed $F_0$'s (red dots) and their $F_0$ model approximation (blue solid line). The $F_0$ model parameters (accent and phrase commands) are also shown.

Figure 1 shows the results of the approximation of an observed $F_0$ contour by the $F_0$ model. Although the approximated $F_0$ contour is close to the observed one, some discrepancies can be seen. The $F_0$ model takes only phrase and accent components into account, and does not count micro-prosodic $F_0$ movements. In addition, 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. These minor $F_0$ movements are related mainly to phonemes.

### 3. HMM-based speech synthesis

A corpus-based method was already developed for synthesizing $F_0$ contours in the framework of the $F_0$ model, and was combined with HMM-based speech synthesis. Thus, speech synthesis in reading and dialogue styles with various emotions was realized [14]. However, the method simply substitutes $F_0$ contours generated by HMM-based speech synthesis with those generated by the model. Although, a better quality of synthetic speech is obtainable, independent control of segmental and prosodic features violates the maximum likelihood criterion of HMM-based speech synthesis.

It is not easy to introduce the $F_0$ model constraints directly into HMM-based speech synthesis, since the $F_0$ model commands cannot be well represented in a frame-by-frame manner. We developed a simple technique: we approximate $F_0$ contours of speech with the $F_0$ model, and use these $F_0$'s for HMM training [15].

As mentioned previously, one of the major advantages of the $F_0$ model is that it can effectively decompose an observed $F_0$ contour into phrase and accent components. Phrase components represent a gradual $F_0$ decline corresponding to phrasing, while accent ones represent local $F_0$ humps corresponding to word accents. Since they are related differently to the linguistic information of the utterance, a better control of prosody is expected to be achieved by handling them separately. We have already realized this idea by predicting $F_0$ model commands, first phrase commands and then accent commands, taking the predicted phrase commands into consideration [14]. Huang et al. developed a similar method for the Chinese language [16]. 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 [17]. They introduced a syllable level $F_0$ layer, which is considered to be suitable for the Chinese language. The modeling is based on approximating global $F_0$ movements with Legendre polynomials, which cannot represent phrase...
components well. These methods generate global $F_0$ movements outside HMM-based speech synthesis processes. $F_0$ contours were decomposed into several layers by arranging level-dependent questions for the context clustering in HMM-based speech synthesis [18, 19]. However, the studies were aimed to realize better $F_0$ control in HMM-based speech synthesis, and the resulting decomposition results were not analyzed well in relation to the linguistic information.

Taking the above into consideration, a method was developed to decompose $F_0$ contours into three layers by using the $F_0$ model, and to handle each layer as different streams in the training and synthesis processes of HMM-based speech synthesis [4]. The three layers comprise phrase component $F_0$, accent component $F_0$, and the $F_0$ residual. All $F_0$ values are handled in logarithmic scale. Since the three layers are related differently to the linguistic information of utterances, the contexts are expected to be clustered differently for each layer. In fact, questions regarding longer time spans, such as breath group length and sentence length, are selected for phrase component $F_0$’s (Fig. 2), while questions regarding (accent types of) accent phrases are selected for accent component $F_0$’s. Questions on phoneme identities are selected for $F_0$ residuals. The advantages of the method as compared to ordinary HMM-based speech synthesis were shown through objective and subjective evaluations.

One of the major issues of HMM-based speech synthesis is the handling of voiceless phoneme periods, where $F_0$ values are unavailable. Although multi-space probability distribution HMM (MSD-HMM) is commonly used [20], it has been noted that it is limited in terms of representing $F_0$ movements around voiced/voiceless boundaries. When using the $F_0$ model, since continuous $F_0$’s are obtainable, continuous $F_0$ HMM [21] is an attractive alternative.

In the proposed method, generated $F_0$ contours are represented as the sum of three contours, two of which are generated from HMM’s trained using the phrase and accent components of the $F_0$ model, and one from HMM’s trained using $F_0$ residuals. The 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 [22, 23].

4. Computer-aided pronunciation learning

The recent development of internationalization has increased to a great extent the number of situations where a person should speak in his/her non-mother tongue. For many foreigners, it is not an easy task to reach an affordable level of making conversations with native speakers. When a person is looking for a job in a foreign country, the situation is serious; his/her mental ability is sometimes ranked at a lower level because of his/her accented pronunciation. Therefore, the need for foreign language learning is increasing. The ideal situation is that sufficient “good” human teachers exist, but this is not the case. Supported by the developments of information technology, a number of computer-aided language (pronunciation) learning systems are now available. By virtue of advanced speech processing technologies, some systems score learners’ proficiency level. However, a low score does not necessarily indicate the learner’s low proficiency; systems may even judge a native speaker’s utterance as an accented pronunciation if his/her voice quality differs widely from that of the reference speech. This is because the scoring is based on the distances of the acoustic features of native speakers’ utterances (or the likelihood to correspond to native speakers’ acoustic models), and is not based on the criterion of native speakers’ perception. In many cases, it is sufficient for learners if their proficiency reaches a level enabling smooth conversation with native speakers. Systems for pronunciation learning should indicate whether the learner’s pronunciation is acceptable to native speakers or not.

The above function can be realized through experiments involving listening to synthetic speech where the acoustic parameters in question are systematically changed. Although such a listening test is unrealistic for analyzing phone quality, since it is related to a number of parameters, it becomes possible if we focus on word accents. In the Japanese language, words with the same phoneme constitution have different meanings according to the accent type. In the framework of the generation process model, accent types are represented as the positions of accent commands with respect to mora boundaries. It is possible to interpolate two accent types by manipulating accent commands. Figure 3
shows an example of a listening experiment using the synthetic speech samples thus obtained. In this figure, the perceptual categorical boundaries between a low-high-high type accent and a high-low-low type accent can be seen [5]. Foreigners may often speak using $F_0$ contours that cannot be categorized into any possible accent types of the Japanese language. The boundary between acceptable/unacceptable pronunciation can also be obtained through listening tests using synthetic speech where the $F_0$ contours are controlled appropriately [6]. Based on the results of these experiments, a system for accent type pronunciation training can show the correctness (in % scores) of a learner’s pronunciation as perceived by native speakers [5, 7].

![Figure 3: Perceptual boundary of accent types of three-mora words using synthetic speech.](image)

### 5. Conclusion

Appropriate modeling is crucial for showing the relation between the linguistic information of utterances and prosodic features. The generation process model for $F_0$ contours has attracted research attention, since it has two important features of prosody, a super-positional (hierarchical) and command response. The relation to linguistic information becomes clear when viewing $F_0$ contours as model commands. The model is useful for a wide range of speech research: speech analysis, synthesis, and recognition. In this paper, our studies on HMM-based speech synthesis and computer-aided language learning were presented. Currently, deep learning techniques are “invading” the speech research area. Supported by a huge corpus, good results are being obtained for speech recognition and synthesis. However, prosodic features are not effectively addressed. Because of their hierarchical and long-time span nature, they cannot be easily handled using only statistical schemes. We should not forget to examine their physical background.

### 6. References


