

# APPLYING A HYBRID INTONATION MODEL TO A SEAMLESS SPEECH SYNTHESIZER

*Takashi Saito, Masaharu Sakamoto*

IBM Research, Tokyo Research Laboratory, IBM Japan Ltd.  
1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502, Japan  
saito@jp.ibm.com, sakamoto@jp.ibm.com

## ABSTRACT

We present a speech synthesizer to seamlessly concatenate recorded and synthetic phrases to produce natural sounding and highly expressive speech. Not only the acoustic units, but also the F0 contours are seamlessly concatenated together from recorded and synthetic phrases. When mixed with recorded phrases, the F0 contours of synthetic phrases are generated *adaptively* relative to the actual surrounding F0 shapes of the recorded phrases. Although the intonation generation scheme was originally developed for unlimited speech synthesis, it is quite naturally extended to a hybrid intonation generation.

## 1. INTRODUCTION

Recent remarkable advances in text-to-speech methodologies such as unit-selection-based techniques and signal processing algorithms have greatly contributed to improving the overall speech quality of text-to-speech systems. In spite of the improvements, the fact remains that the naturalness and expressiveness of synthetic speech is, in general, far from recorded speech. Therefore, pre-recorded speech prompts are still used in various speech applications, particularly when their speech output messages can be covered with relatively limited vocabularies, such as voice response systems via telephone, car navigation systems, speech dialogue systems for robots, etc.

At the same time, even in such applications, text-to-speech is commonly used to support variable phrases or out-of-domain words. Simple concatenations of synthetic and recorded speech do not, however, provide the best solution in terms of the overall speech quality. Synthetic speech when mixed with recorded speech is sometimes heard as being even worse than when used independently, because the synthetic speech is directly compared with its ideal version, the recorded speech. Therefore, some strategies are needed to maximize the overall speech quality when mixing synthetic and recorded speech.

Given this background, there have been several studies [1,2,3] trying to optimize the use of recorded speech in a generic text-to-speech framework in order to improve the naturalness of the overall speech quality. In the previous studies, even combined with text-to-speech technology, the improvement foci are still limited mainly to fixed phrases or sentences, which correspond to the recorded speech portions. In other words, the synthetic speech portions are still treated as almost independent and standard text-to-speech sections, and the effect of interactions

with the recorded speech is not seriously considered, particularly as it effects intonation. However, the actual intonation patterns of individual speech utterances are affected by degrees of emphasis or emotion and by conveyed intentions of the speaker, as well as by the fluctuations of utterances themselves. Considering these various behaviors observed in actual utterances, the fixed intonation of synthetic speech predicted using only linguistic factors might not always be appropriate for combination with really expressive recorded speech.

In this paper, we describe a speech synthesis system, *Seamless Speech Synthesizer*, which produces natural sounding and highly expressive speech by seamlessly concatenating synthetic phrases and recorded ones. In order to realize a flexible connection of the two different kinds of speech, it is important to bridge the two in terms of not only acoustic features, but also prosodic features. Hence, we devised a method of generating a hybrid intonation for the seamless speech synthesizer to generate appropriate intonation for a sentence composed of synthetic and recorded phrases. The intonation of the synthetic phrases is not determined only from the linguistic context, but is determined depending on the actual surroundings of the recorded phrases.

The following sections are organized as follows. Section 2 describes the outline of the seamless speech synthesizer. In Section 3, we describe the proposed method of generating a hybrid intonation, and some experiments using actual speech databases are provided in Section 4. A summary and future plans are given in the last section.

## 2. SEAMLESS SPEECH SYNTHESIZER

Figure 1 shows the outline of the seamless speech synthesizer, which is designed to handle recorded phrases as an extension of a baseline generic text-to-speech system. The baseline system is a corpus-based speech synthesizer that can generate highly speaker-dependent voices by using both acoustic and prosodic features of a donor speaker. The donor speaker's speech features are extracted from a speech corpus for use in speech synthesis, and prepared as a common (domain-unlimited) speech database, a *VoiceFont* [5], to be used in the three parts (phoneme duration prediction, F0 contour generation, and acoustic unit selection) of the baseline system.

Recorded phrases or sentences, frequently used for a specific application, are collected together in a domain-specific speech database. There are normally several hundred phrases, although the number is dependent on the applications used and there is no apparent limitation on the size. The domain database is defined as an extended user dictionary, which includes not only the ordinary lexical information for each phrase, but also the acoustic data (waveform) and prosodic data (duration and F0 values of the phonetic segments) of the utterances just like the features in the VoiceFont.

As in a previous study [3], the two databases are not necessarily separated and can be handled with the same unit selection algorithm, which takes into account both phonetic and prosodic context. The current implementation here is to handle the two separately except for the acoustic unit selection of the joint portions of recorded and synthetic phrases. This is primarily because the domain database is assumed to include not only normal phrases, but also various expressive ones, which might have fairly different acoustic and prosodic features from normal ones. By appropriately using the two databases, the speech synthesizer is able to seamlessly mix even expressive utterances like emotive or emphatic ones with synthetic speech in the generic text-to-speech framework.

If a recorded phrase registered in the domain database is found in the input text at the linguistic analysis stage, the prosodic and acoustic units for the phrase are read from the database, and are used instead of those prepared for a purely synthetic phrase.

The point here is how to build up a natural hybrid sentence from the two kinds of phrases. The most important thing is to realize a natural and smooth connection between them in both acoustic and prosodic senses. For the acoustic connections, a seamless transition between the two kinds of phrases can be obtained quite simply by applying the ordinary unit selection used for synthetic phrases to both ends of the recorded phrases, but not using the real acoustic unit data. For the prosodic connections, we used a slightly more sophisticated method to obtain a natural and smooth intonation in such hybrid sentences. This method is described in the next section.

### 3. HYBRID INTONATION MODEL

We present here a method of generating *hybrid* intonation for a sentence composed of recorded and synthetic phrases. For simplification, as shown in Figure 2, we assume here a typical case where a synthetic phrase is to be inserted between recorded phrases. Considering the intrinsic variations of F0 contours in recorded phrases as mentioned in the introduction, one fixed F0 contour of a synthetic phrase predicted only from linguistic factors is not always appropriate for insertion between the recorded phrases. Therefore, it is quite reasonable to take into account the actual F0 shape information of the surrounding recorded phrases in predicting the F0 contour of the synthetic phrase.

#### 3.1. Basic Framework of Generating F0 Contours

We describe here the basic framework of generating F0 contours used in this synthesizer, which was initially presented in [6]. In this approach, an F0 contour of a sentence is generated by concatenating the natural F0 shapes, from left to right, as optimally selected from the F0 shape inventory. An F0 shape unit is defined as a real F0 contour fragment segmented from natural F0 contours. The unit corresponds approximately to a segment of a minor prosodic phrase. The generation method consists of three main stages, (1) F0 shape target prediction, (2) F0 shape selection, (3) F0 shape concatenation. The details of the three stages are given below.

##### (1) F0 Shape Target Prediction

First, the F0 shape target of a minor phrase segment is predicted by using linguistic information (accent type, phrase length in morae, phoneme class, etc.) of the current and neighboring phrases, and the prosodic levels of the preceding and following phrase boundaries (i.e., sentence start position or end position, pause inserted or not inserted). The prosodic information on phrase boundaries is determined prior to the F0 contour generation by using the local phrase structures of the sentence.

Three parameters that describe an F0 shape target are defined: (1) the top absolute level (TopAbsLvl), (2) the starting-edge offset level relative to the top level (SttRelLvl), and (3) the ending-edge offset level relative to the top level (EndRelLvl). The three-parameter expression for a minor phrase was initially used by Sagisaka in [4] as a final representation of an F0 contour. We regard it as an intermediate skeleton of an F0 contour to be used as a target value in selecting an optimal unit from the natural F0 shape inventory.

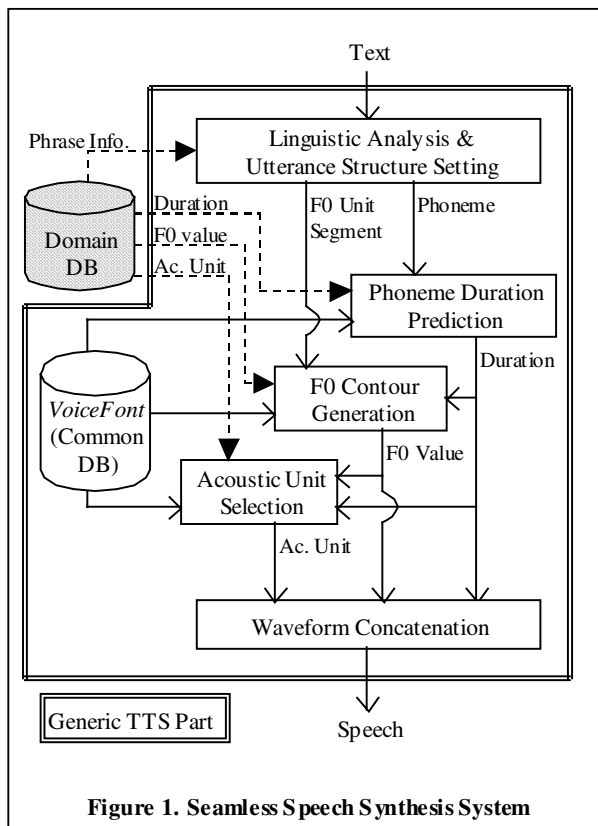


Figure 1. Seamless Speech Synthesis System

We applied a linear regression model, which is based on the well-known quantification theory type-I [7], to predict these three values.

## (2) F0 Shape Selection

In the F0 shape selection stage, the most suitable F0 shape unit for a given target synthesis environment is selected from the F0 shape inventory. The F0 shape selection applied here consists of two steps, prescoring of possible candidates and phonemic matching. In the first prescoring step, all the F0 shape candidates with accent attributes equal to the input are reordered according to their distance from the target F0 shape. The distance is defined as a Euclidean distance of two-dimensional vectors (SttRelLvl, EndRelLvl) in the log domain. Candidates below a predefined threshold are passed to the final step, phonemic matching. The phonemic matching is designed to select the best match from the prescored candidates, based on minimizing the phonemic class difference from the given phrase.

## (3) F0 Shape Concatenation

In the final stage of the F0 contour generation, the selected F0 shapes are concatenated with one another, basically without shape modification. Two kinds of minimal adjustments are applied to the F0 shape units in order to realize smooth connections: one is an F0 level shift, and the other is a time axis modification. The F0 level shift is carried out so as to smooth the levels between concatenated F0 shapes based on the TopAbsLvl parameter obtained in the first stage. The time axis modification is to adjust the duration of each mora for the target phoneme sequence.

## 3.2. Extension to Hybrid F0 Contours

Even when generating the F0 contour of a synthetic phrase concatenated with recorded phrases, the generation procedure is almost the same as that for the purely synthetic phrase sequences described above. The only extension from the basic procedure is in the first stage to predict the F0 shape target

parameters. As illustrated in Figure 2, the real values of the three parameters of the preceding and following recorded phrases are used as additional factors to predict the F0 shape target of the synthetic phrase.

## 4. EXPERIMENTS

In this section, we describe experiments to evaluate the F0 contour generation method applied to the mixture of synthetic and recorded phrases. First, we investigate and discuss the basic performance of the hybrid F0 contour prediction by speech recorded from a normal reading. We also consider some examples of generated hybrid F0 contours seamlessly combined with very expressive recorded phrases.

### 4.1. F0 Shape Target Prediction

#### Speech material

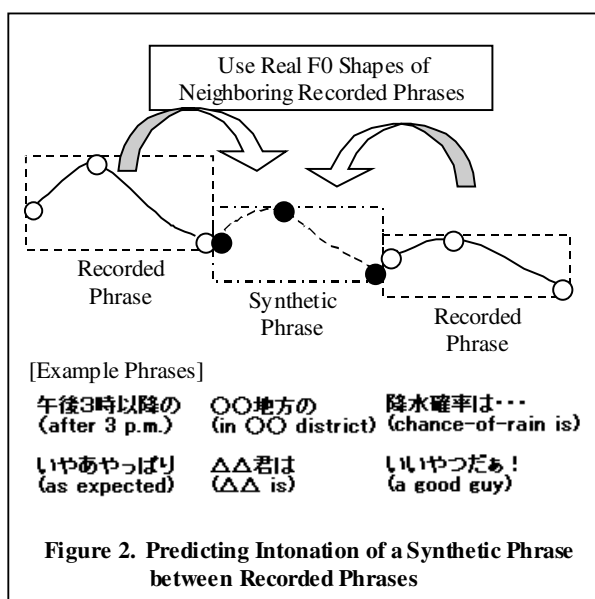
We used a text set of 503 sentences from the ATR continuous speech database, and prepared a speech corpus for text spoken by a professional female speaker. We used 450 sentences of the corpus for training (model parameter estimation) and for building up the F0 shape inventory, and the rest (53 sentences) were used for cross validation. All the acoustic and linguistic features contained in the F0 shape inventory were automatically extracted by an in-house extraction tool, and the extraction was followed by a stage of manual check and correction.

#### Results

Table 1 shows the results of predicting the three parameters of an F0 shape target: top absolute level (TopAbsLvl), start relative level (SttRelLvl), and end relative level (EndRelLvl). First, we can see that the use of surrounding real F0 shapes obviously decreases the prediction root mean square (RMS) errors, consistently for both closed and open data.

The effect on predicting EndRelLvl of using real F0 shapes seems to be smaller than other parameters. One of the reasons for this is that the prediction of EndRelLvl itself is not yet stable, as is seen from the difference in error between open and closed data. Further investigation would be necessary to improve this prediction. On the other hand, for the other two parameters the predictions are more consistent, especially in when using real F0 shapes.

These results show that the use of surrounding real F0 shapes improves the predictions of F0 shape targets for synthetic phrases.



	Use Real F0 Shapes of Rec. Phrases	RMS Error (Closed Data) [oct]	RMS Error (Open Data) [oct]
TopAbsLvl	No	0.146	0.147
	Yes	0.138	0.139
SttRelLvl	No	0.146	0.150
	Yes	0.140	0.141
EndRelLvl	No	0.145	0.154
	Yes	0.144	0.153

Table 1. Results of Target F0 Shape Parameter Prediction

## 4.2. Hybrid Generation Applied to Expressive Speech

In Table 1, we can see the effect of using real F0 shapes, but the degree of decrease in RMS errors is not surprisingly big. This is primarily because the speech utterances used here were read consistently in a normal style by a professional speaker. In more expressive speaking styles, on the other hand, we could expect the effect of using real F0 shapes to be more noticeable since the F0 movements are much more dynamic. In order to confirm this idea, we built a speech inventory set (VoiceFont, and domain database of recorded phrases) for the seamless speech synthesizer, aiming at synthesizing the voice of an animated character. The two speech databases were both produced by a professional female speaker using the voice she used for a small boy in an animated movie. The general characteristics of the speaking style are fairly varied and dynamic. For instance, it has a very wide range of F0 contour movement, varying from 130 to 600 Hz. Figure 3 shows an example of seamlessly mixed phrases. The underlined phrases are purely synthetic ones. The F0 contour of the upper one is generated without the surrounding real F0 shape information, and that of the lower one is generated with the real F0 shape information. It is obvious that the phrasal F0 contour of the lower one has a natural and smooth transition with the surrounding recorded phrases. We also confirmed these distinct differences by listening.

## 5. CONCLUSIONS

We have described a speech synthesizer to seamlessly concatenate recorded phrases and synthetic phrases to produce natural sounding and highly expressive speech. Not only the acoustic units, but also the F0 contours are seamlessly

concatenated together from recorded and synthetic phrases. When mixed with recorded phrases, the F0 contours of the synthetic phrases are generated adaptively according to the surrounding real F0 shapes of the recorded phrases. Although the F0 generation scheme was originally developed for unlimited speech synthesis, it is quite naturally extended to hybrid F0 contour generation. In this research, we have only studied generating synthetic F0 contours adaptively considering the neighboring recorded phrases. In more general cases, we need to extend the method to even modify the real F0 contours of the recorded phrases, if more flexible use is assumed for the recorded phrases as well as the synthetic ones.

## 6. REFERENCES

- [1] Black, A. W., et al., "Limited Domain Synthesis," *Proc. of ICSLP 2000*.
- [2] Donovan, R. E., et al., "Phrase Splicing and Variable Substitution Using the IBM Trainable Speech Synthesis System," *Proc. of ICASSP 2000*.
- [3] Taylor, P., et al., "Speech Synthesis by Phonological Structure Matching," *Proc. of Eurospeech '99*.
- [4] Sagisaka, Y., "On the Quantification of Global F0 Pattern Control," *IEICE Technical Report*, SP89-111, 1989. (In Japanese)
- [5] Saito, T., et al., "A Method of Creating a New Speaker's VoiceFont in a Text-to-Speech System," *Proc. of ICSLP 2000*.
- [6] Saito, T., et al., "Generating F0 Contours by Statistical Manipulation of Natural F0 Shapes," *Proc. of Eurospeech 2001*.
- [7] Komazawa, T., *Theory of Quantification and Data Analysis*, Asakura Publishing Company, Japan, 1982.

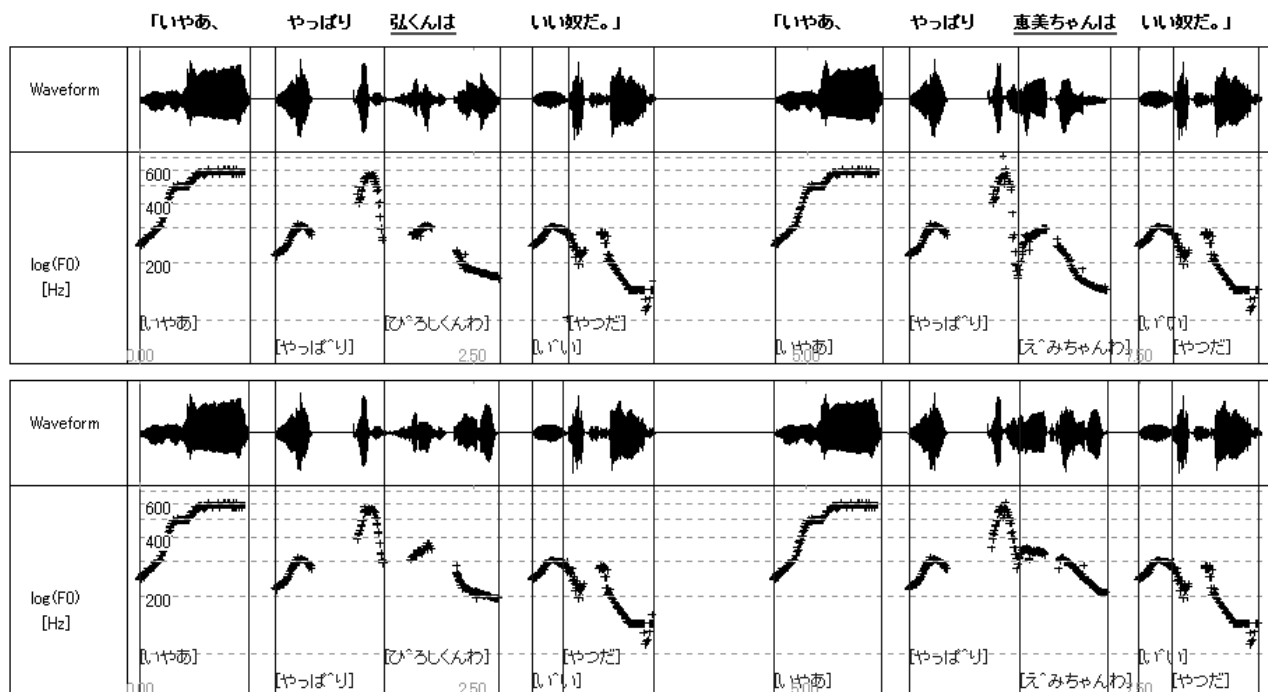


Figure 3. Example of Hybrid F0 Contours with Expressive Recorded Phrases (underlined: purely synthetic phrases) (Upper one: generated without surrounding real F0 shape information, Lower one: generated with the information)