ON THE USE OF F0 FEATURES IN AUTOMATIC SEGMENTATION FOR SPEECH SYNTHESIS

Takashi Saito

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

ABSTRACT

This paper focuses on a method for automatically dividing speech utterances into phonemic segments, which are used for constructing synthesis unit inventories for speech synthesis. Here, we propose a new segmentation parameter called, "dynamics of fundamental frequency (DF0)." In the fine structures of F0 contours, there exist *phonemic events* which are observed as local dips at phonemic transition regions, especially around voiced consonants. We apply this observation about F0 contours to a speech segmentation method. The DF0 segmentation parameter is used in the final stage of the segmentation procedure to refine the phonemic boundaries obtained roughly by DP alignment. We conduct experiments on the proposed automatic segmentation with a speech database prepared for unit inventory construction, and compare the obtained boundaries with those of manual segmentation to show the effectiveness of the proposed method. We also discuss the effects of the boundary refinement on the synthesized speech.

1. INTRODUCTION

Automatic speech segmentation was not essential for conventional text-to-speech synthesis, but is quite recently regarded as a highly attractive function of speech synthesis systems, which provides an efficient way to acquire new speakers' voice characteristics[1-4]. It is expected as a basic technology not only for automatic preparation of synthesis unit inventory but also prosodic feature customization, and it will spread the range of current text-to-speech applications.

This paper focuses on a method for automatic segmentation aimed to construct synthesis unit inventories for speech synthesis. We propose here a new segmentation parameter called, "dynamics of fundamental frequency (DF0)" and apply it to our voice registering system that is integrated in a text-to-speech synthesizer[4].

Fundamental frequency(F0) is one of important speech features utilized to wide variety of speech processing and applications. Especially its global (supra-segmental) structures are commonly understood as prosodic and syntactic representation of speech. It has been commonly used as basic parameters for speech synthesis to model and control accents and [5]. Moreover, it has been also applied to speech recognition field, for example, in phrase boundary detection to obtain syntactic structures for continuous speech recognition[6]. Besides these, there was an interesting study which reported through psychoacoustic experiments that F0 contours contain much information on speaker individualities[7]. In this way, the F0 feature of speech contributes greatly to wide variety of advanced speech processing. All these techniques reported above are, however, based mainly on the global structures of the F0 contours.

On the other hand, as for the local structures of F0 contours, there were a few but informative findings[8,9] from the standpoint of speech synthesis in Japanese.

- Voiced consonants have low tone property: Even in case of type-1 accent, F0 value of voiced consonants in the first syllables of word utterances stays low, and F0 increases immediately in the vowels that follow[8].
- For voiced consonants, F0 value goes down from the preceding vowels and up to the following vowels, i.e., local dips in F0 contours arise at consonant intervals. For unvoiced consonants, F0 goes down at the transition from the consonants to the following vowels[9].
- Intelligibility of voiced plosives was improved by introducing an F0 contour model considering the phonemic fluctuation[9].

They all discussed the relationship of F0 contours with the phonemic events for the purpose of improving prosodic control in speech synthesis.

Here, we attempt to apply the local F0 information related to the phonemic events to a speech segmentation system for speech synthesis. We propose a new segmentation parameter, "F0 dynamics," and use it especially for refining the phonemic boundaries of voiced consonants.

This paper is organized as follows. In section 2, we describe the outline of the segmentation method in the voice registration system. In section 3, we present the derivation of an F0 dynamic feature we propose for segmentation. In section 4, we describe the experiments conducted for evaluating the proposed segmentation method. In section 5, we discuss the effects of the proposed method on the synthesized speech. Finally, we summarize the results obtained here.

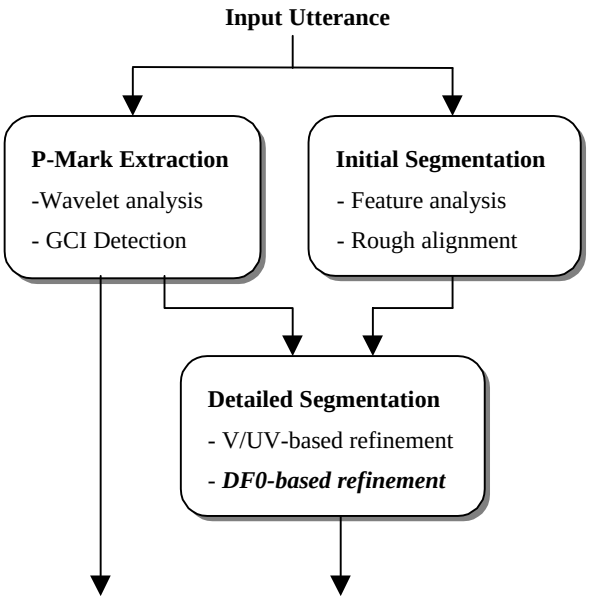


Figure 1: Speech Segmentation Part in the Voice Registration System

2. SEGMENTATION METHOD

The speech segmentation system is a part of our voice registration system[4], which is integrated within a text-to-speech synthesizer. Figure 1 shows the block diagram of the segmentation system proposed here. The procedure for the automatic segmentation consists of three stages: initial segmentation, pitch mark extraction, and detailed segmentation. Each stage is described below in details.

2.1 Initial Segmentation

In the initial segmentation, a word utterance of a new speaker is roughly divided into phonemic segments by DP alignment with a reference speaker's segmented utterance. Most of automatic segmentation systems for speech synthesis are using powerful HMM speech alignment techniques[1-3]. Currently, we use here a typical DP matching as an alignment tool since it is capable of segmenting word utterances for this rough alignment purpose, although the more powerful aligner would be the better. This rough alignment gives initial values for phonemic boundaries of the input utterance which are used in the detailed segmentation.

2.2 Pitch Mark Extraction

Next is the pitch mark extraction, and this procedure is required in our system to extract waveform control parameters, since we use a waveform-concatenation-based technique for speech synthesis method[10].

In our system, pitch mark information is obtained by a GCI(Glottal Closure Instant) detection method based on a wavelet signal analysis. The wavelet-based GCI detection method was originally proposed by Kadambe[11]. The dyadic wavelet transform applied in the method shows local maxima around the points of discontinuity of a signal. The GCI detector is based on the property to find discontinuity, since glottal closure causes abrupt changes in the derivative of the air flow in the glottis. The original algorithm[11] is quite simple and effective, but there were some problems found in our preliminary experiments.

- V/UV detection is not satisfactory since it is based only on the amplitude of the wavelet transform output.
- Post processing is needed to select reliable GCIs from candidates because it tends to suffer from insertion errors.

In our implementation, a decision on whether a frame should be voiced or unvoiced is taken prior to the GCI detection so as to reinforce the detection procedure. The voiced/unvoiced decision is made by using the information on the log power and zero crossing rate of speech signal. We also refined the original method to improve its detection accuracy and robustness by checking the continuity of the GCI candidates.

As the result of the GCI detection, precise F0 values are obtained, as well as the pitch mark information. The F0 values are utilized in the final stage of segmentation described below.

2.3 Detailed Segmentation

The purpose of the detailed segmentation is to refine phonemic boundaries obtained by the automatic alignment so as to realize the speech synthesis procedure in a simple and straightforward way, such as unit concatenation, duration manipulation.

In the detailed segmentation, two kinds of segment-boundary refinement procedures are carried out according to the boundary types: one is V/UV-based refinement for phonemic boundaries at unvoiced-to-voiced or voiced-to-unvoiced transitions, and the other is DF0-based refinement for boundaries at voiced-to-voiced transitions between consonants and vowels. Both refinements are carried out based on the precise F0 information obtained in the pitch mark extraction.

For the former type of boundaries, the initial boundaries are adjusted simply to the nearest starting or ending points of pitch-marks in voiced segments obtained in the previous stage. For the latter type of boundaries, a powerful boundary refinement is carried out by using the new F0 feature, DF0, defined in the next section.

3. F0 DYNAMICS

3.1 Derivation of F0 Dynamics (DF0)

First, obtain F0 (= 1/T0, T0: pitch period) value by taking the interval between adjacent GCIs as T0. Smoothed logarithmic F0 value, SF0, is then obtained for each fixed-length segment by calculating a mean value of log(F0) in the segment. This smoothing operation is needed to eliminate perturbations irrelevant to phonemic events. Finally, the dynamics of F0 pattern,

DF0, is obtained as the slope of the regression line of SF0 by the following equation.

$$DF0(i) = \left(\sum_{i=-N}^{i=N} i\, W(i)\, SF0(i)\right) / \left(\sum_{i=-N}^{i=N} i^2\, W(i)\right)$$

where W(i) : symmetric weighting function

3.2 Use of DF0 in Detailed Segmentation

DF0 is applied to the boundary refinement of voiced consonants in the detailed segmentation as follows.

1. Search range setting: The search range for segment boundaries to be refined is set as (+Ts,-Ts) from the initial boundaries obtained in the rough alignment. Ts is a fixed time length, and it was set 30 ms here through a preliminary experiment.

2. Find a phonemic event: First, find a local minimum point in the search range of the starting point of a voiced consonant. If found, then find a local maximum point that follows the local minimum point in the search range of the ending point of the consonant. If the differential value between the two points is greater than a predefined threshold, the interval from the local minimum point to the local maximum point is detected as a phonemic event and selected as refined boundaries for the consonant.

In figure 2, an example of DF0 calculation for a word utterance is shown in the bottom frame. In the voiced consonant regions(/n/,/b/,/r/), the F0 contour has fairly distinct dips that indicate phonemic events, and these dips can be captured by DF0 as an interval from a local minimum point to a local maximum point.

4. EXPERIMENTS

We conducted experiments on the automatic segmentation using speech database prepared for unit inventory construction, and compared the obtained boundaries with those of manual segmentation to examine the effectiveness of the proposed segmentation parameter.

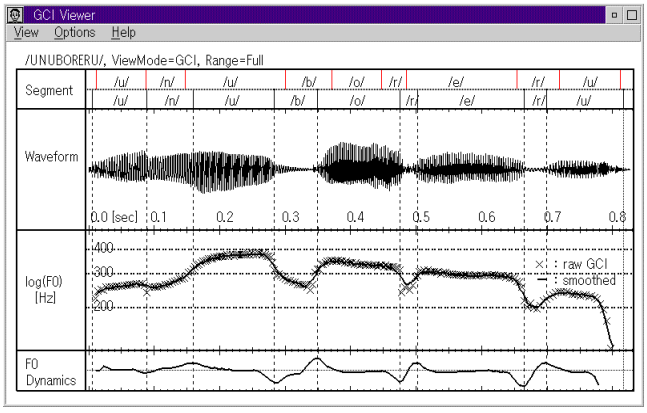


Figure 2: Example of Analyzed Utterance
Segment: (upper: initial segmentation, lower: detailed seg.)
logF0: (x: F0 raw data, solid line: smoothed F0(SF0))
F0 Dynamics: (Refined boundaries are indicated by vertical lines.)

4.1 Speech Data Preparation

We use the reference unit inventory developed for the registration system[4] as reference template for the rough alignment in the initial segmentation. As a new speaker's data for this experiment, we have prepared speech data of a 1530-word set for two female speakers (speaker-Fa and speaker-Fb). The vocabulary set contains phonemically-balanced words which was prepared for constructing context-dependent syllabic unit for Japanese speech synthesis[12]. The two sets of speech data were segmented manually to compare with the results of the automatic method. The manual segmentation was carried out on the basis of finding the center of spectral transition between phonemes.

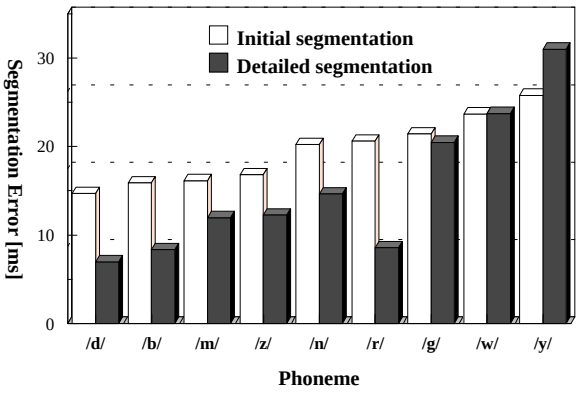
4.2 Comparison with Manual Segmentation

We applied the DF0-based refinement in the detailed segmentation with the same threshold parameters to all the voiced consonants in the database except at word-top position. This is intended to investigate the difference in the effects of DF0 among consonant types. Prior to this experiment, GCI errors(1% for speaker Fa, 2% for speaker Fb) for voiced regions were manually corrected so as to focus on the effect of DF0 parameter on the refinement. As a result, the average boundary error in the detailed segmentation decreased 28.9 % from the initial segmentation for speaker Fa, and 16.3 % for speaker Fb. (The boundary error is defined as the average value of left-hand(vowel-to-consonant) and right-hand(consonant-to-vowel) boundary errors of a consonant.) The proposed parameter DF0 seems to work successfully in the boundary refinement.

The difference in the improvement rate between the two speakers is caused by the difference in the results of the initial segmentation, since the average errors for both speakers in the detailed segmentation are on the same level; 13.6 ms for Fa, and 12.3 ms for Fb, respectively. In other words, this refinement might be robust to the rough alignment performance.

Figure 3 shows the average segmentation errors for speaker Fa in

Figure 3. Average Segmentation Error



initial and detailed segmentation, classified according to the type of consonants. In particular, the improvement rates for /r/,/d/,/b/ sounds are distinct. The result of speaker Fb has the same tendency, and it also corresponds well with the results obtained in Takeda's study of an F0 contour generation model[9]. DF0 parameter is likely to be particularly effective for these kinds of sounds. On the other hand, as for semivowels such as /w/,/y/, DF0-based method does not seem to work at all. This is because the fluctuation of F0 contour around semivowels are generally very small.

5. DISCUSSION

5.1 Algorithm Improvement

The DF0-based refinement algorithm is carried out by the search of local minima and local maxima in DF0 patterns. There are some cases where the correct boundaries fail to be found although the phonemic events for the consonants exist; they are out of the search range. Thus, it might be better to adapt the search range to the scores of the rough alignment.

On the observation of analyzed data over multiple speakers, local minima in DF0 tend to be found more stable and distinct than local maxima. Therefore, the algorithm of finding the phonemic events of F0 dips might be improved in robustness by trusting only or putting weight on the search of local minima.

5.2 Effects on Synthesized Speech

We conducted an informal listening test that compared three kinds of synthesized speech: (Sa) is synthesized by using synthesis units based on the results of the initial segmentation, (Sb) is synthesized by using synthesis units based on the detailed segmentation, and (Sc) is synthesized by using synthesis units based on the manual segmentation. Speech samples are Japanese words that include the consonants only of /r/, /b/ and /d/, which are highly improved in the refinement. As the result, (Sb) and (Sc) were hard to distinguish. We observed, however, several differences of (Sa) from (Sb) and (Sc) as follows:

- Some consonants were not so clear or were making noises because of spectral gap at unit-connection (vowel-to-consonant) boundaries.
- Particularly in a slow speaking rate, some consonants became very unnatural because inappropriate portions were stretched in time-scale modification.

These problems are caused primarily by the failure in cutting synthesis units from utterances although improvements in the speech generation phase might contribute to solve them, and short consonants like Japanese /r/ sound seem to have a strong tendency to suffer from this type of degradation.

6. SUMMARY

In this paper, we have presented a speech segmentation method which aims to automatically constructing synthesis unit inventories, and have proposed a new segmentation parameter called, "F0 dynamics." It is motivated by the observations on F0 contour movements related to phonemic events. We have conducted experiments on the proposed automatic segmentation for 1530 words uttered by two female speakers, and compared the obtained boundaries with those of manual segmentation. As a result, the F0 dynamics is shown to work successfully in the boundary refinement. In particular, the improvements for /r/,/d/,/b/ sounds are distinct in the detailed segmentation. We have also discussed the effects of the DF0-based boundary refinement on the synthesized speech.

7. REFERENCES

1. S. Pauws et al., "A Hierarchical Method of Automatic Speech Segmentation for Synthesis Applications," Speech Communication, Vol.19, pp.207-220, 1996.
2. R. E. Donovan et al., "Automatic Speech Synthesizer Parameter Estimation using HMMs," Proceedings of ICASSP '95, pp. 640-643, 1995.
3. X. Huang et al., "Whistler: A Trainable Text-to-Speech System," Proceedings of ICSLP'96, pp.2387-2390, 1996.
4. T. Saito, "A Method for Registering New Voices in a Text-to-Speech Synthesizer," Proc. of 3rd ASA & ASJ Joint Meeting, pp.1057-1060, 1996.
5. H. Fujisaki , K. Hirose, "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese," Journal of ASJ (E), 5, 4, 1984.
6. A. Sakurai, K. Hirose, "Detection of Phrase Boundaries in Japanese by Low-pass Filtering of Fundamental Frequency Contours," Proceedings of ICSLP'96, pp.817-820, 1996.
7. H. Ohno, M. Akagi, "Speaker Individuality in Fundamental Frequency Contours of Sentences," technical report of IEICE, SP97-128, 1998(in Japanese).
8. H. Sato, "Analysis of Fundamental Frequency Characteristics Related to Phonemes," Proc. of ASJ Annual Meeting, 2-3-18, pp.259-260, 1989 (in Japanese).
9. S. Takeda, "A Model for Generating Fundamental Frequency Contours Considering Phonemic Fluctuation and Rules for Speech Synthesis," Journal of IEICE, J73-A, 3, pp.379-386, 1990 (in Japanese).
10. M. Sakamoto et al., "A New Waveform Overlap-Add Technique for Text-to-Speech Synthesis," IEICE technical report, SP95-6, 1995 (in Japanese).
11. S. Kadambe et al., "Application of the Wavelet Transform for Pitch Detection of Speech Signals," IEEE Trans. Info. Theory, vol.38, pp. 917-924, 1992.
12. T. Saito et al., "High-Quality Speech Synthesis Using Context-Dependent Syllabic Units," Proceedings of ICASSP '96, pp. 381-384, 1996.