December 24, 1998
RT0293
Human-Computer Interaction 4 pages

Research Report

Use of F0 Features in Automatic Segmentation for Speech Synthesis

Takashi Saito

IBM Research, Tokyo Research Laboratory
IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato
Kanagawa 242-8502, Japan

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

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* that 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 conducted experiments using the proposed automatic segmentation method with a speech database prepared for unit inventory construction, and compare the resulting boundaries with those obtained by manual segmentation to show the effectiveness of the proposed method. We also discuss the effects of the boundary refinement on synthesized speech.

1. INTRODUCTION

Automatic speech segmentation is not essential for conventional text-to-speech synthesis, but has quite recently come to be regarded as a highly attractive function of speech synthesis systems, which provides an efficient way of acquiring new speakers' vocal characteristics [1-4]. It is expected to become a basic technology not only for automatic preparation of synthesis unit inventories but also for prosodic feature customization, and to extend the range of current text-to-speech applications.

This paper focuses on a method of automatic segmentation for constructing 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, which is integrated into a text-to-speech synthesizer [4].

Fundamental frequency (F0) is one of the most important speech features in a wide variety of speech processing and applications. In particular, its global (supra-segmental) structures are commonly understood as prosodic and syntactic representation of speech. It has been commonly used as a basic parameter in speech synthesis, for modeling and controlling accents [5]. It has also been applied in the speech recognition field, for example, in phrase boundary detection to obtain syntactic structures for continuous speech recognition [6]. An interesting study reported on the basis of psychoacoustic experiments that F0 contours contain much information on speaker individualities [7]. In this way, the F0 feature of speech has made strong contributions to advanced speech processing in a wide variety of ways. All the techniques reported above are, however, based mainly on the global structures of the F0 contours.

As regards the local structures of F0 contours, there have been a few informative findings [8,9] from the standpoint of speech synthesis in Japanese:

- Voiced consonants have low tone property: Even in the case of type-1 accents, the F0 value of voiced consonants in the first syllables of word utterances remains low, and F0 increases immediately in the vowels that follow [8].
- For voiced consonants, the 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 transitions from the consonants to the following vowels [9].
- The intelligibility of voiced plosives was improved by introducing an F0 contour model that takes account of the phonemic fluctuation [9].

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

Here, we attempt to apply the local F0 information related to 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 that we propose for segmentation. In section 4, we describe experiments conducted to evaluate the proposed segmentation method. In section 5, we discuss the effects of the proposed method on synthesized speech. Finally, we summarize the results obtained in this study.

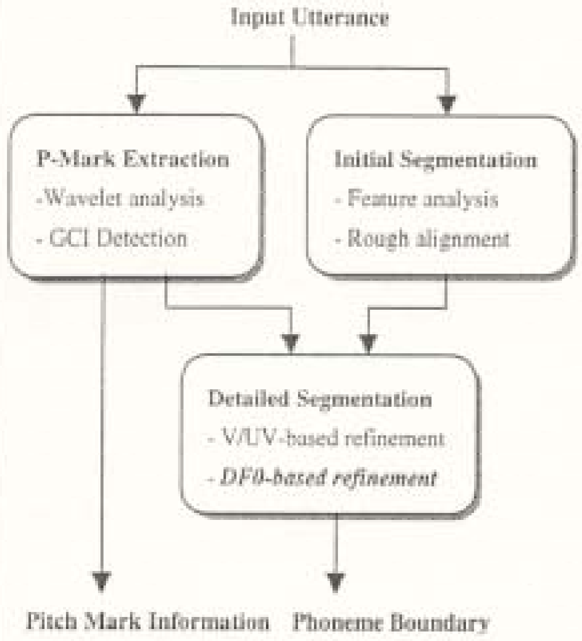


Figure 1: Speech Segmentation Part in the Voice Registration System

2. SEGMENTATION METHOD

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

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 automatic segmentation systems for speech synthesis use powerful HMM speech alignment techniques [1-3]. Currently, we use a typical DP matching technique as an alignment tool, since it is capable of segmenting word utterances for the purpose of this rough alignment, although a more powerful aligner would be better. This rough alignment gives initial values for the phonemic boundaries of the input utterance, which are used in the detailed segmentation.

2.2 Pitch Mark Extraction

The next step is 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 our speech synthesis method [10].

In our system, pitch mark information is obtained by a glottal closure instant (GCI) detection method based on 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 uses this 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 some problems arose 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 the method 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 signals. We also refined the original method to improve its detection accuracy and robustness by checking the continuity of the GCI candidates.

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

2.3 Detailed Segmentation

The purpose of the detailed segmentation is to refine the 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 and duration manipulation.

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

For boundaries of the first type, 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 boundaries of the second type, a powerful boundary refinement is achieved by using the new F0 feature, DF0, defined in the next section.

3. F0 DYNAMICS

3.1 Derivation of F0 Dynamics (DF0)

First, the value of F0 (= 1/T0, T0: pitch period) is obtained by taking the interval between adjacent GCIs as T0. A smoothed logarithmic F0 value, SF0, is then obtained for each fixed-length segment by calculating the 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:

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

where W(i) is a 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 to 30 ms in accordance with a preliminary experiment.
2. Phonemic event location: First, find a local minimum point in the search range of the starting point of a voiced consonant. If one is 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 intervals between local minimum points and local maximum points.

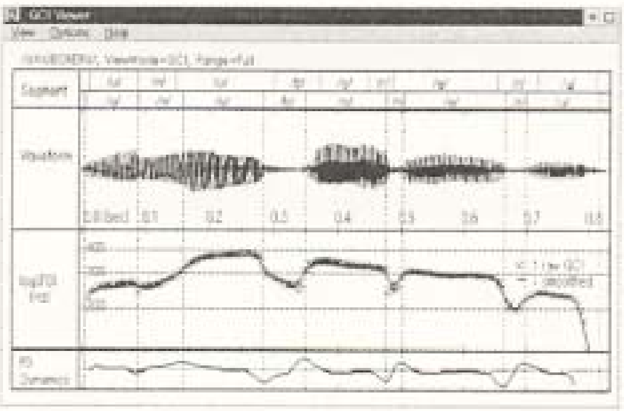


Figure 2: Example of Analyzed Utterance
Segment: (upper: initial segmentation, lower: detailed seg.)
logF0: (x: F0 raw data, solid line: smoothed F0 (SF0))
F0 Dynamics: (Refined boundaries are indicated by vertical lines.)

4. EXPERIMENTS

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

4.1 Speech Data Preparation

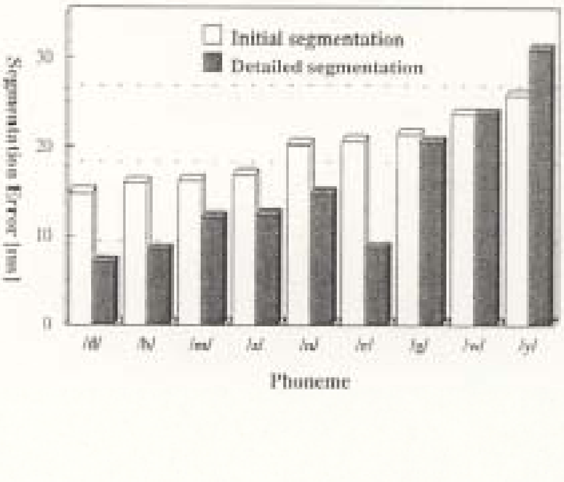
We use the reference unit inventory developed for the registration system [4] as a reference template for rough alignment in the initial segmentation. As a new speaker's data for this experiment, we have prepared a 1530-word set of speech data for two female speakers (speaker-Fa and speaker-Fb). The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

4.2 Comparison with Manual Segmentation

To investigate the difference in the effects of DF0 for various consonant types, we applied the DF0-based refinement in the detailed segmentation with the same threshold parameters to all the voiced consonants in the database except those in word-top position. Prior to this experiment, GCI errors (1% for speaker Fa, 2% for speaker Fb) for voiced regions were manually corrected in order to focus on the effect of the DF0 parameter on the refinement. As a result, the average boundary error in the detailed segmentation decreased by 28.9% from the initial segmentation for speaker Fa, and by 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 be effective for boundary refinement.

The difference in the improvement rate between the two speakers was caused by the difference in the results of the initial segmentation, since the average errors for both speakers in the detailed segmentation were of the same order: 13.6 ms for Fa,

Figure 3. Average Segmentation Error



and 12.3 ms for Fb, respectively. In other words, this refinement might be robust with respect to the rough alignment performance.

Figure 3 shows the average segmentation errors for speaker Fa in initial and detailed segmentation, classified according to the type of consonants. The improvement rates for /r/,/d/,/b/ sounds are particularly distinct. The results for speaker Fb show the same tendency, and also correspond well with the results obtained in Takeda's study of an F0 contour generation model [9]. The DF0 parameter is likely to be particularly effective for sounds of these kinds. On the other hand, for semivowels such as /w/ and /y/, the DF0-based method does not seem to work at all, because the fluctuation of the F0 contour around semivowels is generally very small.

5. DISCUSSION

5.1 Algorithm Improvement

The DF0-based refinement algorithm searches for local minima and local maxima in DF0 patterns. In some cases, the correct boundaries fail to be found even though phonemic events for the consonants exist, because the boundaries lie outside the search range. Thus, it might be better to adapt the search range to the scores in the rough alignment.

Observation of analyzed data for multiple speakers shows that local minima in DF0 tend to be more stable and distinct than local maxima. Therefore, the robustness of the algorithm for finding the phonemic events of F0 dips might be improved by trusting only, or emphasizing, the search for local minima.

5.2 Effects on Synthesized Speech

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

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

These problems are caused primarily by failure to separate synthesis units from utterances although improvements in the speech generation phase might help to solve them, and short consonants such as the 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 for use in automatic construction of synthesis unit inventories, and proposed a new segmentation parameter called "F0 dynamics," which is motivated by observations on F0 contour movements related to phonemic events. We 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 was shown to be effective for boundary refinement. In particular, it yielded distinct improvements for /r/,/d/,/b/ sounds in the detailed segmentation. We also discussed the effects of DF0-based boundary refinement on 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 and K. Hirose, "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese," Journal of ASJ (E), 5, 4, 1984.
6. A. Sakurai and 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 and 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.