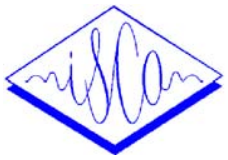

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A NEW F_0 MODIFICATION ALGORITHM BY MANIPULATING HARMONICS OF MAGNITUDE SPECTRUM

Satoshi TAKANO, Masanobu ABE
NTT Cyber Space Labs.
1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa, 239-0847 Japan
Satoshi@nttspch.hil.ntt.co.jp

ABSTRACT

This paper proposes a new speech modification algorithm based on a vocoder framework to synthesize high quality speech. Its innovation is in preserving the fine structure of the magnitude spectrum. A key point is the use of a "compensatory gaussian window" to extract moderate F_0 harmonics structures in the magnitude spectrum. The other key point is, starting from the magnitude spectrum, generating the F_0 harmonics structures that match the target's fundamental frequency. Preference tests show that the proposed algorithm synthesizes higher quality speech than TD-PSOLA if large prosody modification is needed, and that the spectral envelope produced by the proposed algorithm is superior to any other conventional vocoders, especially when modifying the frequency upward.

Keywords: text-to-speech synthesis, phase manipulation, harmonics modification, FFT, fundamental frequency

1. INTRODUCTION

A text-to-speech system based on speech unit concatenation requires prosody modification. Minimizing the speech degradation that normally occurs during modification has been a key issue for improving synthesized speech quality. Although the TD-PSOLA [1] algorithm is the most popular synthesis algorithm in current TTS systems, it still has problems such as a relatively narrow range of modification wherein naturalness is retained; speech distortion is evident if prosody modification is large. This problem is avoided in the other approach of using a vocoder framework. The biggest advantage of this approach is its high flexibility in speech modification. This makes it possible not only to enlarge the dynamic range of prosodic parameter modification to synthesize emotional speech and conversational speech, but also to control voice quality and change speaker identity. We believe that the flexibility offered by the vocoder framework is most important. A new vocoder called STRAIGHT [2] was recently proposed and has attracted attention due to its high quality. This contradicts the common assumption that the vocoder approach could not match the quality of TD-PSOLA.

In this paper, we propose a high quality prosody modification algorithm using an idea taken from STRAIGHT. Our main interest is the analysis part, and we directly apply the synthesis procedure of STRAIGHT. Section 2 describes the relation between fine spectral structure and speech quality. Section 3 proposes a spectral harmonic modification algorithm. Section 4 explains the synthesis performed by STRAIGHT. Section 5 describes the subjective experiments used to evaluate proposed algorithm.

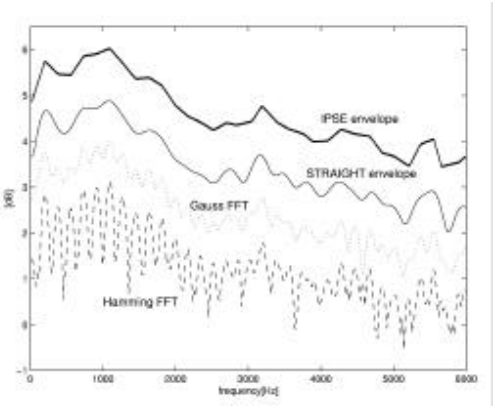

Figure 1: Hamming window FFT magnitude spectrum (moved downward for comparison: broken line) and gaussian window FFT magnitude spectrum (dotted line). IPSE envelope (thick line) and STRAIGHT envelope (thin curve line).

2.SPEECH ANALYSIS AND SPECTRAL ENVELOPE

2.1 Compensatory Gaussian Window

In short period FFT-based analysis, changing the cutting window function can change the spectral harmonics structures extracted. The periodicity of these structures is related to the fundamental frequency (F_0). STRAIGHT uses a compensatory gaussian window as a window function to extract the spectrum. The advantages of the compensatory gaussian window are isometric resolution in both the time and frequency domains, minimum uncertainty, and better temporal smoothing.

Fig. 1 compares the FFT magnitude spectrum gained using the conventionally Hamming window and that output by the compensatory gaussian window with window length of three pitch periods. Though conventional vocoders use the Hamming window because of its high frequency resolution, its spectrum has excessively large harmonics such as illustrated in Fig. 1 and it is too sensitive to F_0. In contrast, the compensatory gaussian window yields a spectrum that is close to the envelope and has both moderate and fine structure. Hence, in this paper, we use the compensatory gaussian window for spectrum extraction.

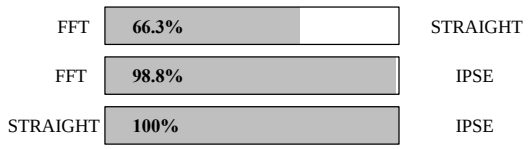

Figure 2: Preference test with copy synthesis from FFT magnitude spectrum (FFT), STRAIGHT envelope (STRAIGHT) and IPSE envelope (IPSE). Score in center bar is percentage of subjects who selected the left one. The results show a significant difference at the 1 % significant level.

2.2 Comparison of Extracting Spectral Envelope by Copy synthesis

Many algorithms can extract an envelope from a spectrum. We start by assuming that the spectrum is output by the compensatory gaussian window and evaluate algorithms to extract the spectral envelope. We adopt the IPSE[3] algorithm as an example of conventional envelope extracting algorithms and compare this to the STRAIGHT algorithm. The IPSE envelope is extracted pitch-synchronously and approximated by interpolating the local peaks of the harmonics of the log-power spectrum using the cosine function. On the other hand, the STRAIGHT envelop is extracted from the spectrum by interpolating the bilinear surface in the time - frequency domain. Figure 1 shows that the non-smoothed spectrum contains more fine structure than STRAIGHT and STRAIGHT contains more fine structure than IPSE.

Next, we ranked the non-smoothed spectrum envelope, IPSE envelope, and STRAIGHT envelope in terms of speech quality. Copy synthesis from natural speech without prosody modification was used to confirm which input spectrum produces higher speech quality. The synthesis procedure of STRAIGHT was used in all cases. Copy synthesis was applied to four natural female speech signals sampled at 22.05kHz. Speech samples were synthesized from the three envelopes. As for analysis parameters, FFT dimension was 1024 and frame shift was 5msec. The pitch extraction algorithm of STRAIGHT was used in all cases. The listeners were 10 females. They were asked to select the preferable sample from randomized speech pairs extracted from the three samples.

The results of this preference test are shown in Fig. 2. With respect to speech quality, FFT is better than STRAIGHT and STRAIGHT is better than IPSE. Figure 1 shows that FFT contains more fine structure than STRAIGHT and IPSE in that order. That is, speech quality increases with the level of fine structure in the spectrum.

3. PROPOSED ALGORITHM

The previous section shows that preserving the fine structure is effective for the copy synthesis of natural speech. However, when prosody modification is required, the spectrum envelope is conventionally smoothed in order to avoid mismatching between the fine structure and the fundamental frequency (F_0).

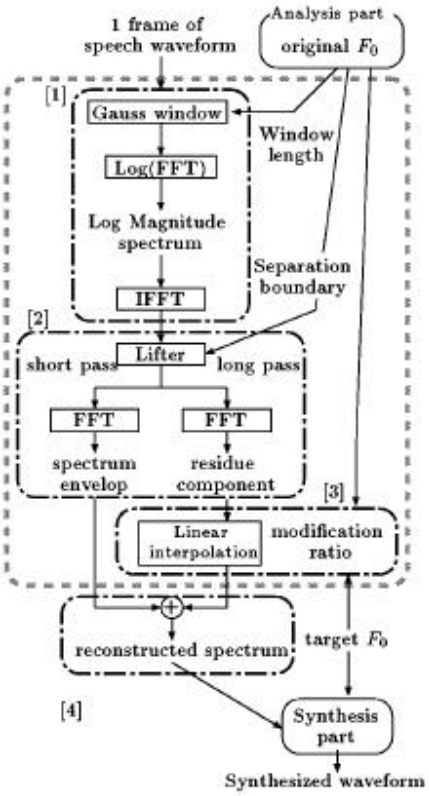

Figure 3: Block diagram of proposed algorithm.

Another approach to preventing mismatch is generating the fine structure to match the target F_0 value. This is possible by modifying the separated residue component and adding it to the spectrum envelope.

The modification algorithm is proposed below. The basic idea is modifying the residue component in the frequency domain according to the ratio of target F_0 to original F_0. A block diagram is illustrated in Fig. 3.

[1] Complex cepstrum is calculated using short-time FFT with the compensatory gaussian window whose length corresponds to three pitch periods.

[2] Spectrum is split by a cepstrum lifter whose boundary is set from F_0. Residue component is output from a long pass lifter; the envelope component is output from a short pass lifter.

[3] Fine spectral structure is generated by linear compression or expansion of the residue component. The factor determining compression or expansion is F_0 modification ratio between intrinsic F_0 and target F_0.

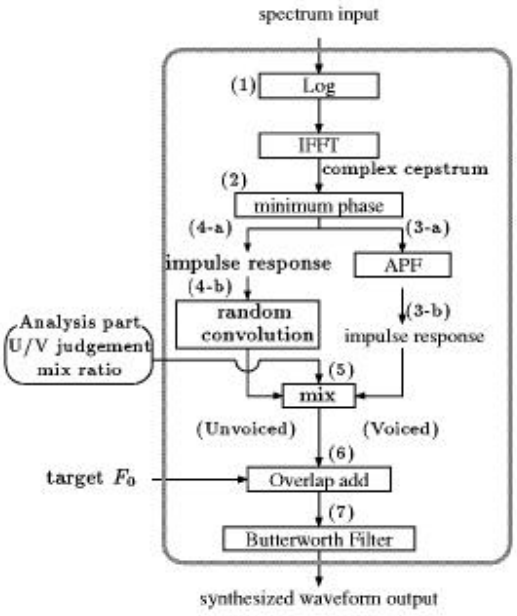

Figure 4: Block diagram of synthesis part in STRAIGHT. APF is sequence of all pass filters.

[4] Generated fine structure is added to the spectrum envelope.

4. STRAIGHT SYNTHESIS

The synthesis procedure of STRAIGHT is illustrated in Fig. 4. It is an pitch asynchronous overlap add synthesis algorithm that manipulates the phase of the impulse response by all pass filters.

(1) Complex cepstrum is calculated.
(2) Minimum phase complex spectrum is calculated.
(3) (Voiced Component)
  (3-a) All pass filters manipulate phase of spectrum
  (3-b) Impulse response is calculated by IFFT.
(4)(Unvoiced Component)
  (4-a) Impulse response is calculated by IFFT.
  (4-b) Impulse response is convoluted by random sequence.
(5)The voiced and unvoiced components are mixed using a mixture ratio parameter taken from analysis part.
(6)Overlap-add synthesis is performed; its interval is determined by F0.
(7)Synthesis waveform is filtered by a butterworth filter.

STRAIGHT exhibits degradation in some unvoiced phonemes, so we developed hybrid synthesis which uses the PSOLA synthesis module for unvoiced phonemes and the proposed synthesis module for voiced phonemes.

5. EVALUATION OF THE PROPOSED ALGORITHM

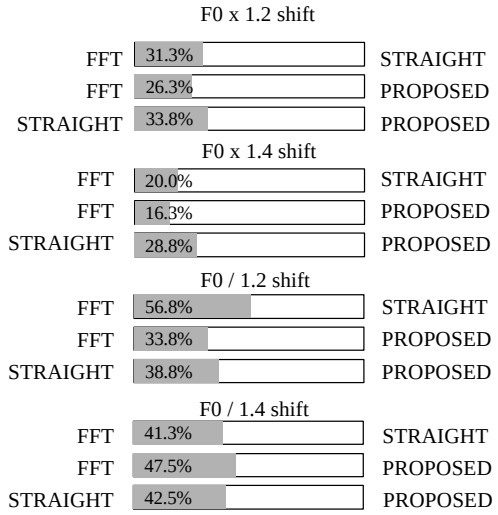

Figure 5: Preference test of uniform F0 modification. A comparison with a double circle has 1 % significant level. A single circle indicates 5 % significant level.

5.1 Uniform prosody modification

This section compares the proposed algorithm to the non-smoothened FFT spectrum and STRAIGHT-envelope.

To compare them as input spectra for synthesis, uniform prosody modification (moving F_0 contour upward and downward) of natural speech was performed. The target prosody is a version of the original F_0 contour shifted by a constant scaling factor. The factors are 1.2, 1.4(upward modification), and 1/1.2, 1/1.4 (downward modification). This modification was applied to four natural female speech segments (1 second to 2 seconds long). Listeners were 10 females; they were asked to select the sample from randomized speech sample pairs extracted from the three samples.

The results of this preference test are shown in Fig. 5. For upward prosody modification, the proposed algorithm and STRAIGHT yield better quality than PSOLA and the difference increases with the degree of modification. For downward modification, the proposed algorithm is better than STRAIGHT. Consequently, the proposed algorithm is the best of the three schemes for uniform modification.

5.2 Synthesis by rule

In this section, we compare the proposed algorithm to TD-PSOLA and STRAIGHT. To evaluate the overall performance of the proposed algorithm for synthesis-by-rule, we composed sample speech using two speech unit concatenation patterns. One was the concatenation of short units (Japanese unit concatenation strategy called CV-VC) involving large modification and the other was the concatenation of long units (Japanese unit concatenation strategy called CVC) involving small modification. The former is the worse case (Fig. 6). In this context, C is Consonant and V is Vowel.

Target prosody of sample speech was given by rule and modified by a prosody editing tool [4]. Four prosody patterns were created by uniformly scaling the target prosody in order to evaluate various F_0 ranges. The factors were 1.0 (no modification), 1.2 (high F_0 contour), 1/1.2 (low F_0 contour). The

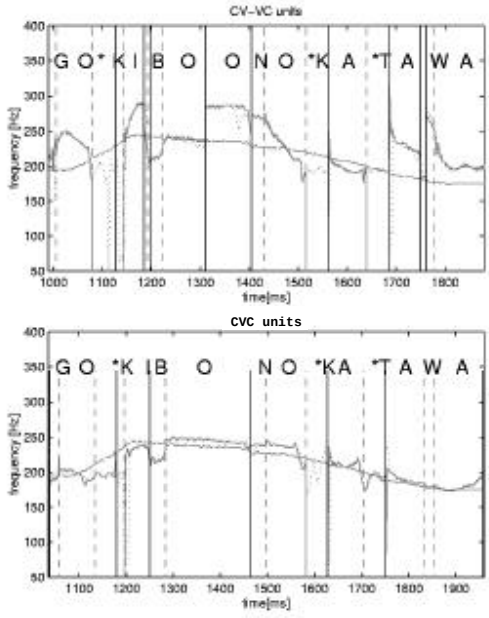

Figure 6: F_0 pattern of example synthesis-by-rule (Japanese text "Gokiboonokatawa"). Vertical straight line shows unit boundary and vertical dotted line shows phoneme boundary. Continuous line shows target F0 and discontinuous line shows intrinsic F0 of units. Top figure shows short units for large modification (worse case) and bottom figure shows long units for small modification (better case).

sample speech was three phrases (1 second to 2 seconds long). Listeners were 9 females; they were asked to select the preferable sample from randomized speech sample pairs.

Preference test results are shown in Fig.7 and 8. For the CV-VC samples, the proposed algorithm yielded better synthesized speech than PSOLA, especially for downward prosody modification. On the other hand, for the CVC samples, the three algorithms yield similar performance. As a result, the proposed algorithm is effective if the prosody modification is large.

6. CONCLUSION

This paper investigated synthesizing high quality speech by vocoder framework. We showed that the fine structure of the magnitude spectrum strongly influences synthesized speech quality. To preserve the fine structure during prosody modification, we proposed harmonic modification of the residue component and generation of fine structure for the new F_0. Preference tests were done to evaluate the proposed algorithm. Compared with PSOLA synthesis, the proposed algorithm offers better quality if the modification is large. That is, the proposed algorithm is better than PSOLA for synthesis-by-rule.

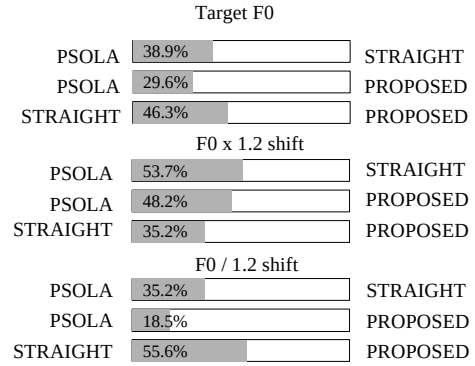

Figure 7: Preference test's result of synthesis-by-rule. CV-VC units. (worse case)

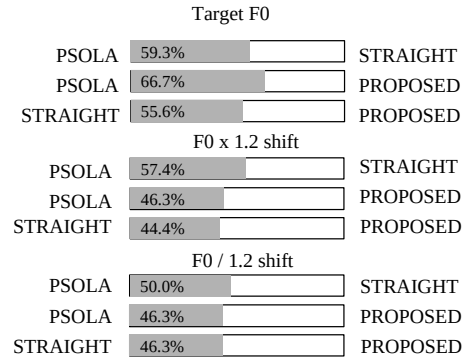

Figure 8: Preference test's result of synthesis-by-rule. CVC units. (better case)

ACKNOWLEDGEMENT

We are grateful to the members of the Speech Synthesis Group for their helpful discussions. We also thank Mr. Yamamori, director of the Media Processing Project, for his continuous support of this work.

REFERENCES

[1] E. Moulines, F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, Vol.9 No.5/6, pp.453-467, 1990.

[2] H. Kawahara. Speech representation and transformation using adaptive interpolation of weighted spectrum : Vocoder revisited. *In Proceedings of ICASSP*, Vol.2, pp.1303-1306, Muenich, 1997.

[3] K. Tanaka, M. Abe. A new fundamental frequency modification algorithm with transformation of spectrum envelope according to F_0. *In Proceedings of ICASSP*, Vol.2, pp.951-954, 1997.

[4] H. Mizuno, M. Abe, S. Nakajima. Development speech design tool "Sesign99" to enhance synthesized speech. *Eurospeech'99*, 1999.