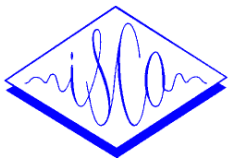

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

NEW SPECTRUM INTERPOLATION METHOD FOR IMPROVING QUALITY OF SYNTHESIZED SPEECH

Takashi Endo & Shun'ichi Yajima

Central Research Laboratory, Hitachi, Ltd.
1-280, Higashi-koigakubo, Kokubunji-shi, Tokyo 185, Japan

ABSTRACT

This paper proposed a new spectrum interpolation method for improving quality of synthesized speech. This method named SSFL(Spectrum Sliding on Formant Loci) overcomes one of the most serious problems in the OLA(Overlapping Addition) method that the continuity of formant loci at a transition part between phonemes is not guaranteed. This discontinuity of formant loci decreases quality of synthesized speech.

In the SSFL method, waves are transformed into spectra to interpolate in the spectrum domain and then spectra interpolated are transformed into waves to produce synthesized speech.

We have evaluated the proposed SSFL method by synthesizing continuous Japanese speech /aiueo/. The evaluation results have confirmed that the proposed SSFL improved the quality of synthesized speech because it guaranteed the continuity of formant loci.

I. INTRODUCTION

As a high quality speech synthesis algorithm, the OLA(Overlapping Addition) method has been studied[1][2][3]. The OLA method has two advantages. First, the spectrum envelope is exactly same as that of natural voice. Second, this method requires smaller computational power than that of other methods.

However, it has two disadvantages. First, it requires a lot of storage for storing synthesis units. This problem has become not serious according to the storage technology progress. Second, continuity of formant loci at a boundary part between synthesis units is not guaranteed. This problem is still serious.

Let $\mathbf{S}(t_1)$, $\mathbf{S}(t_2)$ and $\tilde{\mathbf{S}}(t)$ be the spectrum of phoneme before a boundary, the spectrum of phoneme after the boundary and the spectrum at the boundary, respectively. In the OLA method, $\tilde{\mathbf{S}}(t)$ is estimated as follows.

$$\tilde{\mathbf{S}}(t)=\frac{(t_2-t)\mathbf{S}(t_1)+(t-t_1)\mathbf{S}(t_2)}{t_2-t_1} \qquad (1)$$

Fig. 1 shows a spectrum shape estimated by the OLA method. Comparing the spectrum estimated by the OLA $\tilde{\mathbf{S}}(t)$ with the ideal spectrum $\mathbf{S}(t)$ drawn in dotted line, the spectrum estimated by OLA has formants at incorrect positions.

This discontinuity of formants by the OLA is one of the most serious problems.

To solve this problem, there are two research issues. The first one is how to guarantee the continuity of formant loci and the second one is how to determine the formant loci. For the first issue, we propose a new method called the SSFL.

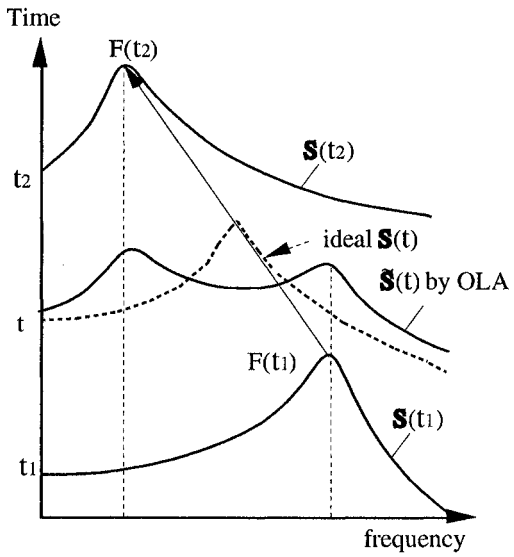

Fig. 1 Interpolated spectrum by OLA method

This method produces interpolation spectra that have continuity formant loci between given two adjacent phonemes' spectra. For the second subject, the authors analyze the spectrum transition of Japanese syllable of V/r/V in order to clarify how to construct rules that control formant loci at a phoneme boundary.

In the 2nd chapter, we explain the detailed algorithm of the SSFL and the evaluation of it. In the 3rd chapter, we describe the analysis of spectrum transition. The 4th chapter is for conclusion.

II. SSFL

2.1 Overview of SSFL

This new spectrum interpolation method proposed in this paper aims to control formant positions without discarding spectrum information. To achieve this aim, the SSFL method divides spectrum into the formant position-based feature and the formant position-independent feature, and interpolates these two features independently. The smooth formant loci are created between two spectra by a linear interpolation. The formant position-based spectrum feature is completely replaced with the spectrum envelope calculated from the formant loci. The formant position-independent spectrum feature is interpolated by the linear interpolation. After the interpolation, two features are re-unified to form a whole spectrum. As the spectrum contour looks like sliding along the formant loci, this new method is named as Spectrum Sliding on Formant Loci(SSFL).

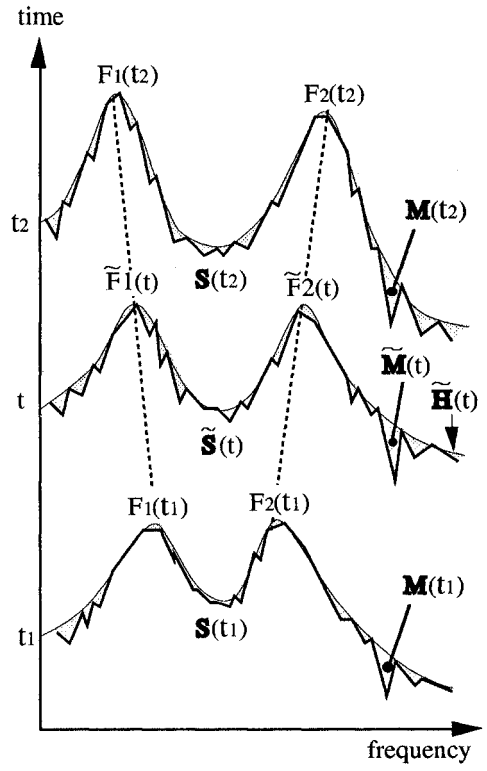

Fig. 2 Algorithm of SSFL

2.2 Algorithm

Fig. 2 shows the details of SSFL algorithm. SSFL algorithm has 5 steps as follows.

Firstly, calculate spectrum from given waves. Let $\mathbf{S}(t_1)$ and $\mathbf{S}(t_2)$ be the spectra at time t_1 and t_2 respectively.

Secondly, estimate formant frequencies and formant band widths from $\mathbf{S}(t_1)$ and $\mathbf{S}(t_2)$. And let $Fn(t_1)$ and $Fn(t_2)$ be the n-th formant frequency at time t_1 and t_2 respectively.

Thirdly, calculate interpolated formant frequency $\tilde{F}n(t)$ from $Fn(t_1)$ and $Fn(t_2)$ by the following equation.

$$\tilde{F}n(t)=\frac{(t_2-t)Fn(t_1)+(t-t_1)Fn(t_2)}{t_2-t_1} \qquad (2)$$

Fourthly, calculate spectrum $\tilde{\mathbf{S}}(t)$ at time t from $\tilde{F}n(t)$, $\mathbf{S}(t_1)$ and $\mathbf{S}(t_2)$. The spectrum envelope of $\tilde{\mathbf{S}}(t)$ is calculated by the following equation:

$$\tilde{H}(t,z^{-1})=\left|\prod_{i=1}^{n}\left(1-exp\frac{-\pi\tilde{B}n(t)+j2\pi\tilde{F}n(t)}{Fs}\cdot z^{-1}\right)\right|^{-1} \qquad (3)$$

where F_s, $\tilde{F}n(t)$ and $\tilde{B}n(t)$ are a sampling frequency, n-th formant frequency and n-th formant band width respectively.

The micro-structure of spectrum $\tilde{\mathbf{S}}(t)$ calculated by subtracting spectrum envelope from whole spectrum is estimated by the following equation:

$$\tilde{\mathbf{M}}(t)=\frac{(t_2-t)\mathbf{M}(t_1)+(t-t_1)\mathbf{M}(t_2)}{t_2-t_1} \qquad (4)$$

where $\mathbf{M}(t_1)$ and $\mathbf{M}(t_2)$ are the micro-structures of $\mathbf{S}(t_1)$and $\mathbf{S}(t_2)$, respectively.

The estimated spectrum $\tilde{\mathbf{S}}(t)$ can be calculated from $\tilde{\mathbf{H}}(t)$ and $\tilde{\mathbf{M}}(t)$ shown in Eq. (5).

$$\tilde{\mathbf{S}}(t)=\tilde{\mathbf{M}}(t)\cdot\tilde{\mathbf{H}}(t) \qquad (5)$$

Lastly, estimated spectrum $\tilde{\mathbf{S}}(t)$ is transformed into time domain through IFFT and added up to wave overlappingly.

2.3 Evaluation Experiment

We evaluate the proposed SSFL method by synthesizing continuous Japanese speech (/aiueo/, /oeuia/, /auieo/ and / oeiua/) consisting of vowels from isolated vowels uttered by a male announcer.

Spectra are estimated by PSE(Power Spectrum Envelop) analysis method and formant frequencies are estimated by peak picking method.

Fig. 3 and Fig. 4 show spectra of speech synthesized by the OLA method and the OLA with the SSFL method, respectively. In the speech synthesized by the OLA method, the formant loci are discontinuous at the boundaries of phonemes, while the formant loci of the speech synthesized by OLA with SSFL method are continuous.

In a listening test of speech /aiueo/ synthesized by the SSFL, neither noises nor deteriorations are detected at interpolated parts of speech. The smoothness at vowel boundaries is improved, especially the boundary between / u/ and /e/. However these are results of the primary experiment. We are currently preparing the detailed evaluation.

III. ANALYSIS OF SPECTRUM TRANSITION

Further improvement of quality depends on improving the quality of formant loci. It is well known that there are some relations between pitches and formant frequencies. However, when analyzing relations between pitches and formants, most researches deal with formants at stable parts of speech such as a center of vowel. To clarify how to control formant loci at a transition part of speech is necessary to improve the quality of synthesized speech. As the first step, we analyzed a duration required for formant transition on Japanese Vowel-/r/-Vowel syllables under two kinds of accent types:(1) putting accent at first vowel, (2) putting accent at final vowel.

3.1 Method of analysis

Fig. 5 shows the definition of the transition time. The transition period begins at the boundary of phoneme and finishes at the point where the slope of the 2nd formant locus has become flat. The boundary of phoneme and the end of formant transition are determined by the observation. In Japanese speech, there are 5 vowels, so the 25 VCV (C is /r/) syllables are analyzed. A speaker is the same male announcer used at the evaluation experiment of SSFL. Each syllable is uttered twice, and the formant transition duration is measured as an averaged duration length of two utterances.

3.2 Results

We calculate ratio of the duration for formant transition between syllables which differ in accent type. If the accent types have no relation with formant loci, the ratio of transition length would be 1.0. Fig. 6 shows the result of analysis. Especially at the syllables whose final vowels are front vowel (/i/, /e/), transition duration differs significantly by an accent type. The results of analysis indicate that the length for transition is related to an accent position. And syllables whose stress is put on the final vowel show a tendency to have longer transition duration than syllables whose stress is put on the first vowel.

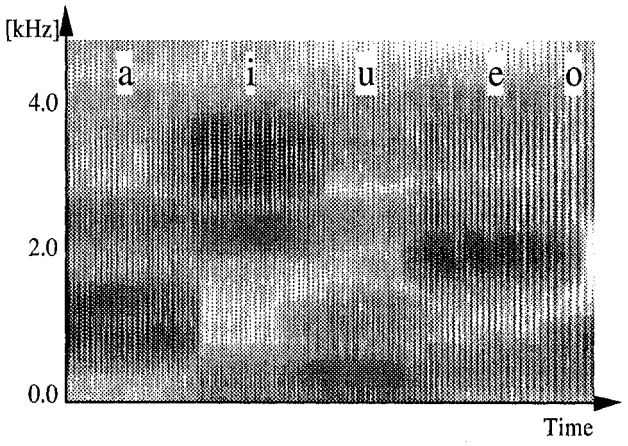

Fig.3 Synthesized speech by OLA method

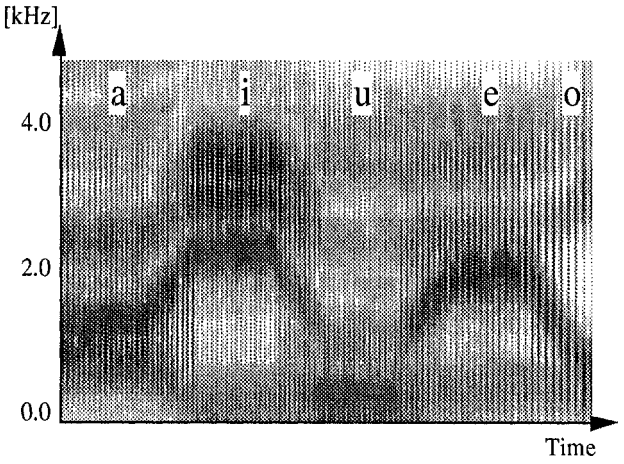

Fig.4 Synthesized speech by OLA with SSFL method

3.3 Evaluation of an influence of transition time

We evaluated the influence of transition time upon the quality of synthesized speech. We synthesized a syllable of /ari/ which has the transition time for different accent type. And we evaluate the deterioration of quality depending on formant transition duration. 5 speech researchers listened to the synthesized syllables twice with the different order of presentation for the evaluation.

When a syllable has an accent at the first vowel, the deterioration of quality depending on formant transition duration was not detected at all. When a syllable has an accent at final vowel, the deterioration of quality was detected at 90%.

This result shows that the transition time of formant at the phoneme boundary is important when the accent is positioned after the boundary.

IV. CONCLUSION

We have proposed a new spectrum interpolation method for improving quality of synthesized speech. We have evaluated the proposed SSFL method by synthesizing continuous Japanese speech /aiueo/. In the experiment, the formant loci are created by the linear interpolation precisely. The results have confirmed the continuity of formant loci and improvement of quality.

We have analyzed the duration length required for formant transition on Japanese Vowel-/r/-Vowel syllable in order to clarify how to construct rules that control formant loci at a phoneme boundary. And concluded that (1) the transition duration is related to an accent type, and (2) the difference in the transition duration affects synthesized speech in quality. Further improvement of quality needs more precise formant transition rules considering accent types.

REFERENCES

[1] K.Itoh, S. Nakajima and T.Hirokawa, A new waveform speech synthesis approach based on COC synthesis unit, Proc. Autumn Meet. Acoust. Soc. Jpn., pp.253-254, Oct. 1993

[2] T. Kamai and K. Matsui, Investigation of Formant Synthesizer Hybridized by Introduction of Natural Waveform Segments, Proc. Autumn Meet. Acoust. Soc. Jpn., pp.249-250, Oct. 1993

[3] K. Iwata, K. Takahashi, Y. Mitome and K. Nagano, Japanese Text-to-Speech Software for Personal Computers, Proc. Autumn Meet. Acoust. Soc. Jpn., pp.245-246, Oct. 1993

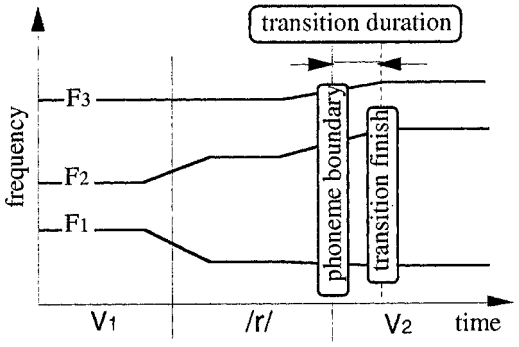

Fig.5 Definition of transition time

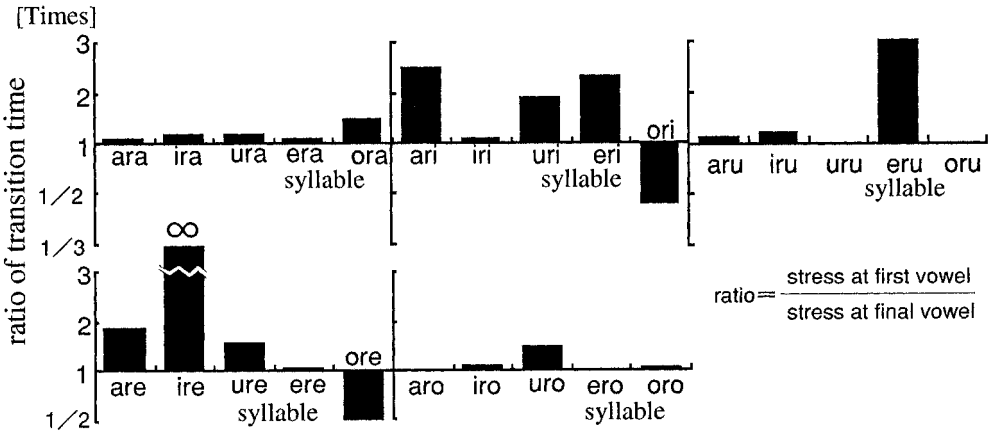

Fig.6 Ratio of transition time of stressed at first vs at final vowel