A VOICE CONVERSION METHOD AND ITS APPLICATION TO PATHOLOGICAL VOICES

H. Kuwabara[1] and T. Takagi[2]

1 The Nishi-Tokyo University, Yamanashi, Japan
2 NHK Science & Tech. Res. Labs., Tokyo, Japan

ABSTRACT

Formant and pitch frequencies are used as the acoustic parameters to be manipulated. These acoustic parameters are first extracted from a speech sound to be modified and changed according to some rules that are to make the original speech clear, and a new speech is synthesized using the modified acoustic parameters. Speech intelligibility is found to reach the maximum when the trajectories are emphasized to some extent. It is also found that our method is capable of improving the so-called "roughness" or "hoarseness" of pathological voices mainly by replacing pitch frequency of the original speech with that of a normal speaker.

1. INTRODUCTION

Using the analysis - synthesis system we have developed [1], voice quality of natural speech has been controlled by changing formant trajectories that are supposed to have a close relation to such voice qualities as intelligibility, clearness and so on. Correlation analysis between psychological and acoustic distances reveals that the formant trajectory has the largest correlation with the voice quality of the announcer's speech sounds, followed by pitch frequency [2]. This result suggests that the quality of speech sound of non-professional speakers may possibly be improved by altering the dynamics of formant trajectory patterns.

Based on the experimental evidence mentioned above, an experiment has been performed to change and improve the quality of natural speech making use of the analysis-synthesis system.

Formant trajectories are extracted first from voiced portions by LPC method and the dynamics of these trajectories are altered depending on the formant pattern itself. The method for altering the formant pattern is the same as that we have proposed earlier for the normalization of vowels in connected speech [3]. This method is applied to the formant and pitch trajectories extracted from natural speech, and the quality-controlled speech sounds are synthesized using the analysis-synthesis system to present to listeners for perceptual judgment.

2. ANALYSIS-SYNTHESIS SYSTEM

Fig. 1 illustrates the block diagram of the analysis - synthesis system. Low-pass filtered input speech was digitized in 12 bits at a rate of 15 kHz. A short time LPC analysis based on the autocorrelation method was performed to obtain LPC coefficients and the residual signals. Formant frequencies and their bandwidths were estimated by solving a polynomial equation. A modification of the spectral envelope is equivalent to a manipulation of the coefficients that would result in a frequency response of the filter equal to the modified envelope. These acoustic parameters (pitch periods, LPC coefficients, formant frequencies, bandwidths, residual signals) were stored for later synthesis.

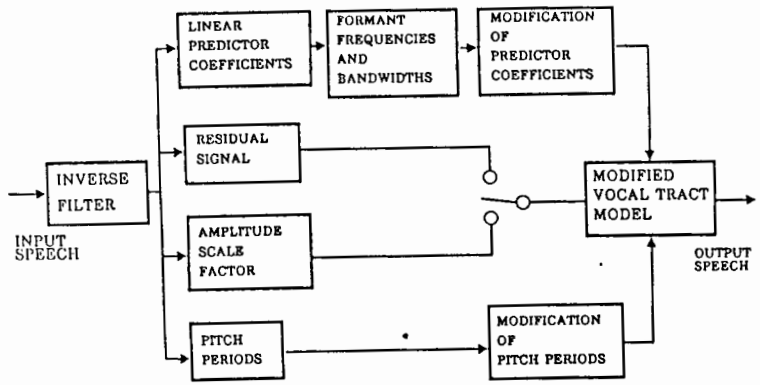

Fig.1 Block diagram of analysis-synthesis system for voice conversion

3. METHOD OF FORMANT TRAJECTORY MANIPULATION

After extracting formant trajectories using the method proposed by Kasuya [4], modification of them was conducted in such a way that the preceding and succeeding acoustic features contributed to the present value with the same weight if the time differences from the present were equal, and that the amount of contribution was proportional to the difference from the present acoustic feature [3]. Suppose x(t) be the time-varying pattern of a formant frequency, the new value y(t) is defined as the sum of the original value x(t) and the additional term of contribution by contextual information. The contribution is assumed to be a weighted sum of differences between values at the present time t and at different time t±τ

Thus, y(t) is given by

$$y(t)=x(t)+\int_{-T}^{T} w(\tau)(x(t)-x(t+\tau))d\tau \quad (1)$$

where w(τ) is the weighting function which is given as

$$w(\tau)=\alpha\cdot\exp(-\tau^2/2\sigma^2). \quad (2)$$

In this study, T=150ms and σ=52 ms were experimentally decided. Given α >0, the dynamics of the original formant trajectory is emphasized, while for α <0, it becomes deemphasized.

Equation (1) is applied to each of the three formant trajectories without vowel/consonant (except for voiceless consonant) distinction. The time interval in equation (1) during which the weighted sum is calculated is 300ms, a 150ms forward and backward each. This is the result for α = 7.3 which, in our previous study, represents a proper value for the purpose of normalizing coarticu-

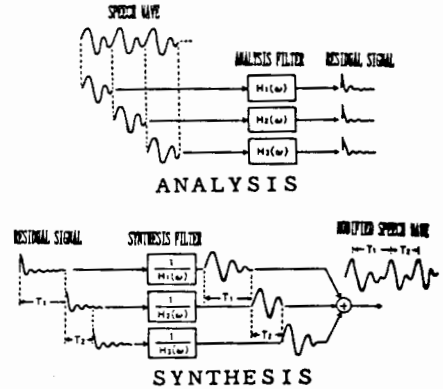

Fig.2 Schematic illustration for changing pitch frequency

lation effects of vowels in continuous speech. It is noticed from the figure that the new formant trajectories are emphasized their up-and-down dynamic movement as compared to those of the raw formants.

4. METHOD OF PITCH MANIPULATION

Pitch frequency manipulation is quite simple as depicted in Fig.2 At the pitch synchronous analysis stage, the residue signal obtained for each pitch period has exactly the same data length as the pitch period. If we give the residue signal as an input to the vocal tract model, exactly the same waveform as the original speech will be obtained. Thus, pitch frequency change can basically be given by controlling the length of the residue signal.
To raise pitch frequency,some data at the last part of the residue are eliminated and to lower the frequency, zero signals are added to the last part of the residue.

5. ENHANCEMENT OF PATHOLOGICAL SPEECH

An attempt has been performed to improve the quality of a pathological speech using the analysis-synthesis system we have developed. The pathological speech used in this experiment is a voice uttered by a patient who has a disease in his vocal cord. Because of malfunction of the vocal cord vibration, the resultant speech wave lacks clear periodicity and its voice quality is "hoarse". The experiment has been designed to create the fundamental frequencies into the pathological speech wave in order to improve the quality as close as normal speech.
Fig. 3 represents the block diagram to improve the quality of pathological speech. It requires two kinds of input speech : a pathological speech to be improved and a normal speech utterance of the same sentence from another speaker. From the pathological speech inputted, voiced portions are at first detected and the spectral envelopes are extracted by LPC analysis. Next, the normal speech is analyzed by the same method and the pitch frequencies are detected to combine with the spectral information extracted from the pathological speech. If the normal speech of the same content can not immediately be available, artificial pulse trains could be used as a voice source.
In the analysis stage, after making voiced/voiceless distinction, the voiceless portions (voiceless consonants and devocalized vowels) are thoroughly kept in memory and the LPC analysis is performed for the voiced portions to obtain LPC coefficients that carry spectral information and the residual signals from which pitch periods can be estimated. For the pathological speech, the frame length (analysis window) is set at 20 ms and the frame shift is a half the window length.
In the feature extraction stage, the residual signals for the pathological speech are discarded after obtaining spectral information. Contrary to this, only the pitch frequency contour is needed from the normal speech.
For the normal speech, however, a process of time alignment has

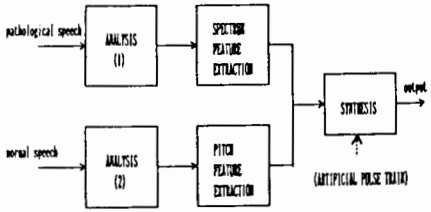

Fig.3 Block diagram for the enhancement of pathological speech

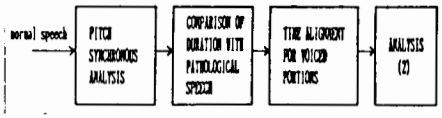

Fig.4 Block diagram of analysis and time alignment for normal speech

been undertaken before feeding to analysis in Fig. 3. This process is shown in Fig. 4. The voiced parts of the normal speech are analyzed pitch synchronously and the length for each part is compared with the corresponding part for the pathological speech in order to make the length equal to that of the pathological speech with accuracy of less than one pitch period. This has been done simply by eliminating or inserting additional pitch periods.
The normal speech, after being time-aligned, is LPC analyzed again and the pitch frequencies are extracted for every voiced portion. This pitch frequencies or the residual signals are fed into the synthesis filter as the voice source. The synthesis filter is made from the predictor coefficients obtained from the pathological speech. The resultant output speech has, therefore, the same spectral characteristics as the pathological speech and the same source characteristics as the normal speech. Fig. 5 depicts an example of speech waveforms for the pathological speech ,synthesized speech by the proposed method and also synthesized speech with an artificial pulse train as the voice source to the filter.
As far as we have tested, the quality of the synthesized speech is has been found to be far better than the original pathological speech, though it is not as good as the normal speech sound.

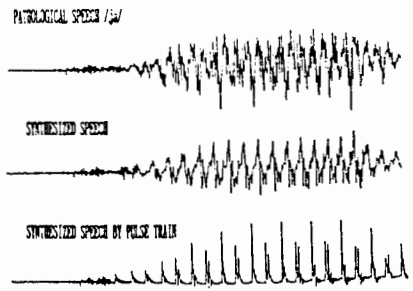

Fig.5 An example of speech waveforms of the pathological and synthesized speech

6. CONCLUSION

Improvement of voice quality has been performed using an analysis-synthesis system capable of modifying pitch, formant frequencies, and formant bandwidths. According to the results of analysis for professional announcers speech sounds, it is obvious that speech intelligibility closely relates to the dynamics of formant and pitch patterns. It has been found to be possible to improve the speech intelligibility without changing voice individuality by emphasizing the movement of time-varying pitch pattern. Another application of this analysis-synthesis system has also been made to enhance a pathological speech which has little periodicity and "hoarse" in voice quality. By adding fundamental frequency component taken from a normal speaker, the voice quality of the pathological speech has been improved to a great extent.

7. REFERENCES

[1] H. Kuwabara, "A pitch synchronous analysis / synthesis system to independently modify formant frequencies and bandwidth for voiced speech,"SPEECH COMMUNICATION, Vol. 3 (1984) pp.211-220
[2] H. Kuwabara, K. Ohgushi, "Acoustic characteristics of professional male announcers' speech sounds," ACUSTICA, Vol.55 (1984) PP.233-240
[3] H. Kuwabara, "An approach to normalization of coarticulation effects for vowels in connected speech," J. Acoust. Soc. Amer., Vol. 77 (1985) pp.686-694
[4] H. Kasuya, "An algorithm to choose formant frequencies obtained by linear prediction analysis method," Trans. IECE Japan, Vol. J66-A (1983) pp.1144-1145