A METHOD FOR CHANGING SPEECH QUALITY AND ITS APPLICATION TO PATHOLOGICAL VOICES

Hisao Kuwabara

Teikyo University of Science & Technology, Uenohara, Kitatsuru-gun, Yamanashi 409-01, Japan
Tel: (0554)63-4411, Fax: (0554)63-4431, E-mail: kuwabara@ntu.ac.jp

***Abstract*: Speech quality is an interesting and very important aspect not only from linguistic and/or phonetic viewpoint but also from the viewpoint of speech technology. This study has been conducted from the latter point of view. A method has been developed based on the analysis-synthesis technique which enables to control voice quality by independently manipulating the voice source and the vocal tract resonant characteristics. Through this method, it is possible to investigate the amount of contribution of individual acoustic parameters to a certain voice quality including voice individuality. Formant frequencies and their bandwidths are used as the acoustic parameters to characterize the vocal tract configuration and the pitch frequency as the voice source. These acoustic parameters extracted from a natural speech are modified or changed to some extent and then a is synthesized making use of the modified acoustic parameters. Speech intelligibility and voice individuality are found to be controlled by this method. An application to a pathological voice has also been made to control the voice quality. It has been found that the method is capable of improving the so-called "roughness" or "hoarseness" of the pathological voice to a certain extent.**

I. 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 with such voice qualities as intelligibility, clearness, articulateness, and so on. Correlation analysis between psychological and acoustic distances reveals that the formant trajectory has the largest correlation with the voice quality of 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 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 coarticu-lated vowels in continuous speech. [3]. This method is applied to the formant and pitch trajectories extracted from a natural speech, and the quality-controlled speech sounds are synthesized using the analysis-synthesis system to present to listeners for perceptual judgments.

II. ANALYSIS-SYNTHESIS SYSTEM

Fig. 1 illustrates the block diagram of the analysis-synthesis system. Low-pass filtered input speech is digitized in 12 bits at a rate of 15 kHz. A short time LPC analysis based on the autocorrelation method is performed to obtain LPC coefficients and the residual signals. Formant frequencies and their bandwidths are 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) are stored for later synthesis. This analysis-synthesis system is capable of analyzing input speech either pitch synchronously or non-synchronously dependent on the type of input speech and the aim of speech analysis or synthesis. When we analyze a pathological speech, as described in a later section, which is difficult or impossible to define pitch periods from the input speech, analysis is generally performed with a fixed frame length.

III. METHOD OF FORMANT TRAJECTORY MANIPULATION

Models and analysis of vocal emissions for biomedical applications. 4th international workshop. October 29-31, 2005 – Firenze, Italy.
Edited by C. Manfredi. ISBN 88-8453-320-1 (online)

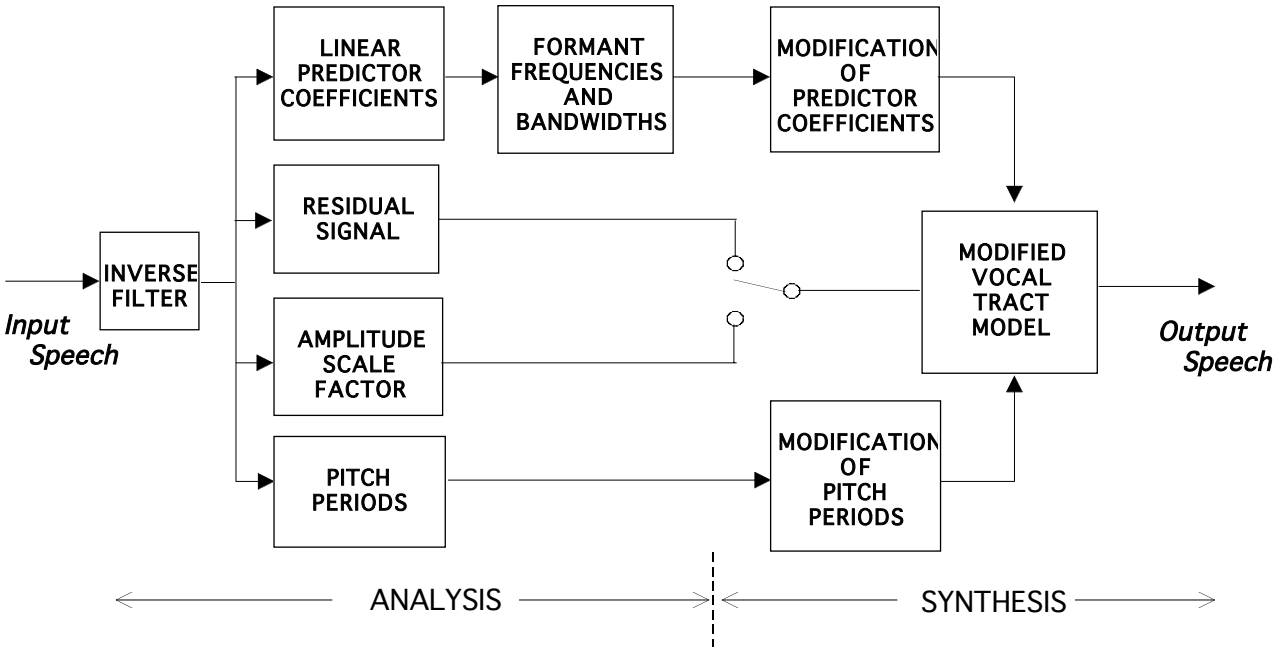
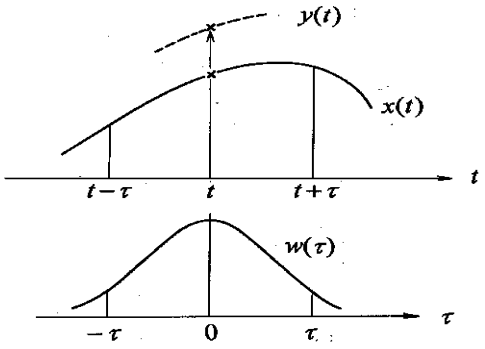


Fig. 1 *Block diagram of analysis-synthesis system for voice conversion.*

The method of formant modification has already been reported in an article [4]. The outline of the method is the following. For each pitch period, formant frequencies and their bandwidths are calculated first by solving the polynomial equation. Then, some modification is made on the original formant frequencies and/or bandwidths and accordingly on the predictor coefficients. A synthesis filter (vocal tract resonance filter) is formed using the modified coefficient. The residual signal has also been used as the input to this filter.

After extracting formant trajectories using the method proposed by Kasuya [5], modification of them has been conducted in such a way that the preceding and succeeding acoustic features contribute to the present value with the same weight if the time differences from the present are equal, and that the amount of contribution is proportional to the difference from the present acoustic feature [3]. This process is illustrated in **Fig. 2**. 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 $t \pm \tau$. Thus, $y(t)$ is given by,



$$y(t) = x(t) + \int w(\tau) \cdot \{x(t) - x(t-\tau)\} d\tau$$

$$w(\tau) = 7.3 \cdot \exp\left\{ -\frac{\tau^2}{2 \cdot (0.52)^2} \right\}$$

Fig. 2 *Graphic illustration for using time-varying dynamic pattern of acoustic feature.*

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

where $w(\tau)$ is the weighting function which is given as

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

The time interval (-T, T) from which the contextual information should be taken into account is theoretically infinite. But actually it must take a finite number and is not determined theoretically but is decided empirically or experimentally. In this study, $T = 150\,ms$ and $\sigma = 52\,ms$ have been experimentally determined. Given $\alpha > 0$, the dynamics of the original formant trajectory is emphasized, while for $\alpha < 0$, it becomes de-emphasized.

Equation (1) is applied to each of the three formant trajectories without vowel/consonant distinctions except for voiceless consonant. The time interval in equation (1) during which the weighted sum is calculated is 300 ms,

a 150 ms forward and backward each. This is the result for $\alpha = 7.3$ which, in our previous study, represents a proper value for the purpose of normalizing coarticulation effects of vowels in continuous speech. It is noticed from the figure that the new formant trajectories are emphasized their up-and-down dynamic movements as compared to those of the raw formants.
As far as we have tested, this method of incorporating contextual information is capable of not only normalizing the coarticulation effect of vowels in continuous speech but also has some advantage for vowel recognition using the formant frequencies with conventional Euclidean distance from the reference vowel. The method is also capable of improving the intelligibility of some lazily spoken continuous speech.

IV. METHOD OF PITCH MANIPULATION

Pitch frequency manipulation is quite simple as described in **Fig.3**. 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

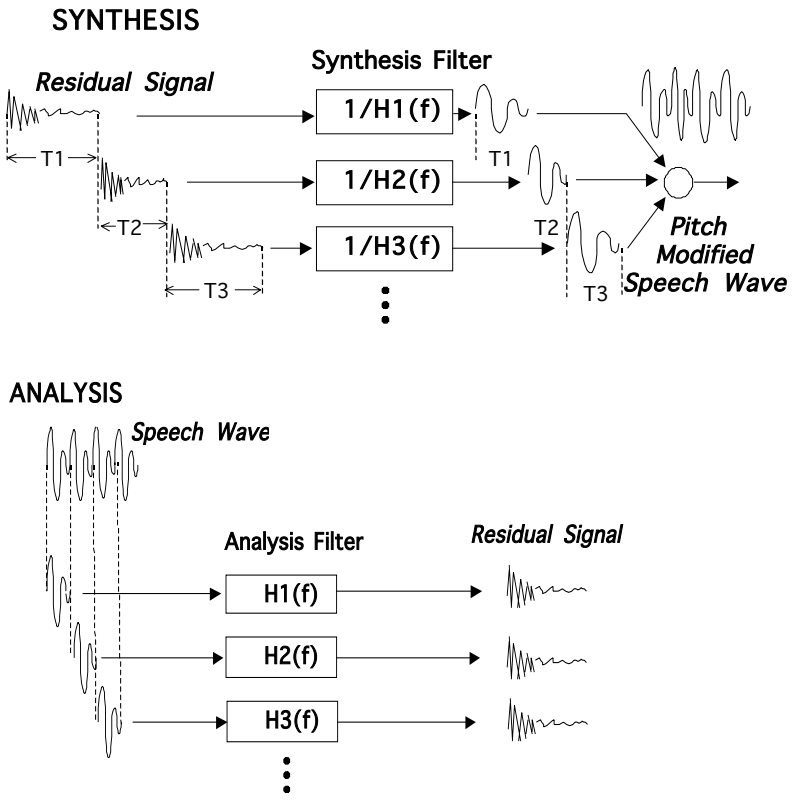


Fig.3 *A method of manipulating fundamental frequency.*

will be obtained. Thus, pitch frequency change can basically be given by controlling the length of the residual 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.
Of course there are some discrepancies on the frequency domain between the pitch-modified speech and the original speech. However, there is no serious voice change in terms of perceptual voice quality when the pitch frequency change is less than 50% form the original.

V. ENHANCEMENT OF PATHOLOGICAL SPEECH

An attempt has been made 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 waves in order to improve the quality as close to a normal speech as possible.
Fig. 4 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 through 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 the 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 the 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 of 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 been undertaken before feeding to analysis in Fig.4. This process is shown in **Fig.5**. The voiced parts of the normal speech are analyzed pitch synchronously and the length for each part

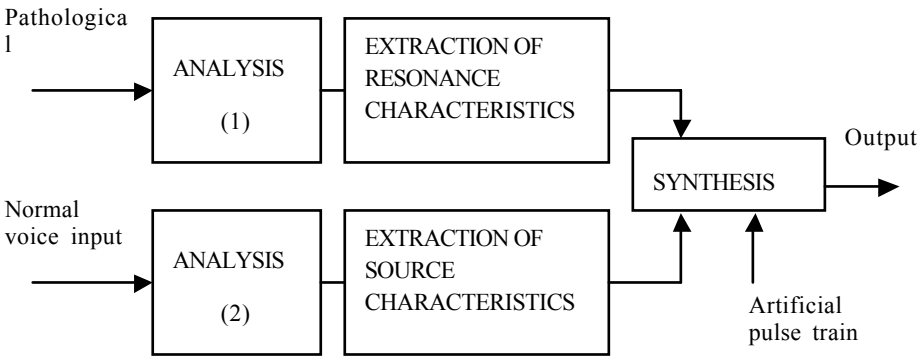


Fig. 4 *Block diagram of speech analysis for improving pathological voice.*

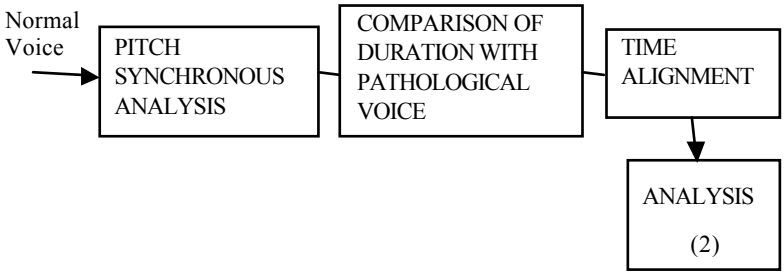


Fig. 5 *Time alignment process with a normal speech voice.*

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. The 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. As far as we have tested, the quality of the synthesized speech has been found to be far better than the original speech, though it is not as good as the normal speech.

VI. CONCLUSIONS

Improvement of voice quality has been achieved 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 patterns. 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.

REFERENCES

[1] H. Kuwabara, (1984) "A pitch synchronous analysis/synthesis system to independently modify formant frequencies and bandwidths for voiced speech," SPEECH COMMUNICATION, Vol.3, pp.211-220

[2] H. Kuwabara, and K. Ohgushi, (1984) "Acoustic characteristics of professional announcers' speech sounds," ACUSTICA, Vol.55, pp.233-240

[3] H. Kuwabara, (1985) "An approach to normalization of coarticulation effects for vowels in connected speech," J. Acoust. Soc. Am., Vol.77, pp.686-694

[4] H. Kuwabara, and K. Ohgushi, (1987) "Contributions of vocal tract resonant frequencies and bandwidths to the personal perception of speech," ACUSTICA, Vol.63, pp.120-128

[5] H. Kasuya, (1983) "An algorithm to choose formant frequencies obtained by linear prediction analysis method," Trans. IECE Japan, Vol.J66-A, pp.1144-1145