QUALITY CONTROL OF SPEECH BY MODIFYING FORMANT FREQUENCIES AND BANDWIDTHS

Hisao Kuwabara

ATR Interpreting Telephony
Research Laboratories
Twin 21 Bldg. MID Tower
2-1-61 Shiromi Higashi-ku
Osaka 540 Japan

Tohru Takagi

NHK Science and Technical
Research Laboratories
1-10-11 Kinuta Setagaya-ku
Tokyo 157 Japan

ABSTRACT

An analysis-synthesis system which is capable of independent manipulation of acoustic parameters has been developed to investigate the contribution of these individual parameters to the speech quality. Formant frequencies and their bandwidths were used as the acoustic parameters to characterize the vocal tract configuration, and pitch frequency as the voice source. This paper describes a way how to control the voice quality of natural speech by manipulating the formant frequencies. Formant trajectories extracted from a natural speech were modified to alter their up-and-down oscillation to some extent, and the resultant speech wave was synthesized by the above mentioned method to present to listeners for the judgment of voice quality. It was found that the speech intelligibility was improved to some extent when the movement of time-varying formant pattern was slightly emphasized, but too much emphasis would cause degradation of the voice quality.

1. INTRODUCTION

This paper deals with a way of controlling the voice quality of natural speech. An analysis-synthesis method has been developed which is capable of independent manipulation of such acoustic parameters as formant frequencies, their bandwidths and pitch frequency [1]. Using this system, 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.

According to our previous study [2], acoustic characteristics of professional announcers speech, which is considered to be the most intelligible or the clearest, lies in the dynamics of pitch and formant frequencies. The dynamic range of these features for the announcers speech is significantly large compared to that for the non-professional speakers. 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. 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 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

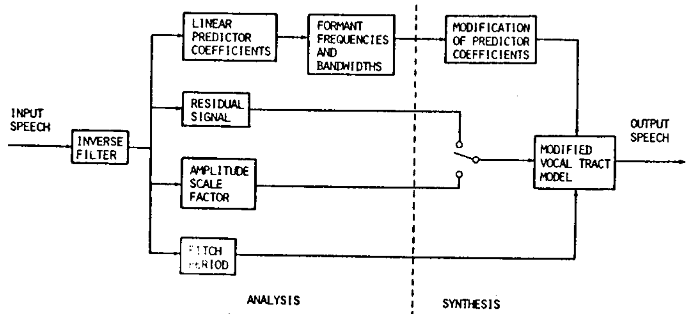

Fig. 1 Block diagram of the analysis-synthesis system to modify formant frequencies.

(pitch periods, LPC coefficients, formant frequencies, bandwidths, residual signals) were stored for later synthesis.

Let $z_i = r_i \exp(j\omega_i)(i = 1,2.....,p)$ stand for the roots corresponding to the formants to be changed. Formant frequencies and/or their bandwidths are modified by changing related angular frequencies ω_i and/or the factors r_i. LPC coefficients are modified so that the modified poles $\tilde{z}_i$ become the roots of a new polynomial,

$$z^p + \tilde{a}_1 z^{p-1} + \ldots + \tilde{a}_{p-1} z + \tilde{a}_p = 0. \qquad (1)$$

Calculation of $\tilde{a}_i(i=1, 2, \ldots, p)$ is performed simply by comparing terms of the same order on both sides of the following equation,

$$(z-\tilde{z}_1)(z-\tilde{z}_2)\ldots(z-\tilde{z}_p) = z^p + \tilde{a}_1 z^{p-1} + \ldots + \tilde{a}_{p-1} z + \tilde{a}_p . \qquad (2)$$

The modified vocal tract model $\tilde{V}(z)$ is then given by,

$$\tilde{V}(z) = 1 \,/\, (1 + \sum_{i=1}^{p} a_i z^{-i}) \qquad (3)$$

where $\{\tilde{a}_i\}$ are the solutions of equation (2). The modified vocal tract model $\tilde{V}(z)$ has the desired frequency characteristics. If the spectral manipulation is too large, some discontinuities are found to occur at the boundary of each frame, which eventually cause a typical buzzing. To cope with this, a simple time domain manipulation has been performed. In this experiment, half the analysis window is set as the period of frame shift. Output speech wave from the modified vocal tract model $\tilde{V}(z)$ for each frame is multiplied by a triangular time window. The amplitude of this triangular window is composed so that the sum of the gain at any instant within the overlapped portion between two successive frames becomes 1. The resultant speech is obtained by adding

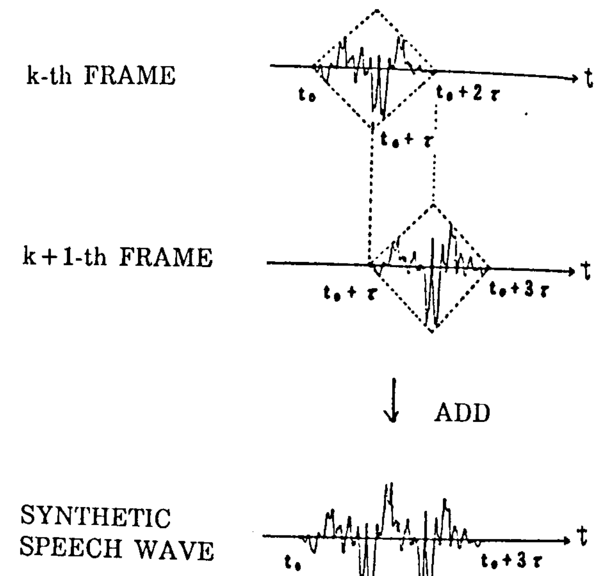

Fig. 2 Method of obtaining high quality synthetic speech

successively speech waves between adjacent two frames. This process is illustrated in Fig. 2.

3. ALTERATION OF FORMANT TRAJECTORY

3.1 Formant Trajectory Extraction

Low-pass filtered input speech was digitized at a rate of 15kHz, and the linear predictive analysis was made to find formant frequencies. Autocorrelation method for the inverse filter was adopted with order 14, the analysis window 20 ms, and the frame period 10 ms. Silent intervals and voiced/voiceless distinctions were made based on the speech power and the first order PARCOR coefficient, respectively. Formant frequencies for each frame were extracted from a set of 7 poles using the method proposed by Kasuya et al [4] and a smoothing was made by averaging formant data over three consecutive frames.

3.2 Formant Trajectory Change

Modification of formant trajectory 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]. This process is illustrated in Fig. 3. Suppose the curve $x(t)$ be an actual 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

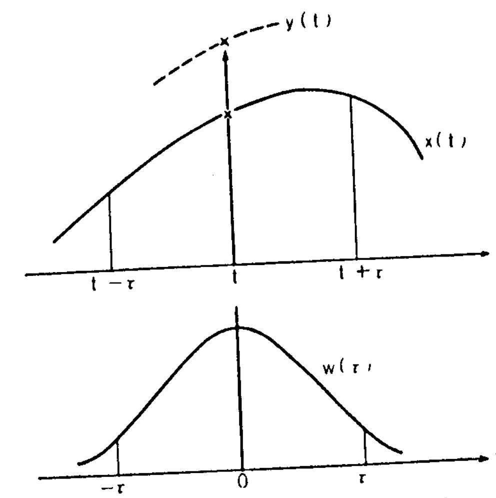

Fig. 3 Illustration of how to change the formant trajectory.

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

$$y(t) = x(t) + \int_{-T}^{T} w(\tau)(x(t)-x(t+\tau))d\tau \qquad (4)$$

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

$$w(\tau) = \alpha \cdot \exp(-\tau^2/2\sigma^2). \qquad (5)$$

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

Equation (4) is applied to each of the three formant trajectories without vowel/consonant (but except voiceless consonant) distinction.

4. PERCEPTION OF QUALITY-CONTROLLED SPEECH

As described in the previous section, dynamic movement of formant frequency is one of the most important acoustic factors that characterize clear and intelligible voice. Change of voice quality to improve the clearness of intelligibility should, therefore, be done by modifying the dynamics of formant frequency. In this section, we describe a perceptual experiment on voice quality for formant-modified speech.

4.1 Synthesis Method

Based on the analysis-synthesis system described above, several formant-modified speech signals were obtained to present to listeners for quality judgment. Following is the process of speech synthesis.

(1) Speech waves are digitized with 15kHz sampling rate and 12bits accuracy. Analysis is made based on the system shown in Fig. 1 with a 20ms analysis frame multiplied by the Hamming window and 10ms frame period. The orders for analysis are 14 for male and 10 for female voices, and the predictor coefficients and the residual signals for each frame are stored.

(2) Formant frequencies of the first three are calculated from the predictor coefficients for each frame, and their trajectories over the entire word are estimated using a tracking algorithm [4].

(3) Equation (4) is applied independently to each formant trajectory and new frequencies down to each frame are calculated, and the resultant new coefficients are obtained by the method described in section 2. However, formants higher than the fourth and voiceless consonants remain unchanged.

(4) A vocal tract model is formed using the new predictor coefficients, given in equation (3), and the speech signals are obtained by inputting the residual signals to the model.

A nonsense word /a o i u e/ which consists of a concatenation of five Japanese vowels was used as the speech material. As mentioned before, some discontinuity would occur at the boundary between two successive frames if we simply connect speech signals from the two frames without overlap, which may cause degradation of speech quality. Fig. 2 shows a method how to avoid this sort of degradation.

In equation (5), constant α represents a scale factor which controls the amount of formant modification when it is applied to a formant trajectory as in equation (4). The dynamic pattern of formant movement is emphasized for α being positive, unchanged for α=0, and de-emphasized for negative value. Fig. 4 represents an example of formant trajectories of a speech sample used in the perceptual experiment before and after applying Eq.(4).

4.2 Result of Perceptual Experiment

The above mentioned nonsense word was used as the speech material and two speakers, male and female each, read the word with a normal speed. Seven different values, ranging from -15.3 to 15.3 including zero were selected as the factor α to

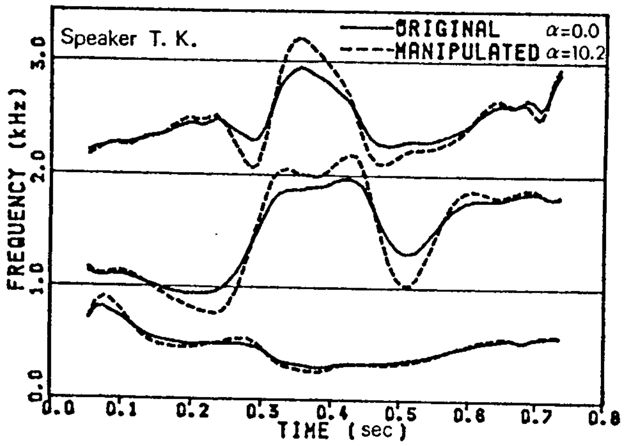

Fig. 4 An example of formant trajectory modification: original(solid lines) and modified(dashed lines).

get synthetic speech samples to be examined. Five female listeners, who never heard the speakers voices before, participated in the experiment. For each speaker, seven speech samples were paired and the listeners were asked to judge which one of a dyad sounded more intelligible or clear by comparison.

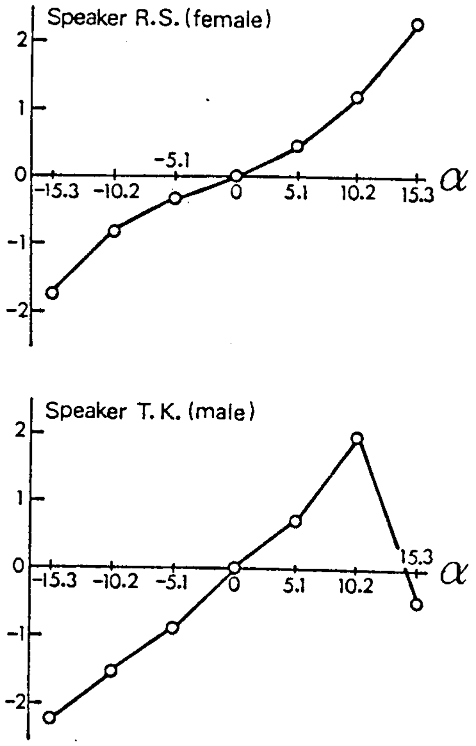

Fig. 5 Results of perceptual experiment on voice quality for male and female speakers.

Fig. 5 shows the result for speech samples of each speaker. The abscissa represents the factor α and the ordinate is a psychological distance. This distance is similar to JND (Just Noticeable Difference) distance, and 1 means that the perceptual difference between the two stimuli is greater than 50 percent chance level.

Being α =0 the reference of comparison, the results show that, in general, the voice quality becomes intelligible as the factor α increases. For male speaker's voice, however, it goes maximum when α =10.2 and goes down rapidly for larger α. This speaker dependency is caused by the degradation of quality by emphasizing the frequency movement too much and partially losing the phonetic quality.

In general, voice quality was found to be improved for the factor somewhere between 5 to 10. The factor greater than 10, however, sometimes gives the speech an improved quality but sometimes degraded quality depending on speakers.

CONCLUSIONS

Time-varying dynamic pattern of formant frequencies which is the main factor to contribute to the clearness or intelligibility has been modified using an analysis-synthesis system and perceptual experiment has been performed on the voice quality. It was found that the voice quality was improved to some extent when the dynamics was properly emphasized.

References

[1] H. Kuwabara, "A pitch synchronous analysis/synthesis system to independently modify formant frequencies and bandwidths 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 et al, "An algorithm to choose formant frequencies obtained by linear prediction analysis method," Trans. IECE Japan Vol.J66-A (1983) pp.1144-1145