TR-I-0023

Quality Control of Speech by Modifying Formant Frequencies and Bandwidths

ホルマント周波数、バンド幅の変形による声質制御

Hisao Kuwabara

桑原尚夫

1988· 3

Abstract

An analysis-synthesis system has been developed which is capable of independent manipulation of acoustic parameters to investigate the contribution of these individual parameters to 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 method 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 movement to some extent, and the resultant speech wave was synthesized by this method to present to listeners for the judgment of voice quality. It was found that the speech intelligibility or articulateness was improved to some degree when the movement of time-varying formant pattern was slightly emphasized, but too much emphasis would cause degradation of the voice quality.

ATR Interpreting Telephony Research Laboratories
ATR自動翻訳電話研究所

1. Introduction

It is well known that speech waves carry two kinds of informations as illustrated in Fig. 1. One is 'linguistic information' which has the primary importance in communicating each other and the other is 'non-linguistic information' which is usually neglected and occasionally becomes stumbling block as long as speech technologies are concerned. Historically, a lot of emphasis is placed on the linguistic aspect of speech, but the present study is concerned with the other aspect of speech, especially, with the articulateness of speech sounds.

This paper deals with a method of controlling the voice quality of natural speech. An analysis-synthesis technique 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, articulateness, 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 articulate, lies in the dynamic aspects 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 announcers 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 benn 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 then 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 continuous speech [3]. This method is applied to the formant trajectories extracted from natural speech, and the quality-controlled speech are synthesized using the analysis-synthesis technique to present to listeners for judgment of quality.

2. Characteristics of Professional Announcers Speech

Before proceeding to the analysis-synthesis system we used for this study, let take a brief look at the results of an experiment we have done before concerning the acoustic and psychoacoustic characteristics of professional announcers speech. Speech samples from 10 male announcers and 5 non-professional male speakers were collected for this experiment.

First, perceptual experiment was performed based on similarity judgment by paired comparison. Listeners were asked to rate each dyad on a 5-point scale: 1 quite similar, 2 similar, 3 uncertain, 4 different, and 5 quite different. The experimental data were analyzed by coventional MDS method developed by Kruskal. Fig. 2 represents the result of two dimensional solution of this analysis. Stimulus numbers from 1 to 10 stand for the announcers' speech and those from 11 to 15 stand for non-professionals. This figure depicts relative psychological distance of speech quality between individuals. The closer the distance between points, the more similar their voice quality. It is obvious that the announcers' speech sounds are very close to each other compared to the other five speakers and they are clearly separated by the vertical axis, indicating that the horizontal axis is the key factor for separation. The psychological constituent is the articulateness of pronounciation, and that of the vertical axis corresponds roughly to tone-hight (pitch). From this experiment, it was revealed that the articulateness was the most important factor to distinguish the quality of professional announcer's speech sound from that of non-professional speakers.

Next, a correlation analysis was performed to find acoustic features to which the quality of announcer's speech sounds is closely related. Four acoustic parameters, (1) pitch contour and (2) mean pitch frequency, as source characteristics (3) formant trajectory and (4) mean spectrum envelope, as spectral characteristics, were considered. Correlation analysis between psychological distance which is represented by Fig. 2 and acoustic distance which is calculated using each of the four acoustic feature has been made. The result is shown in Table 1. Formant trajectory was found to have the highest correlation with the distribution of psychological distance followed by pitch contour. In other words, dynamic aspect of acoustic features plays the most important role in the articulateness of speech sound.

3. Analysis-Synthesis System

Fig. 3 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 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.

Let $z_i = r_i \exp(j\omega_i)$, $(i = 1, 2, \cdots, p)$, 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 z'_i become the roots of a new polynomial,

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

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

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

The modified vocal tract model V'(z) is then given by,

$$V'(z) = 1 / (1 + \sum a_i z^{-i}) \qquad (3)$$

where $\{a'_i\}$ are the solutions of equation (2). The modified vocal tract model 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. The output speech wave from the modified vocal tract model 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 gains at any instant between two frames is always kept 1. The resultant speech is obtained by adding successively speech waves between adjacent two frames. This process is illustrated in Fig. 4.

4. Alteration of Formant Trajectory

In this section, we describe a way of formant tarjector extraction from natural speech and a method of changing the dynamics of their movement. The method of changing formant trajectories is exactly the same as that we used in the study of vowel normalization for coarticulation [3].

4.1 Formant trajectory extraction

Low-pass filtered input speech was digitized at a rate of 15 kHz, 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 the voiced/voiceless distinctions were made based on the speech power and the first order PARCOR coefficients, 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.

4.2 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 differences from the present were equal, and that the amount of contribution was proportional to the difference from the present acoustic feature. This process is illustrated in Fig. 5. 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 additional term of contribution by contextual information. The contribution was assumed to be a weighted sum of differences between values at the present time t and at different time t±τ. Thus, the new value y(t) is given by

$$y(t) = x(t) + \int w(\tau)\cdot\{x(t)-x(t+\tau)\}\, d\tau \qquad (4)$$

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

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

In this study, constants, $T=150$ ms and $\sigma=52$ ms, were decided based on some perceptual experiment we have done before [5]. 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. Fig. 6 represents an example of changing formant dynamics in a short Japanese sentence “a no hi to wa i a i nu ki no me i ji N de su.” Small circles represents the row data of the first three formants extracted by the LPC analysis and crosses are the new formant frequencies calculated according to the equation (4) giving α a positive value ($\alpha=7$). It is observed that the movement of formant trajectories is emphasized to some extent.

5. 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 articulate voice. Change of voice quality to improve the cleaness or articulateness 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.

5.1 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 15 kHz sampling rate and 12 bits accuracy. Analysis is made based on the system shown in Fig. 1 with a 20ms analysis frame multiplied by the Hamming window and 10 ms 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 3. 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. 4 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 positive value α, unchanged for $\alpha=0$, and de-emphasized for negative α. Fig. 7 represents an example of formant trajectories of a speech sample used in the perceptual experiment before and after applying equation (4).

5.2 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 get synthetic speech samples to be examined.

Five female listeners, who never heard the speakers voice before, participated in the experiment. For each speaker, seven speech samples were paired and the listeners were asked to judge which one, first stimulus or second one, sounded more clear or articulate by comparison.

Fig. 8 shows the result for speech samples of each speaker. The abscissa represents the factor α and the ordinate is a kind of psychological distance. This distance is is similar to JND (Just Noticiable Difference) distance, and 1 means that the perceptual difference between the two stimuli is greater than 50 percent chance level. The listeners never heard the voices of speakers in Fig. 8 (a), but they know very well about those in (b).

Being $\alpha=0$ the reference of comparison, the results show that, in general, the voice quality becomes articulate as the factor α increases. For male speaker's voice, however, it goes maximum when $\alpha=10.2$ 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 articulateness 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 vooiced speech," SPEECH COMMUNICATION, Vol. 3 (1984) pp.211-220

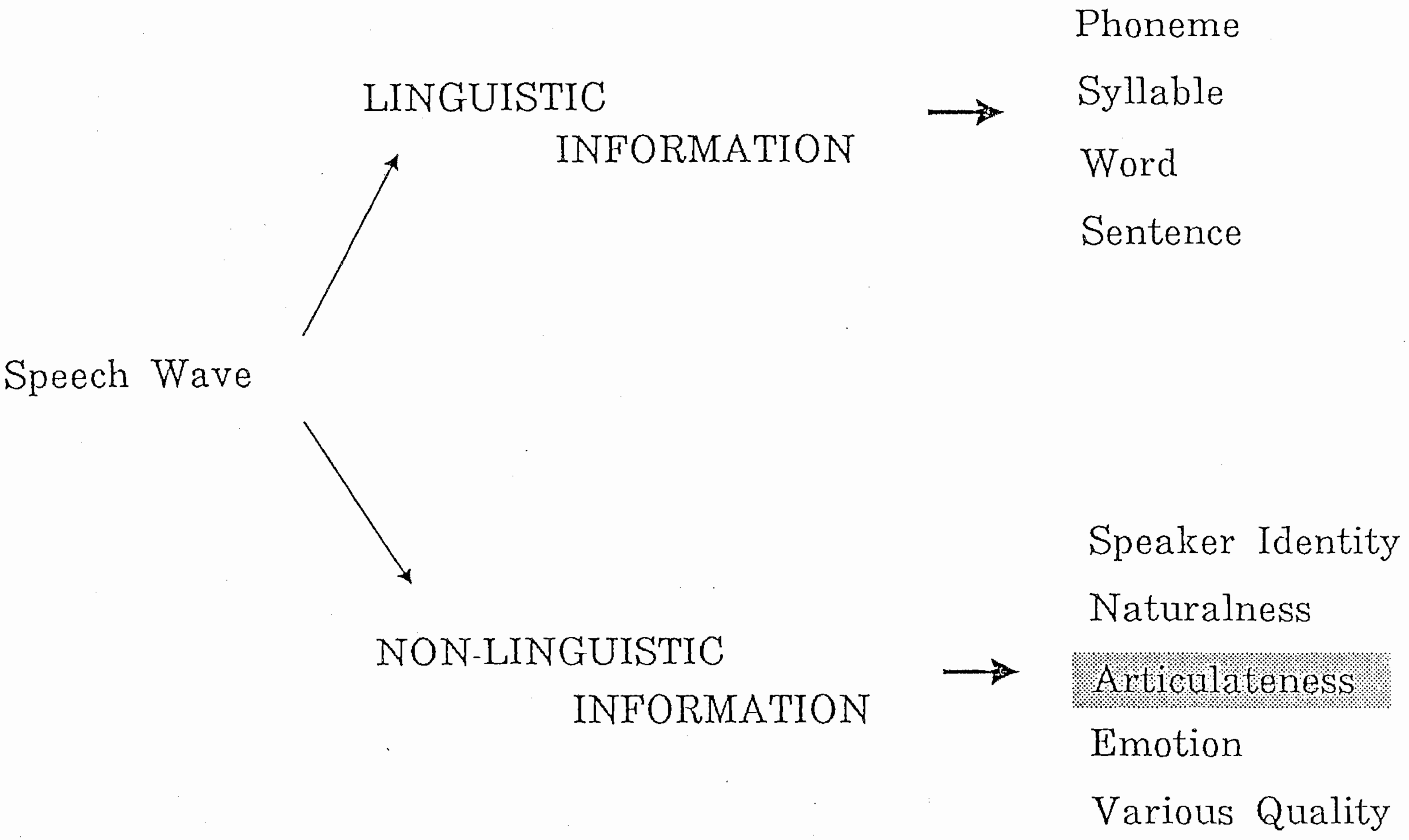
[2] H. Kuwabara and K. Ohgushi, "Acoustic characteristics of professional male announcers' speech sounds," ACUSTICA, Vol. 55 (1984) pp.233-240

[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

[5] H. Kuwabara and H. Sakai, “Perception of vowels and CV-syllables segmented from connected speech,” J. Acoust. Soc. Japan, Vol. 28 (1972) pp.225-234

INFORMATION CONVEYED BY SPEECH SIGNAL



Speech Wave

LINGUISTIC INFORMATION

Phoneme

Syllable

Word

Sentence

NON-LINGUISTIC INFORMATION

Speaker Identity

Naturalness

Articulateness

Emotion

Various Quality

Fig. 1 Information conveyed by speech signal

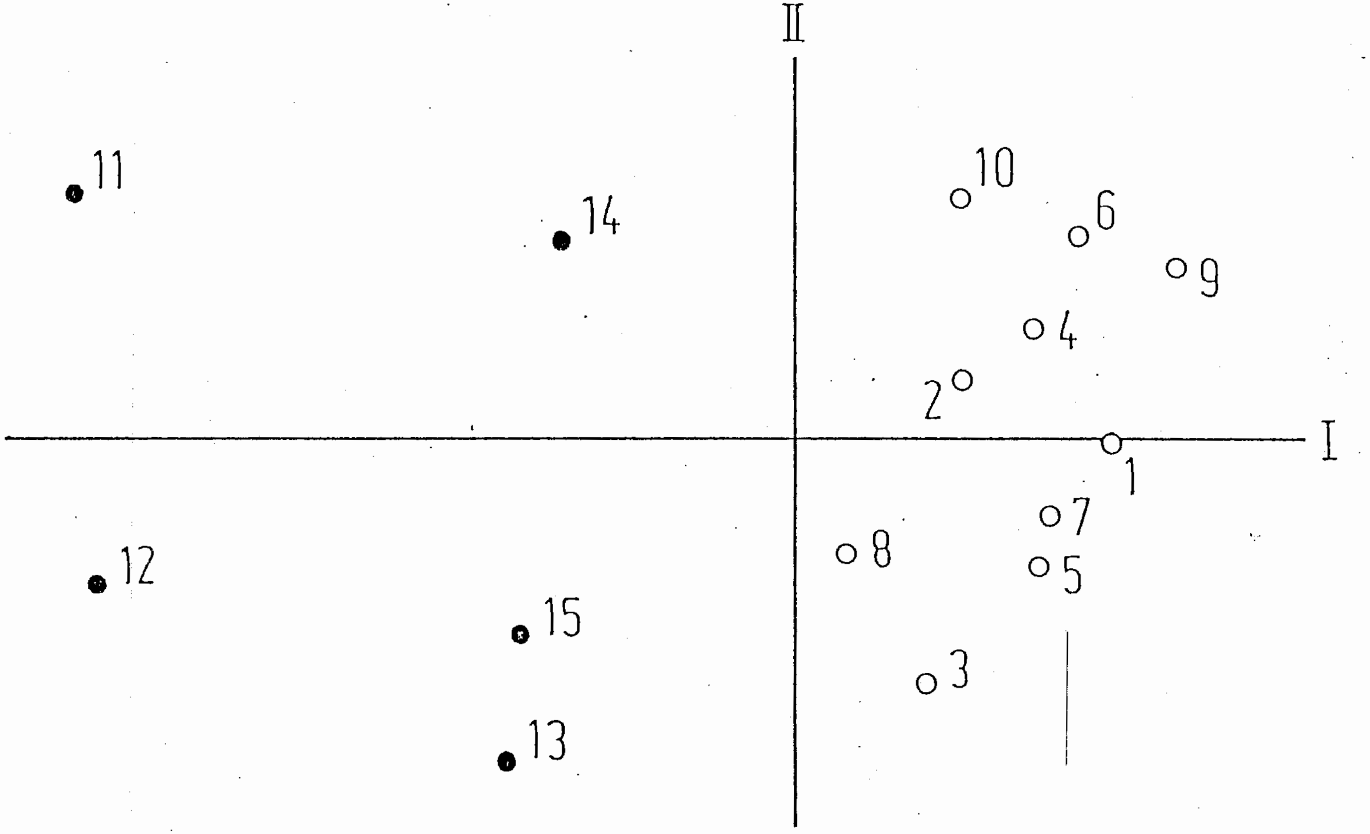


Fig. 2 Distribution of psychological distance of similarity

Table 1 The result of correlation analysis between psychological and physical distances

SOURCE CHARACTERISTICS	
(1) Pitch Contour	0.40
(2) Mean Pitch Frequency	0.24
RESONANCE CHARACTERISTICS	
(3) Formant Trajectory	0.52
(4) Mean Spectrum Envelope	0.32

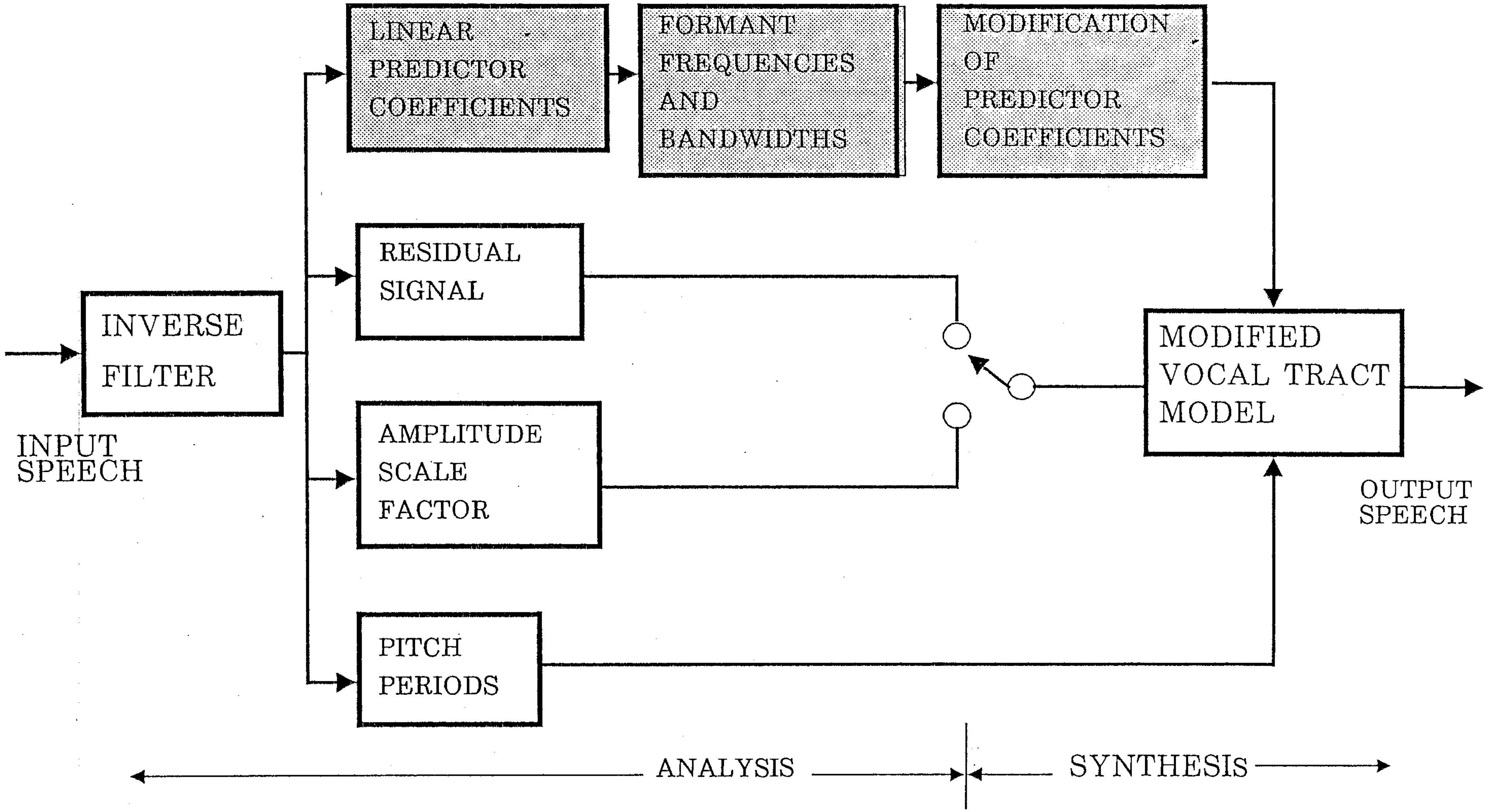

LINEAR PREDICTOR COEFFICIENTS
FORMANT FREQUENCIES AND BANDWIDTHS
MODIFICATION OF PREDICTOR COEFFICIENTS
RESIDUAL SIGNAL
INVERSE FILTER
INPUT SPEECH
AMPLITUDE SCALE FACTOR
PITCH PERIODS
MODIFIED VOCAL TRACT MODEL
OUTPUT SPEECH
ANALYSIS
SYNTHESIS

Fig. 3 Block diagram of analysis synthesis system

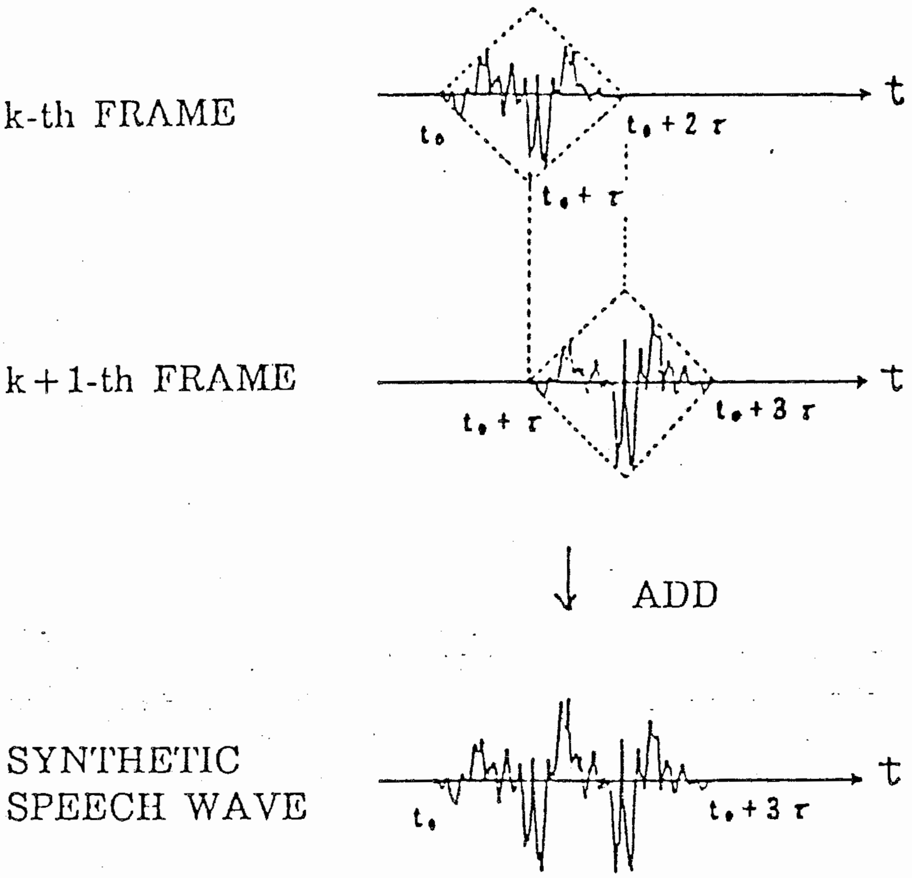
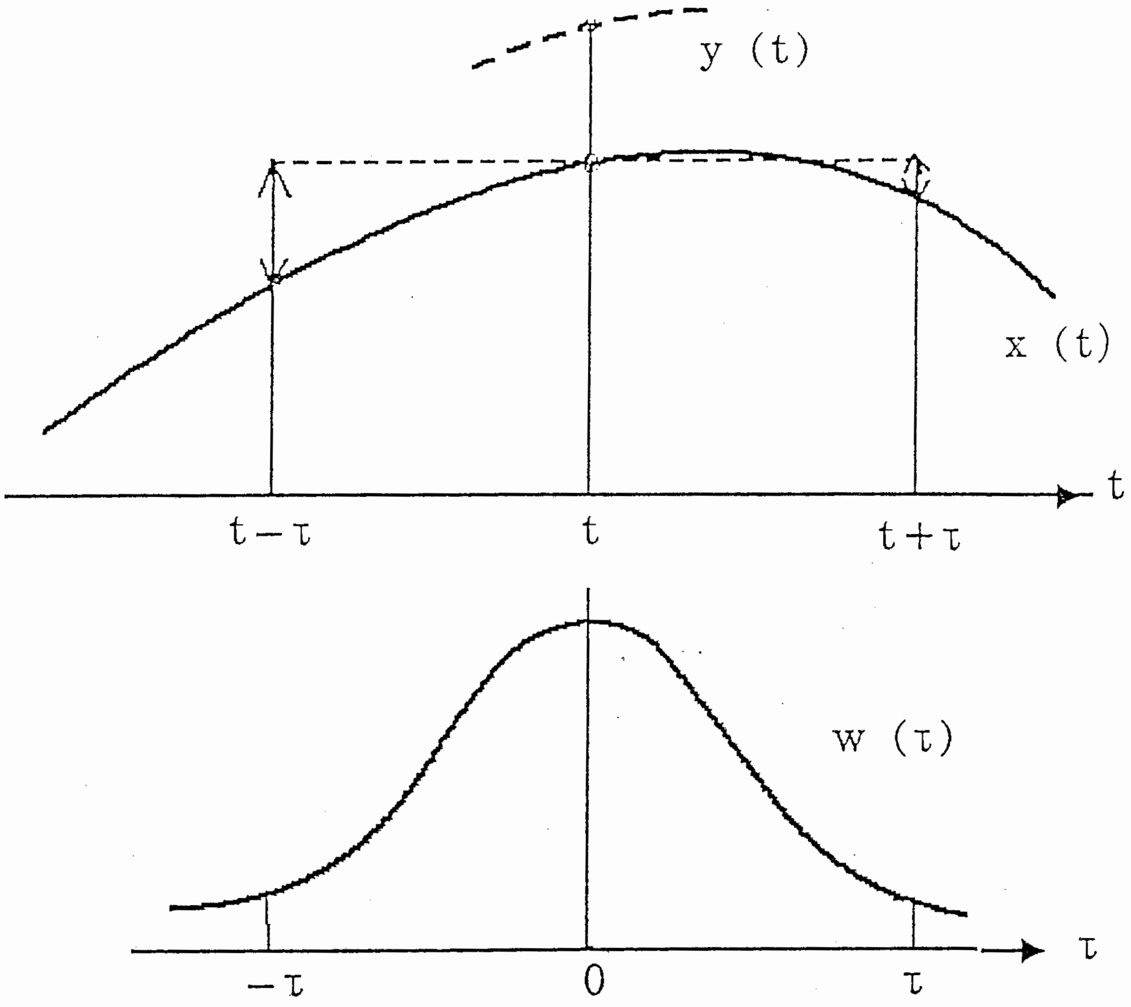


Fig. 4 Method of obtaining high quality synthetic speech



$$y(t) = x(t) + \int w(\tau)\ [x(t) - x(t+\tau)]\ d\tau$$

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

$$\sigma = 52\,\mathrm{ms}$$

Fig. 5 Schematic illustration of how to change the formant dynamics

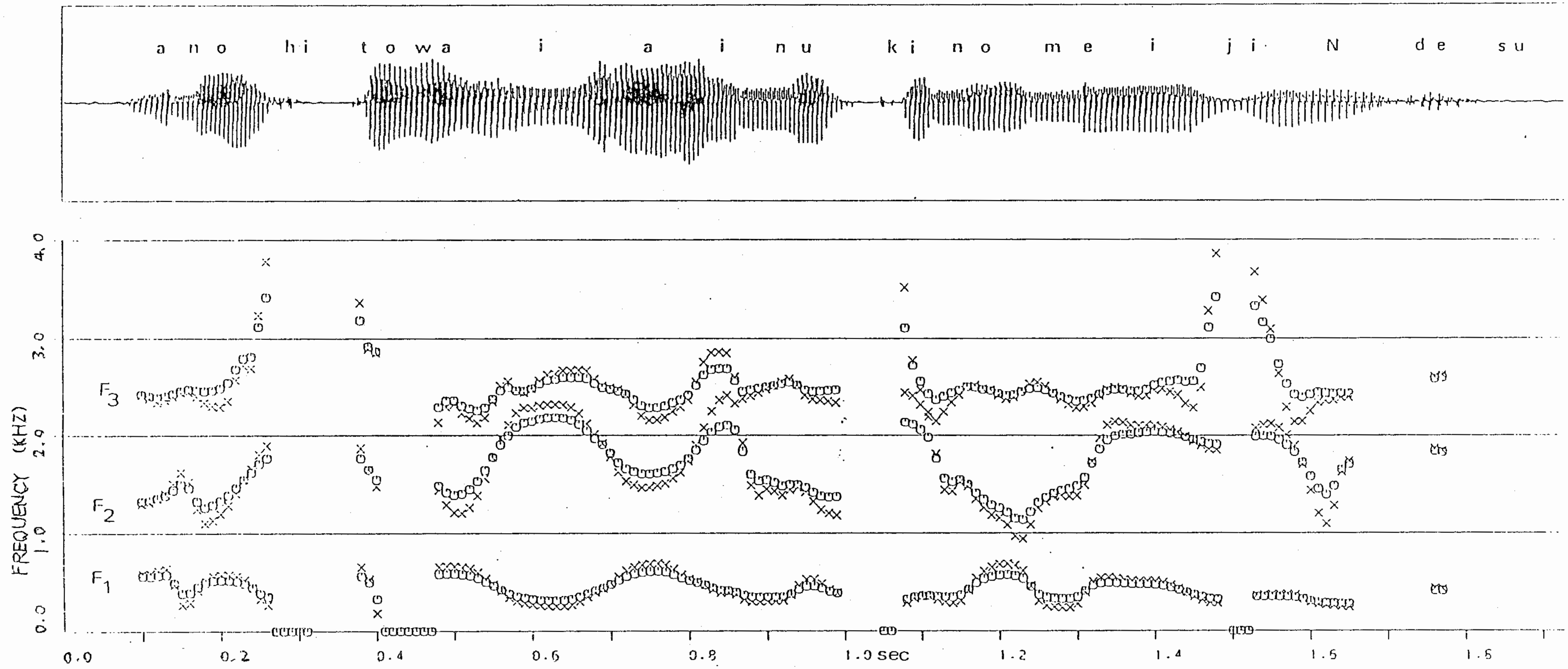


Fig. 6 An example of changing formant movement for a short sentence

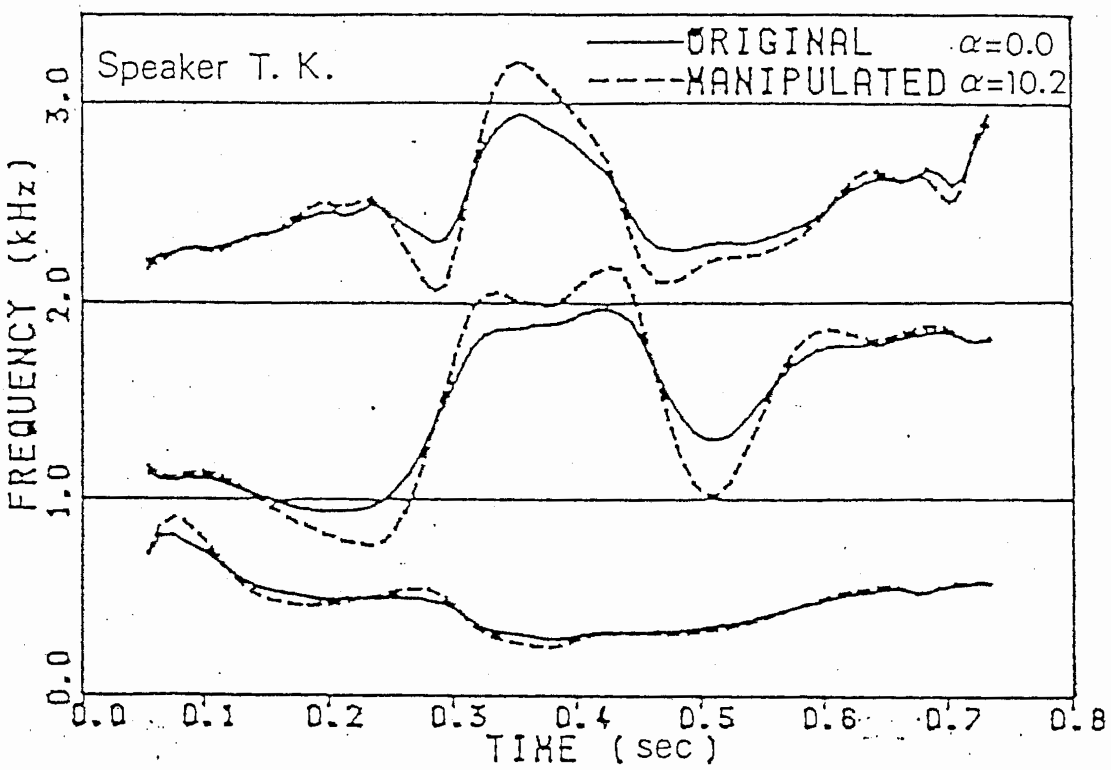


Fig. 7 Change of formant movement for the word /a o i u e/ used for the perceptual experiment

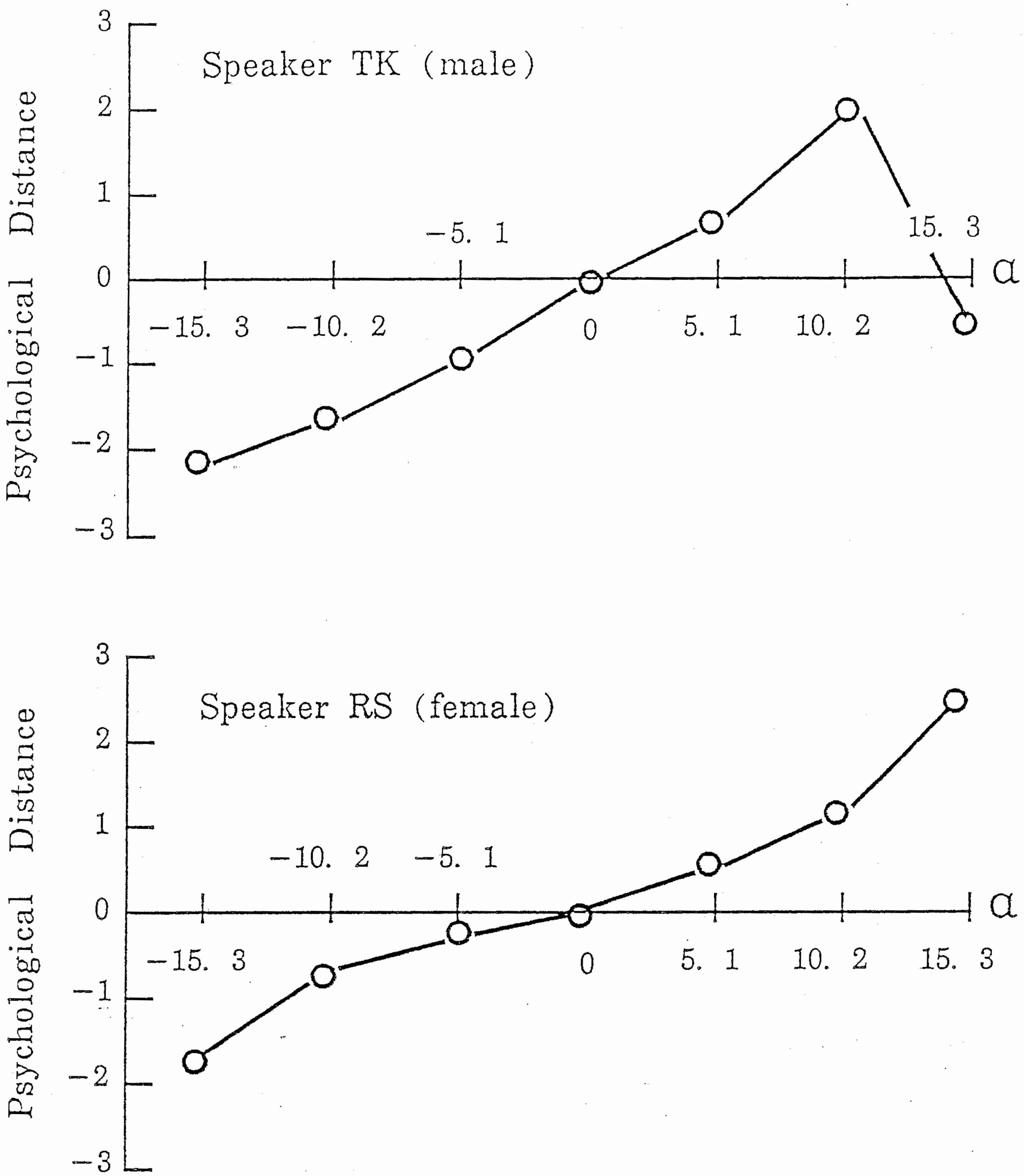
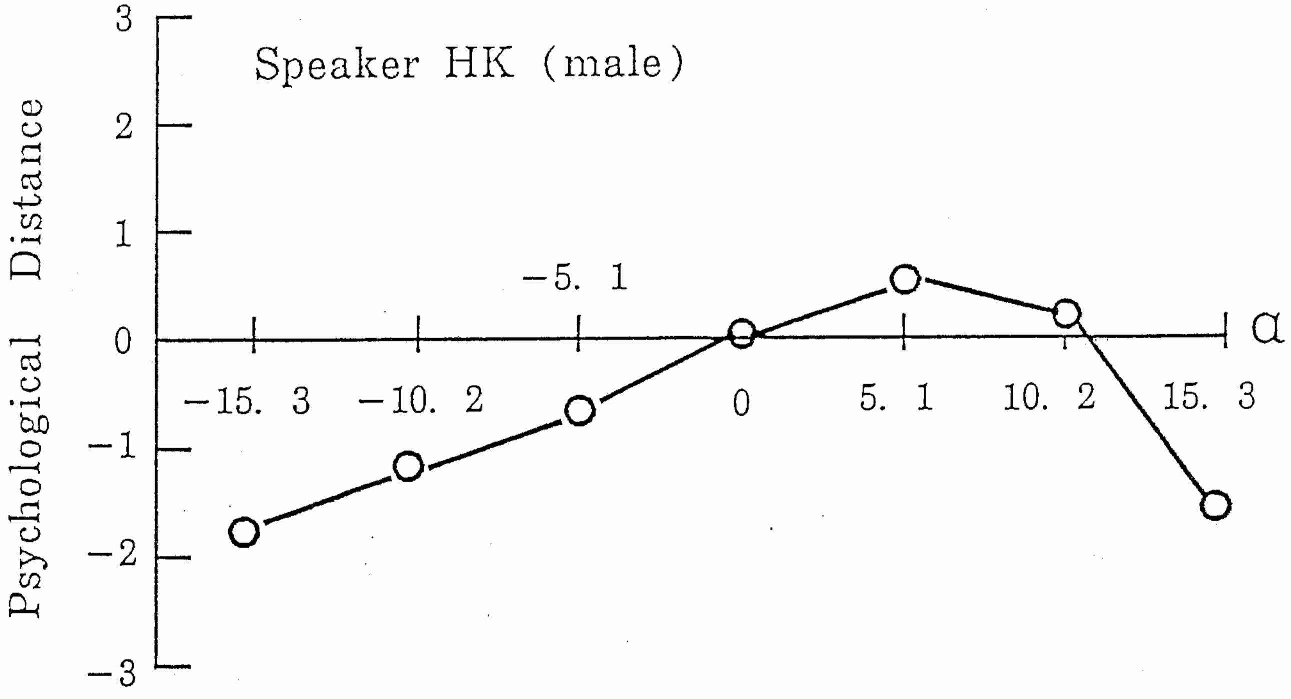


Fig. 8 (a) Experimental results of hearing test for two speakers



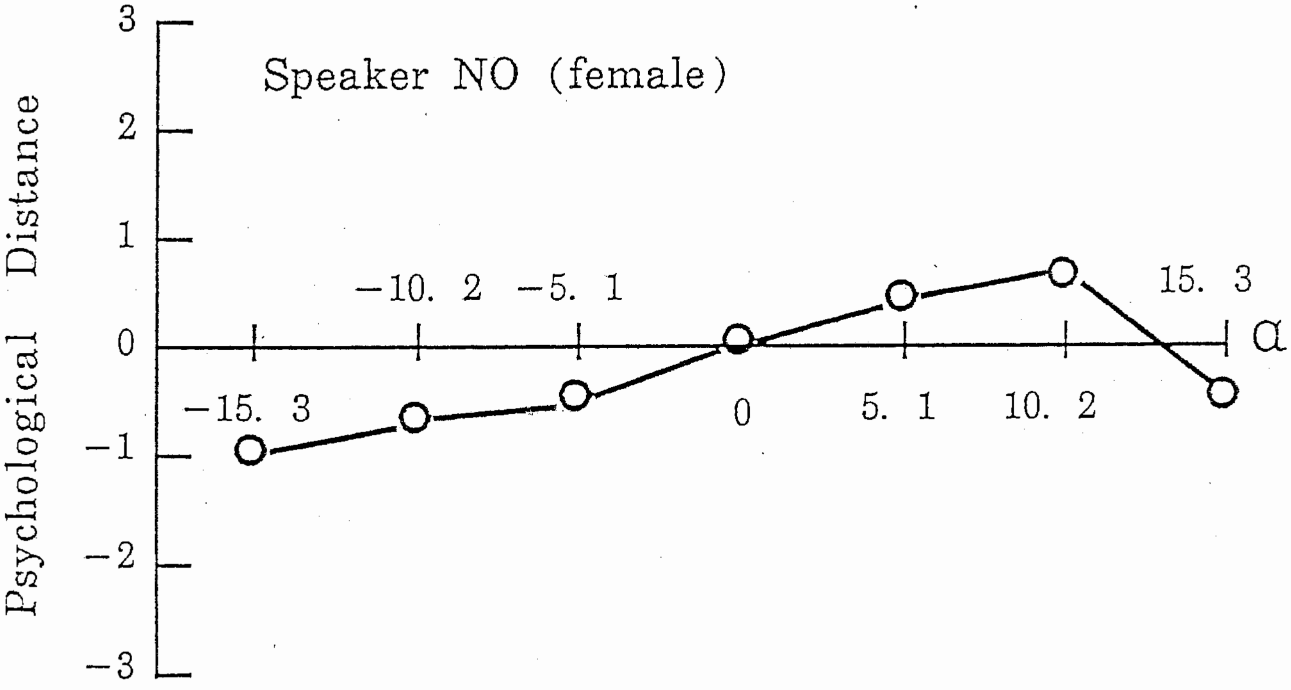


Fig. 8 (b) Experimental results of hearing test for two speakers