Speech Communication 10 (1991) 491–495
North-Holland

Acoustic parameters of voice individuality and voice-quality control by analysis–synthesis method

Hisao Kuwabara
The Nishi-Tokyo University, Uenohara, Kitatsuru-gun, Yamanashi 409-01, Japan

Tohru Takagi
NHK Science and Technical Research Laboratories, Setagaya, Tokyo 157, Japan

Received 15 June 1991

Abstract. Experiments on voice individuality have been performed using an analysis–synthesis system capable of modifying pitch, formant frequencies and formant bandwidths. The results show that the perception of voice-individuality is significantly affected by formant shifts, especially of the lower three, and it is completely lost for a uniform shift of five percent. Pitch frequency and bandwidth manipulation, on the other hand, is less important to the individuality perception.

Zusammenfassung. Es wurde experimentiert mit der Individualität der Stimme mit Hilfe eines Analyse- und Synthesesystems welches fähig war Stimmhöhe, Formantenfrequenzen und Bandbreiten zu verändern. Die Resultate zeigen, daß die Wahrnehmung der Individualität der Stimme stark beeinflußt wird von Verschiebungen der Formantfrequenzen, im besonderen der ersten drei. Die Wahrnehmung der Individualität geht verloren nach einer gleichmäßigen Verschiebung von fünf Prozent. Veränderungen der Stimmhöhe und der Bandbreiten scheinen, dahingegen, wenig Einfluß auf die Wahrnehmung der Individualität der Stimme zu haben.

Résumé. Des expériences sur l'individualité de la voix ont été effectuées en utilisant un système d'analyse–synthèse qui était à même de modifier la hauteur de la voix, les fréquences formantiques et les largeurs de bande. Les résultats montrent que la perception de l'individualité vocale est affectée de façon significative par des translations des fréquences formantiques, particulièrement des trois premières, et qu'elle est complètement perdue après une translation uniforme de cinq pourcent. La hauteur de la voix, ainsi que les largeurs de bande, par contre, semblent être moins importantes pour la perception de l'individualité.

Keywords. Voice individuality, analysis/synthesis, spectral manipulation.

1. Introduction

One of the great advantages of the analysis–synthesis method is that it can separate the voice source and the characteristics of the vocal tract from a speech sound and reconstruct speech, very close to the original speech sound in quality, by handling them independently. In a recent study (Itoh and Saito, 1982), it was shown that the characteristics of the vocal tract are more responsible for the voice individuality of speech than the source excitation. The results of our experiments (Kuwabara and Ohgushi, 1984) support this finding, though indirectly.

To investigate the relationship between acoustic parameters and voice individuality, a pitch synchronous analysis–synthesis system was developed, capable of the independent manipulation of formant frequencies and bandwidths in voiced speech (Kuwabara, 1984). In this paper, we first describe this system briefly, and then give the perceptual results on voice individuality, using speech samples whose pitch, formant frequencies and bandwidths were modified (Kuwabara, 1983).

0167-6393/91/$03.50 © 1991 Elsevier Science Publishers B.V. All rights reserved

2. Analysis–synthesis system

Figure 1 shows the block diagram of the analysis–synthesis system we have developed. Four parameters are extracted first for each pitch period. Unlike synthetic speech, waveforms of natural speech differ from one pitch period to another, even within a stationary vowel. This certainly contributes to the so-called "naturalness" of the human voice, together with fluctuations of the pitch period. To keep pitch fluctuations intact, we have constructed a pitch synchronous analysis–synthesis system. Our way of extracting the pitch period has been designed to preserve the one-to-one correspondence between period and the shape of the vocal tract should at least be retained. A relatively simple method to do this is to make use of the residual signal, actually a normalized squared-error signal (Wong et al., 1979). Though the normalized squared-error itself generally has a clear periodic structure, it occasionally fails to exhibit any exact periodicity of pitch period, due mainly to poor prediction of the resonances of the vocal tract, especially for vowel transitions and low power levels. To avoid these errors, glottal pulses estimated from the residual signal are used to extract the pitch period.

3. Method of spectrum manipulation

Formant frequencies and bandwidths are estimated first by solving a polynomial equation. Spectral manipulation is performed by altering the formant frequencies and bandwidths. Since spectral information is contained in the predictor coefficients, the manipulation is done by altering these coefficients.

Let $\{a_i\}$, $i = 1, \ldots, p$, stand for the predictor coefficients in a pitch period. It is well known that the formant frequencies and bandwidths are calculated by solving a polynomial equation which has $\{a_i\}$ as its coefficients. Let $\{z_i\}$, $i = 1, \ldots, p$, stand for the roots of the polynomial, and z_j' be the j-th pole to which we want to change the original. Then the predictor coefficients are modified so that the new polynomial equation has z_j' as one of its roots. Figure 2 represents the process to reconstruct the speech by changing formant frequencies and/or bandwidths.

4. Method of pitch manipulation

Pitch frequency manipulation is quite simple. 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. Therefore, 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. Perceptual experiments on voice individuality

The vocal tract resonance characteristics of one's speech certainly convey one's voice individuality. These resonance characteristics, a spectral envelope as their acoustic appearance, were changed by manipulating formant frequencies and bandwidths using the method described above, and the voice individuality was examined to investigate how the individual parameters contributed to voice quality.

5.1. Experimental procedure

Speech material used here was a short nonsense word, "aoiue", comprising a concatenation of five Japanese vowels. Two male speakers pronounced the word at a normal speech rate. Spectral manipulation was done by shifting the formant frequency to a high/low region, and widening/narrowing the formant bandwidths. For the formant shift, the following four experimental conditions were set:

(1) Uniform shift of all formant frequencies.
(2) Shift of the lower three formants (F_1, F_2, F_3) versus higher formants (higher than F_4).
(3) Shift in combination of two formants, F_1 and F_2, F_2 and F_3, F_3 and F_4.

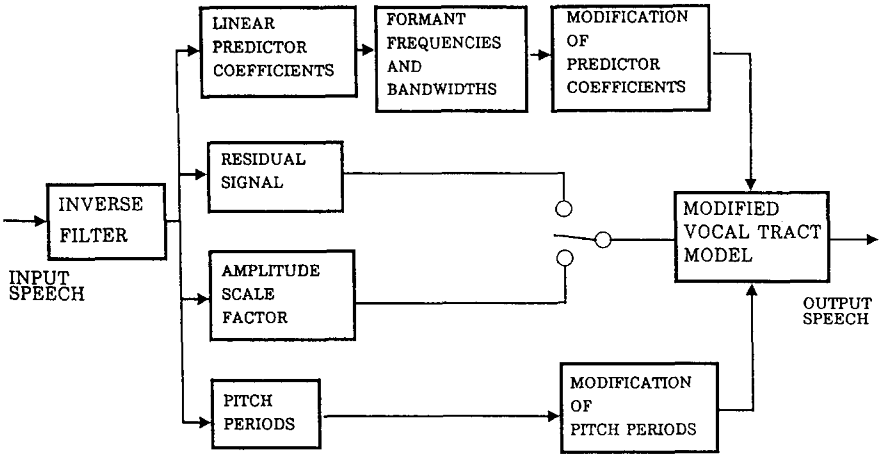
LINEAR PREDICTOR COEFFICIENTS
FORMANT FREQUENCIES AND BANDWIDTHS
MODIFICATION OF PREDICTOR COEFFICIENTS
RESIDUAL SIGNAL
INVERSE FILTER
INPUT SPEECH
MODIFIED VOCAL TRACT MODEL
AMPLITUDE SCALE FACTOR
OUTPUT SPEECH
PITCH PERIODS
MODIFICATION OF PITCH PERIODS

Fig. 1. Block diagram of pitch synchronous analysis–synthesis system.

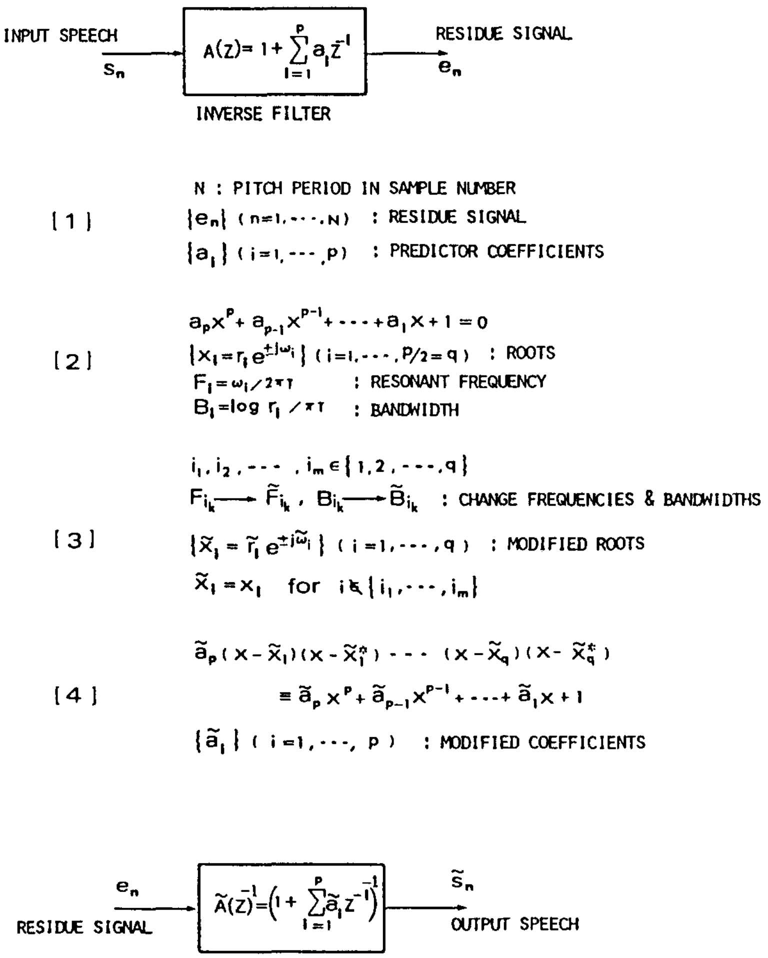
INPUT SPEECH
s_n
$A(z) = 1 + \sum_{i=1}^{p} a_i z^{-i}$
RESIDUE SIGNAL
e_n
INVERSE FILTER
[1]
N : PITCH PERIOD IN SAMPLE NUMBER
$\{e_n\}$ $(n=1,\cdots,N)$: RESIDUE SIGNAL
$\{a_i\}$ $(i=1,\cdots,p)$: PREDICTOR COEFFICIENTS
[2]
$a_p x^p + a_{p-1} x^{p-1} + \cdots + a_1 x + 1 = 0$
$\{x_i = r_i e^{\pm j\omega_i}\}$ $(i=1,\cdots,P/2=q)$: ROOTS
$F_i = \omega_i / 2\pi T$: RESONANT FREQUENCY
$B_i = \log r_i / \pi T$: BANDWIDTH
[3]
$i_1, i_2, \cdots, i_m \in \{1, 2, \cdots, q\}$
$F_{i_k} \longrightarrow \tilde{F}_{i_k}$, $B_{i_k} \longrightarrow \tilde{B}_{i_k}$: CHANGE FREQUENCIES & BANDWIDTHS
$\{\tilde{x}_i = \tilde{r}_i e^{\pm j\tilde{\omega}_i}\}$ $(i=1,\cdots,q)$: MODIFIED ROOTS
$\tilde{x}_i = x_i$ for $i \notin \{i_1,\cdots,i_m\}$
[4]
$\tilde{a}_p (x-\tilde{x}_1)(x-\tilde{x}_1^*) \cdots (x-\tilde{x}_q)(x-\tilde{x}_q^*)$
$\equiv \tilde{a}_p x^p + \tilde{a}_{p-1} x^{p-1} + \cdots + \tilde{a}_1 x + 1$
$\{\tilde{a}_i\}$ $(i=1,\cdots,p)$: MODIFIED COEFFICIENTS
e_n
RESIDUE SIGNAL
$\tilde{A}(z)^{-1} = \left(1 + \sum_{i=1}^{p} \tilde{a}_i z^{-i}\right)^{-1}$
$\tilde{s}_n$
OUTPUT SPEECH

Fig. 2. Process of changing formant frequencies and bandwidths.

(4) Individual formant shift from F_1 to F_3.

Parallel with these, a bandwidth narrowing and/or widening, which we tentatively call "manipulation" for convenience, was chosen as follows:

(5) Uniform manipulation of all formant bandwidths.

(6) Manipulation of the lower three (B_1, B_2, B_3) versus higher formant bandwidth (higher than the fourth).

(7) Manipulation in combination of two bandwidths, B_1 and B_2, B_2 and B_3, B_3 and B_4.

(8) Individual bandwidth manipulation of B_1, B_2 and B_3.

For each of the formant shifts and bandwidth manipulations, test speech samples modified according to the four experimental conditions were presented to listeners through a loudspeaker in a sound-proof chamber. The samples modified from the speech of one speaker were assembled in random order to form a set of stimuli for the experiment, and those from the other speaker were grouped separately. Three listeners, who knew the voices of the two speakers quite well, participated in the experiment. For each set of stimuli, listeners were asked to identify the speaker. The experiment was repeated ten times.

5.2. Result for formant shift

Because of the page limitation, we only give the result for condition (1). Figure 3 shows the result for one speaker, It is obvious that the voice individuality is entirely lost for a formant shift as small as 8 percent towards both the high and the low frequency regions. The boundary dividing the perception of individuality with a 50-percent probability is approximately 5 percent shift on both sides. The range inside which the individuality remains intact is as narrow as 2 percent on both sides.

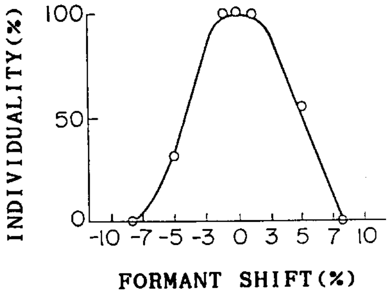

Fig. 3. Result of perceptual experiment on voice individuality for uniform formant shift.

5.3. Result for bandwidth manipulation

Figure 4 shows the result for the uniform change of all formant bandwidths (experiment condition (5)). The perception of voice individuality is more sensitive to widening the bandwidths than to narrowing them. The individuality is relatively well preserved over a wide range of bandwidth manipulation. However, it is completely lost for bandwidths narrower than one-fifth or wider than three times the original speech, but it is completely retained for bandwidths between half and twice as wide as the original.

5.4. Result for pitch frequency shift

Figure 5 depicts the result for the pitch frequency shift. Voice individuality retains over a

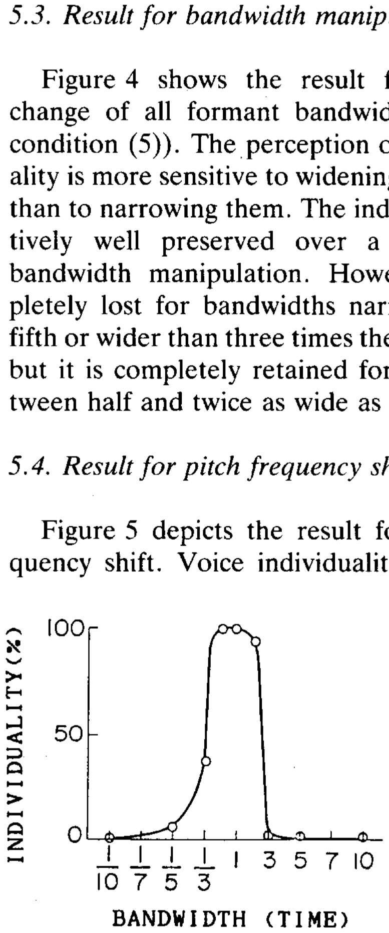

Fig. 4. Result of perceptual experiment on voice individuality for uniform bandwidth manipulation.

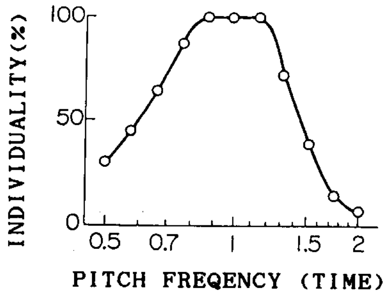

Fig. 5. Result of perceptual experiment on voice individuality for pitch frequency shift.

Table 1
The voice-individuality preserve range with more than 50% correct answer for formant shift and change of bandwidths

	Formant shift (%)	Bandwidth (time)
F_1	-12 to 12	1/30 to 4
F_2	-11 to 10	1/5 to 12
F_3	-7 to 9	1/10 to 30
F_1, F_2	-8 to 7	1/7 to 4
F_2, F_3	-7 to 7	1/4 to 6
F_3, F_4	-6 to 8	1/5 to 6
F_1, F_2, F_3	-6 to 6	1/5 to 4
F_4	-8 to 9	1/3 to 3
ALL	-4 to 5	1/3 to 3

wide range of pitch change, especially towards the low frequency region. Pitch frequency seems to be less sensitive to the voice individuality than any other acoustic parameters.

Table 1 shows the ranges where the voice individuality is preserved with more than 50% of the time for the change of formant frequency and bandwidths. The values are the means of the results for the two speakers. "All" indicates the uniform change for all formants, i.e., for the experiment condition (1) for the formant frequency shift and (5) for the change of bandwidths. It is seen from the table that the individuality is more sensitive to the change in F_3 than those of F_1 and F_2. F_3-and-below have stronger effects than F_4-and-above.

6. Conclusions

Experiments on voice individuality were performed to investigate the contribution of individual acoustic parameters to individuality by manipulating the spectral envelope and pitch frequency of natural speech. It has been found that the voice individuality is more sensitive to formant shifts than to bandwidth manipulation or pitch shift, and it is completely lost for a shift of approximately five percent of the original formants. For bandwidth manipulation, on the other hand, individuality is also lost for bandwidths three times wider than or one-third as narrow as those of the original speech. It has also been found that, for formant shifts, individuality is more sensitive to the lower three formants than to the higher ones, while for bandwidth manipulation it is sensitive to higher formant bandwidths.

References

K. Itoh and S. Saito (1982), "Effects of acoustical feature parameter of speech on perceptual identification of speaker", *Trans. IECE Japan*, Vol. J65-A, pp. 101–108.

H. Kuwabara (1983), "Acoustic modification of the vocal tract characteristics based on analysis/synthesis method and experiment on voice quality", *Trans. Committee on Speech Res. Acoust. Soc. Japan*, S82–73.

H. Kuwabara and K. Ohgushi (1984), "Independent manipulation of the formant frequencies and bandwidths and experiments on voice quality", *Trans. Committee on Speech Res. Acoust. Soc. Japan*, S84–40.

H. Kuwabara (1984), "A pitch-synchronous analysis/synthesis system to independently modify formant frequencies and bandwidths for voiced speech", *Speech Communication*, Vol. 3, No. 3, pp. 211–220.

D.Y. Wong, J.D. Markel and A.H. Gray Jr. (1979), "Least squares glottal inverse filtering from the acoustic speech waveform", *IEEE Trans.*, Vol. ASSP-27, pp. 350–355.