

Perception of Voice-Individuality for Distortions of Resonance/Source Characteristics and Waveforms

Hisao Kuwabara

Teikyo Univ. of Science & Technology
Uenohara, Kitatsuru-gun, Yamanashi 409-0193, Japan
kuwabara@ntu.ac.jp

Abstract

A perceptual study has been performed to investigate relationship between acoustic parameters and the voice individuality making use of the pitch synchronous analysis-synthesis system. Voice-individuality is involved in many acoustic parameters and the aim of this experiment is to examine how individual parameters affect the voice-individuality by separately giving them some distortions. Formant-frequency shift and bandwidth manipulations are given for spectral distortion, F_0 -shift for source manipulation. As the waveform distortion, zero-crossing and center-clipping techniques are used. It has been found that formant-shift is very sensitive to voice-individuality change and F_0 -shift and bandwidth manipulations are rather tolerant to the voice-individuality. The results of waveform manipulation reveal that the voice-individuality is kept more than the phonetic information for zero-crossing distortion and the results for center-clipping distortion are reverse.

1. Introduction

A lot of studies have so far been made about speech individuality [1-3]. However, many problems still remain unsolved and need to be investigated. Acoustic characteristics of individual speech, if their real time extraction is possible, are of very im-

portant for speech technology. Their immediate application to speech recognition and synthesis can be imagined. Recent development of speech analysis and synthesis techniques enables to study acoustic parameters more accurately than before. One of the great advantages of the analysis-synthesis method is that it can separate the voice source and the resonance characteristics of speech sounds and reconstruct speech very close to the original sound in quality by handling them independently.

A perceptual study has been performed to investigate relationship between voice-individuality and those acoustic parameters as formant frequency/bandwidth, fundamental frequency and others by making use of the pitch synchronous analysis-synthesis system [4]. Three manipulations on the original speech sounds have been taken into account; 1) Spectral manipulation, 2) Fundamental frequency manipulation, and 3) waveform manipulation. Experimental results will be shown later together with the results of a similar experiment for waveform distortions. Before proceeding to the perceptual experiment, let's take a brief look at the analysis-synthesis system used in this experiment.

2. Analysis-synthesis system

Fig. 1 illustrates a block diagram of analysis-synthesis system. First, low-pass filtered speech was digitized in 16 kHz at a rate

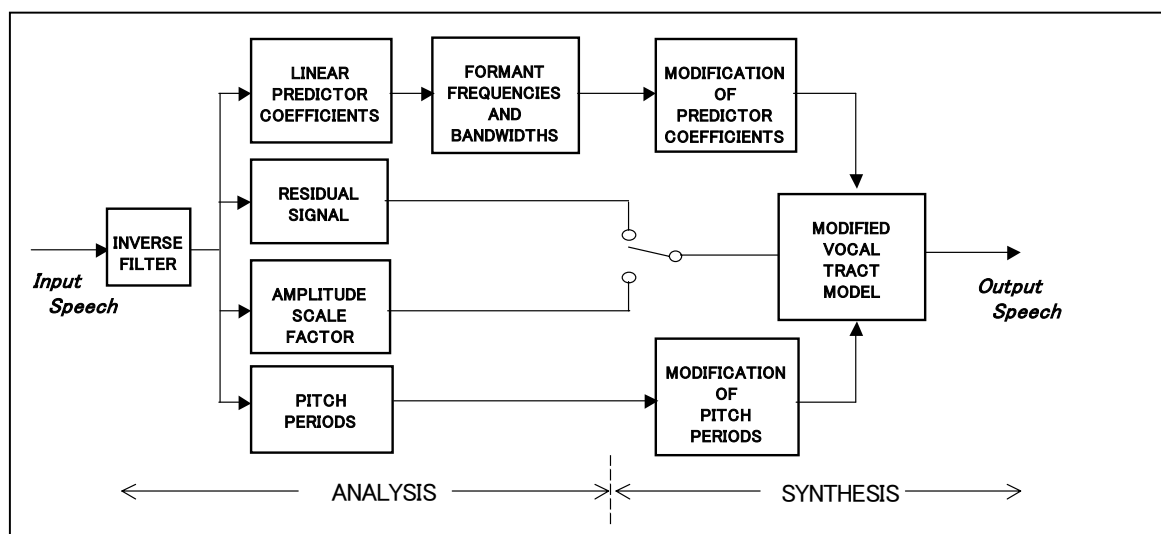


Fig. 1 Block diagram of analysis-synthesis system used for generating source/resonance manipulated speech.

of 15kHz. A short time LPC analysis based on the covariance method was carried out to obtain LPC coefficients and the residual signals. Pitch periods were obtained through peak-to-peak picking of the glottal waveforms estimated using the method proposed by Wang et al [5].

2.1. Spectral manipulation

The method of formant modification has already been reported in the article [4]. There are three steps in modifying formant frequencies and bandwidths as shown in Fig. 2. The outline of the method is as follows. For each pitch period, formant frequencies and bandwidths are calculated first by solving the polynomial equation. Then some modification is made to the original formant frequencies or bandwidths and accordingly to the predictor coefficients. A synthesis filter is formed using the modified coefficients. The residual signal, no change in length of course, has also been used as the input to the modified synthesis filter.

2.2. Fundamental frequency manipulation

$$\begin{aligned}
 & a_p z^p + a_{p-1} z^{p-1} + \dots + a_1 z + 1 = 0 \\
 [1] \quad & \{z_i = r_i e^{\pm j\omega_i}\}, \quad (i = 1, 2, \dots, p/2) : \text{roots} \\
 & F_i = \omega_i / 2\pi T : \text{Formants}, \quad B_i = \log r_i / \pi T : \text{Bandwidths} \\
 [2] \quad & \text{for } k \in \{1, 2, \dots, p/2\}, F_k \rightarrow \tilde{F}_k, B_k \rightarrow \tilde{B}_k \\
 & \tilde{z}_k = \tilde{r}_k e^{\pm j\tilde{\omega}_k} : \text{modified roots} \\
 [3] \quad & \tilde{a}_p (z - \tilde{z}_1)(z - \tilde{z}_2) \dots (z - \tilde{z}_p) \equiv \tilde{a}_p z^p + \tilde{a}_{p-1} z^{p-1} + \dots + \tilde{a}_1 z + 1 \\
 & \text{where } \{\tilde{a}_i\}, (i = 1, 2, \dots, p) : \text{modified coefficients}
 \end{aligned}$$

Fig. 2 Algorithm for changing formant frequencies and bandwidths

Fundamental frequency manipulation is quite simple. For each pitch period, the residual signal obtained at the analysis stage was used as the input to the synthesis filter which is exactly the reciprocal of the analysis filter or the vocal tract inverse filter. Pitch frequency change can be given simply by controlling the length of the residual signal. To raise the pitch frequency, some data at the last part of the residue are eliminated and to lower frequency, zero signals are added to the end of the residue. Finally, output signal from each synthesis filter are simply added to make pitch manipulated speech sound. In the resulting speech, acoustic parameters other than pitch frequency are of course kept intact.

2.3. Waveform manipulation

As the waveform manipulation, we adopted two distortions, 1) zero-crossing wave, and 2) center-clipping distortion. It is well known that zero-crossing waves carry enough phonetic information to be understood when the waveforms are meaningful words and sentences. However, when it comes to voice individuality no one knows whether zero-crossing waves carry enough information to identify the speaker. On the other hand, center-clipping is a kind of reverse operation. It eliminates waveform data very close to the zero-level, that is, it cuts off speech

speech data symmetrically to the zero-level. This operation is a sort of elimination of the zero-crossing information while other waveform information is kept intact.

3. Perceptual experiments

It is quite obvious that the acoustic information for voice individuality is involved not in a single acoustic parameter but distributed over several parameters. In this experiment, an investigation has been made on how individual distortions affect to the voice-individuality to find which acoustic parameters are sensitive to the individuality. Perceptual experiments on voice-individuality for speech materials that are reproduced through the above mentioned manipulation. Fine structures of pitch patterns are kept unchanged when their average values change. Comparisons with the results for formant manipulated speech reported before [5] have been made.

3.1. Experimental procedures

3.1.1. Spectral manipulation

Japanese five vowels have been used as the speech material except for the waveform-manipulation speech materials. Two adult male speakers uttered five Japanese vowels /a, i, u, e, o/ successively with a short silent interval between every two vowels. A set of these five vowels as a whole, without separating each other, has been used as the test material. Spectral manipulation has been performed by uniformly shifting the formant frequencies to a high/low frequency regions or uniformly widening/narrowing the formant bandwidths. Formant frequency shift has been given up to 10 percent both towards

high and low frequency regions. The bandwidth manipulation, on the other hand, has been altered up to ten times as wide as and 1/10 times as narrow as the original formant bandwidths. The test speech stimuli were presented to listeners through a loud speaker in a sound-proof chamber. Three listeners who knew the two speakers voices quite well participated in the experiment. For each stimulus, the listeners were asked to identify the speaker. Trials were repeated ten times.

3.1.2. Fundamental frequency manipulation

The same series of five vowels, but for eight speakers including the above mentioned two, has been used as the test material. The F_0 manipulation has been made by increasing/decreasing the average frequency while keeping the duration, the pitch-contour, the intensity pattern as close to the original ones as possible. Pitch change has been made from -40%, that is 40% lower than the original value, to +40% with a step of 5%, which makes a total of 17 stimuli including the original voice (0%) per speaker.

Twelve listeners participated in this experiment. A set of 50 stimuli which include one speaker's 17 samples and the rest from different speakers, all random in order, has been arranged. Just before the 50-stimuli set, listeners hear the speaker's original voice and they are asked whether each stimulus in the set is the same speaker's voice they heard immediately before or not. Ex-

periment has been repeated three times per listener.

3.1.3. Waveform manipulation

For both zero-crossing and center-clipping operations, speech materials are 1) isolately spoken 100 CV-syllables, 2) 10 meaningful words and 10 nonsense words, 3) 10 short sentences. Two speakers uttered these speech samples and ten listeners participated in the hearing test. In addition to the speaker identification test, phonetic identification, word/sentence intelligibility tests, which are not included in other manipulation speech, have also been conducted. In this experiment, a different strategy was taken from the rest of the experiments. Since the listeners didn't know the two speakers voices at all, upon hearing each stimulus they were forced to choose one speaker from three original voice samples that followed.

3.2. Experimental results

The listeners' responses are separately pooled and analyzed. For some listeners, judgment of speaker is somewhat difficult task and it depends on the listener's familiarity to the speaker's voice to be identified. The followings are the experimental results.

3.2.1. Spectral manipulation

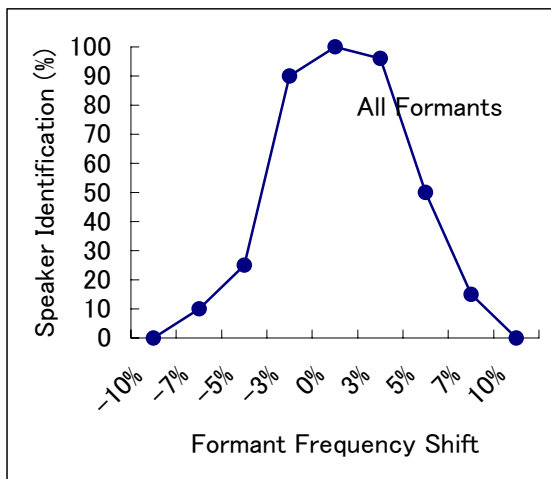


Fig. 3 Result of speaker identification for formant shift speech

Fig. 3 shows the result of speaker identification experiment for formant-shift speech. The first through sixth formant frequencies are uniformly shifted toward high and low frequency regions by up to 10%. The center point on the abscissa (0% point) represents the original speech and the plus direction stands for the shift toward high frequency and the minus direction lower them. It is clear, from the figure, that the voice individuality is almost completely kept within 5% shifts toward both high and low frequency regions but is completely lost when the shift exceeds 10%. The response curve shows nearly symmetric towards high and low frequency regions. Generally, it seems that the voice individuality is rather sensitive to the formant frequency.

Fig. 4, on the other hand, represents the result for the bandwidth widening/narrowing. Though it appears there is an abrupt change in response curve, the voice-individuality is well

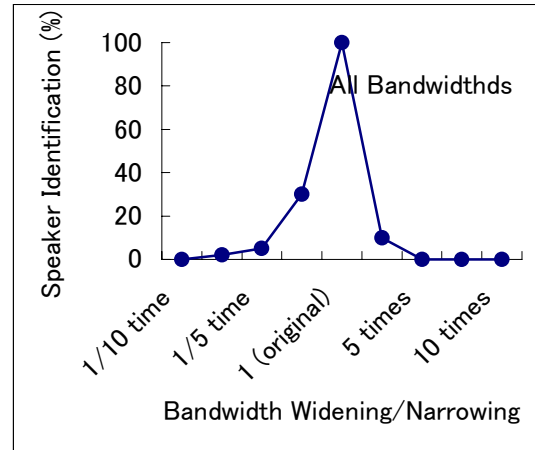


Fig. 4 Result of speaker identification for bandwidth manipulated speech.

retained for the bandwidths up to 3 times as wide as and 1/3 times as narrow as the original. It may be concluded that the voice-individuality is not sensitive to the formant bandwidth than to the frequency itself.

3.2.2. Fundamental frequency manipulation

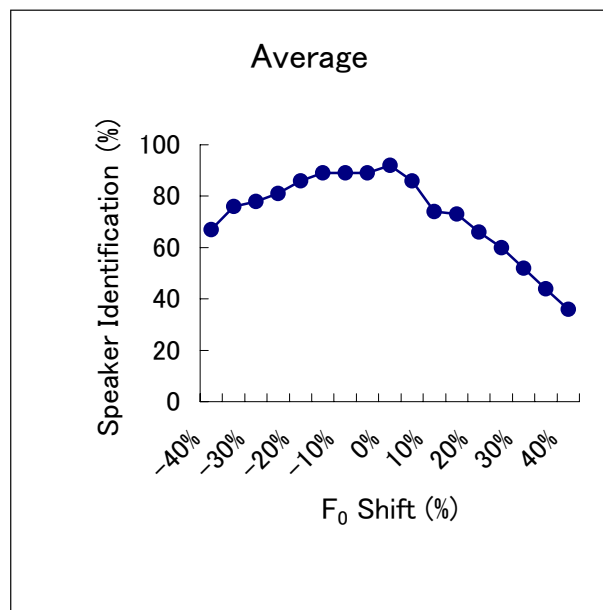


Fig. 5 Result of speaker identification for fundamental frequency shift speech averaged over nine speakers.

Fig. 5 shows the results for the pitch frequency shift averaged over the eight speakers. The abscissa represents pitch frequency shift and the ordinate depicts the percentage of voice individuality. The point 0% represents the original speech again and plus region shows the F₀ shift towards high-pitch-voice and minus region towards low-pitch-voice. It has been found that the individuality does not seem to be affected very much by pitch change, especially towards low frequency. It retains almost 70% of individuality for -40% shift while it goes down to approximately 30% for 40% pitch shift. Generally, pitch frequency shift has little effect to the voice individuality especially for lowering it.

3.2.3. Waveform manipulation

Unlike the former two manipulations, the waveform distortions, one is zero-crossing and the other is center-clipping, are quite simple. It is often said that phonetic information, especially sentence intelligibility, is largely retained by the zero-crossing operation. In addition to the speaker identification, conventional phonetic identification tests are also conducted for this waveform distortion speech samples. **Figs. 6** and **7** represent the results for the zero-crossing and center-clipping speech samples, respectively.

Figures 6 and 7 look almost the same but are slightly different in details. In both figures, speaker identification scores are higher for all speech materials as compared with the phoneme identification, which implies that the voice-individuality is rather robust against this kind of waveform distortions. The

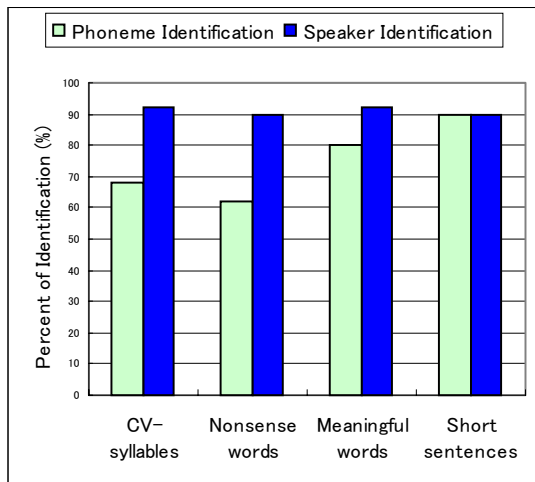


Fig. 6 Result of phoneme/speaker identification for zero-crossing speech.

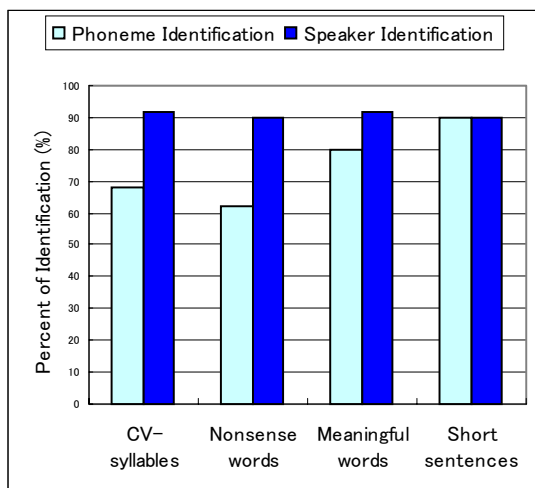


Fig. 7 Result of phoneme/speaker identification for center-clipping speech.

elimination levels of the center-clipping operation were three kinds; -10, -15, 120dBs being the average power level of a speech sample 0dB. The results in Figure 7 show the average scores over the responses for the three clipping-level speech

samples. It is clear from the figures that, for short sentences, both distortions give almost no damage to the voice-individuality as well as to the phonetic information. Linguistic information would certainly contribute to the result. Other speech materials, such as CV-syllables and nonsense words, exhibit a small difference on the scores between phoneme and speaker identifications. Unlike to the formant frequency/bandwidth manipulation, waveform distortions have little effect on the voice-individuality even for CV-syllables and nonsense words which are considered to lack much of linguistic information.

4. Conclusions

A perceptual experiment has been performed on voice individuality for acoustic-parameter-manipulated speech of human voice making use of an analysis-synthesis system that is capable of independent manipulation of fundamental and formant frequencies and bandwidths as well as those for waveform distortions. This experiment has been performed to examine which parameter carry more information and/or more robust, in terms of voice individuality, against acoustic distortions. Experimental results reveal that the perception of voice individuality is far more sensitive to the formant frequencies than to the F_0 frequency. Only a few percent shift of all formant frequencies, either towards high or low frequency regions, from the original values has been found to cause a 100% loss of the voice individuality while F_0 frequency shift does not. The fundamental frequency has been found to be far more robust than the formants. A few percent shift of F_0 frequency has no damage on individuality. It has been found that the individuality is not entirely lost until F_0 shift reaches as large as 50%, either towards high or low frequency regions, from the original value. Formant bandwidths, on the other hand, have been found to be less sensitive than the frequency itself. Little damage on the voice individuality has been found to occur when the bandwidth widening/narrowing is less than 3 times of the original widths.

Compared to the formant and fundamental frequency manipulations, perception of voice individuality has been found to be far less sensitive to the waveform distortions. Conventional zero-crossing and center-clipping operations have been done for four types of speech materials; 1) CV-syllables, 2) nonsense words, 3) meaningful words, and 4) short sentences. None of the four exhibits different perceptual result. They all show consistently high percentage of speaker identification, though they have a little difference in scores of phoneme identification. This implies that the voice individuality is very robust against waveform distortions while the phonetic information is not.

5. References

- [1] Matsumoto, H., Hiki, S., Sone, T., and Nimura, T., "Multi-dimensional representation of personal quality of vowels and its acoustical correlates," *IEEE Trans., AU-21*, 428-436, 1973
- [2] Furi, S., Itakura, F., and Saito, S., "Talker recognition by the long-time averaged speech spectrum," *Trans. IECE Japan*, 55-A, 549-556, 1972
- [3] Itoh, K., and Saito, S., "Effects of acoustical feature parameters of speech on perceptual identification of speaker," *Trans. IECE Japan*, J65-A, 101-108, 1982
- [4] Kuwabara, H., "A pitch-synchronous analysis/synthesis system to independently modify formant frequencies and bandwidths for voiced speech," *Speech Communication*, 3, 211-220, 1984
- [5] Wong, D. Y., Markel, J. D., and Gray, Jr. A. H., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans., ASSP-27*, 350-355, 1979