音楽情報科学 19－13
(1997. 2. 21)

声道特性の静的特徴と動的特徴が個人性の知覚に与える寄与の検討

朱 偉中 粕谷 英樹
宇都宮大 工学部

宇都宮大学工学部電気電子工学科

〒321 宇都宮市石井町2753
E-mail:zhu@klab.ishii.utsunomiya-u.ac.jp

あらまし ARX音声生成モデルに基づいて、音声「青い上 /aoiueoie/」から、音源パラメータと声道パラメータを抽出する。音声のホルマント軌跡に対して、Discrete Cosine Transform(DCT) を用いて、話者の声道の静的特徴と動的特徴を分離する。複数の発声者間における、声道の静的パラメータと動的パラメータを複合処理した合成音に対して、個人性の知覚に与える寄与を検討する。声道の動的特徴より、静的特徴が個人性の知覚に与える寄与は大きいことを示す。

キーワード 個人性、ホルマント軌跡、ARXモデル

Study of Perceptual Contributions of Static and Dynamic Features of Vocal Tract Characteristics to Speaker Individuality

Weizhong ZHU, Hideki KASUYA

Faculty of Engineering, Utsunomiya University 2753 Ishii-machi, Utsunomiya 321

E-mail:zhu@klab.ishii.utsunomiya-u.ac.jp

Abstract The ARX speech production model was used to extract voice source and vocal tract parameters from a Japanese speech, /aoiueoie/ ("say blue top" in English). The Discrete Cosine Transform (DCT) method was applied to separate formant trajectories of the speech signal into static and dynamic features. The perceptual contributions of static and dynamic vocal tract features to speaker individuality was quantitatively studied by systematically replacing the corresponding acoustic parameters extracted from three Japanese male utterances. Results of the experiments show that the static (average) characteristic of the vocal tract is a primary cue to speaker individuality.

key words Speaker individuality, Formant trajectories, ARX model

1 Introduction

As the output of the built-in limits of the speech production system, speech waves convey linguistic, paralinguistic information and extra-linguistic [1]. Extra-linguistic information carries the speaker's idiosyncratic features that show the differences between speakers in the physiological structure of the vocal apparatus and talking behavior. Where speech communication is concerned, speaker individuality appears in the form of anatomical, articulatory, acoustic and perceptual differences. The study of speaker individuality can be classified into a horizontal study and a vertical study. Horizontal study focuses on speakers differences in each of anatomical, articulatory, acoustic or perceptual categories, while vertical study focuses on the relationship of these speaker differences between categories. In an actual application, such as an automatic speaker recognition system, usually only one category is examined, in this case the acoustic. Much effort has been made to find a robust speaker's features, to decrease identification error rate. In order to understand the true nature of speaker individuality, it is necessary to accumulate knowledge of speaker individuality by vertical study. A better understanding of speaker individuality not only can improve performance of speaker recognition and verification, but also can create a better speaker adaptation method in speech recognition; can help synthesize more natural speech; and can even help to find more efficient methods for voice conversion.

From perceptual experiments, it was reported that the difference in fundamental frequency is one of the important features distinguishing male and female speakers [2]. Itoh et al. [3] showed, using a PARCOR analysis-synthesis system, that the spectral envelope is more responsible than pitch, or dynamic characteristics in five male subjects. Kuwabara et al. [4] found that the voice individuality is more sensitive to formant shift than to bandwidth manipulation or pitch shift. Furui et al. [5] showed that the individuality of the spectral envelope mainly exists in frequency band between mean 2.5 and 3.5 kHz. Recently Kitamura et al. [6] reported that speaker individuality in spectral envelopes is mainly above 22 ERB (2212 Hz). They suggested that people identify speakers of vowels mainly by using the high frequency band in the spectral envelopes. In automatic speaker recognition process, Furui [7] suggested combining dynamic features with statistic features. Soong and Rosenberg [8] showed that instantaneous and transitional spectral features are fairly uncorrelated, and can used jointly to improve speaker recognition performance.

Those studies, however, dealt with spectral envelopes that included both voice source and vocal tract characteristics. In this paper, an ARX speech production model was used to separate voice source characteristics from vocal tract characteristics [9]. The purpose of this paper is to quantitatively study the perceptual contributions of voice source and vocal tract characteristic to speaker individuality, specifically that of the static and dynamic features of vocal tract characteristics. The Discrete Cosine Transform (DCT) method was applied to separate formant trajectories of the speech signal into static and dynamic features. The perceptual contributions of voice source features; static and dynamic vocal tract features; and lower and higher formant trajectories, to speaker individuality was quantitatively studied by systematically replacing the corresponding acoustic parameters extracted from three Japanese male utterances /aoiueoie/ ("Say blue top") in English. The speech stimulus was synthesized by an ARX-speech-production-model based speech synthesis system that was developed on the MS-Windows platform [10].

2 Analysis-Synthesis Method

2.1 ARX model

The speech production process can be modeled as a time-variant IIR system with an equation error as follows,

$$s(n) + \sum_{i=1}^{p} a_i(n)s(n-i)$$

$$= \sum_{j=1}^{q} b_j(n)u(n-j) + u(n) + e(n) \qquad (1)$$

where $s(n)$ and $u(n)$ denote a speech signal and a glottal waveform at time n, respectively. In the above equation, $a(n)$ and $b(n)$ are time-varying coefficients. p and q are model orders, and $e(n)$ is an equation error. When $e(n)$ is assumed to be white, the equation represents an ARX model. By performing the Z-transform on the equation (assuming time invariance), we get the following equation,

$$S(z) = \frac{B(z)}{A(z)}U(z) + \frac{1}{A(z)}E(z) \qquad (2)$$

where $S(z)$,$U(z)$ and $E(z)$ are the Z-transforms of the speech signal $s(n)$, voice source $u(n)$, and equation error $e(n)$, respectively. The ARX model consists of an IIR filter and an AR filter. The vocal tract transfer function of voiced sounds is represented by $B(z)/A(z)$, whereas the production process of unvoiced sounds is approximated by an AR model with a transfer function $1/A(z)$ driven by white noise.

2.2 Voicing source model

The RK model is used to represent a differentiated glottal wave form because of its capability of adjusting independently, both the waveform and spectral slope, as well as it's relative easy of implementation. This model uses a generator of a rudimentary waveform defined as

$$g(n) = \begin{cases} 2an - 3bn^2, & 0 \leq n \leq T \cdot OQ, \\ 0, & T \cdot OQ < n < T, \end{cases} \tag{3}$$

$$a = \frac{27 \cdot AV}{4 \cdot (OQ^2 \cdot T)}, \quad b = \frac{27 \cdot AV}{4 \cdot (OQ^3 \cdot T^2)},$$

where T is the fundamental period, AV the amplitude parameter and OQ the open quotient of the glottal open phase divided by the duration of a complete glottal cycle. The value of $g(n)$ is 0 in the close period. $g(n)$ is filtered by a low-pass filter to adjust the tilt of its spectral envelope using a spectral tilt parameter TL.

2.3 Analysis algorithm

In the analysis scheme, the kalman filter algorithm is used to estimate the formant parameters from the coefficients of the ARX model, and the simulated annealing method is employed as a nonlinear optimization approach to estimate the voice source parameters (refer [9] for details).

2.4 Cascade formant synthesizer

A cascade formant synthesizer that was constructed in the speech synthesis system is used to synthesize the voiced and unvoiced speech signal. The RK model is used to synthesize the voiced sound, whereas the M-series white noise is used to synthesize the unvoiced sound. The synthesizer is composed of second-order resonators and antiresonators in cascade form. The spectrum for each resonator is expressed as

$$H(z) = \frac{a}{1 - bz^{-1} - cz^{-2}}, \tag{4}$$

$$c = -exp(-2\pi B/f_s),$$
$$b = 2exp(-\pi B/f_s) \cdot cos(2\pi F/f_s),$$
$$a = 1 - b - c,$$

and that for each antiresonator as

$$H(z) = a' + b' z^{-1} + c' z^{-2}, \tag{5}$$

$$a' = 1/a, \quad b' = -b/a, \quad c' = -c/a,$$

where F, B and f_s are the formant frequency, the bandwidth and the sampling frequency, respectively.

3 Perceptual Similarity Experiments

3.1 Speech materials and acoustic analysis

Three adult males (MHK, MMM, MSH) participated in the experiments. We choose this three subjects because we have already studied their speaker individualities in anatomical structures of sustained vowels measured by magnetic resonance images[11]. A simple Japanese sentence /aoiueoie/ ("say top blue in English") was recorded on DAT tape in a sound proof room. First MHK's speech was recorded 5 times. We selected one of the speech that most seems to be MHK's natural voice. Then MMM and MSH listened by headphone to the selected MHK's speech with silent intervals inserted and during each silent interval would repeated the same sentence, but as much possible using their natural voice. The silent interval was set to be slightly longer than the speech interval so that MMM and MSH would produce speech intervals of similar length to MHK. Finally we selected one speech interval from each speaker ensuring the selected intervals were as equal in length as possible. ARX analysis was performed at pitch-synchronous mode on each speech signal at a sampling frequency of 14700 Hz. Acoustic measurements were made to determine the formant frequencies and bandwidths which represent the vocal tract characteristics. Measurement were also made to determine the fundamental frequency contours ($F0$), the voicing amplitude parameter (AV), the noise amplitude parameter (NA), the open quotient (OQ) and the spectral tilting parameter (TL) which represent voice source characteristics. Final acoustic parameters for each speaker were obtained by resampling each parameter in a 5 ms period. MHK's parameter trajectories of AV, F0 and formants are shown in Fig.1.

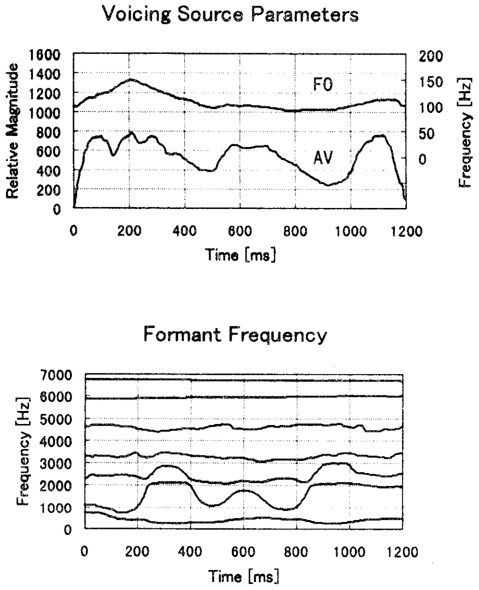


Fig.1 MHK's parameter tracjectories of F0, AV and formants.

3.2 Method of perceptual judgment

An X-A-B judgment method was used, where X was the test stimulus which synthesized by using a combined parameter set arranged from two different speakers, and A and B were the synthetic speech signal generated from original speaker parameters. Listeners were asked to determine whether X was A's voice or B's voice. The stimulus intervals are shown in Fig.2. Ten male listeners participated in the following three experiments. All were native speakers of Japanese and had no known hearing impairments. Every pair of stimuli was presented 10 times (5 times in the order of X-A-B, 5 times in the order of X-B-A). All the pairs of stimuli were presented randomly in a quiet room through a speaker (DIATONE professional, AS-1051) at a comfortable loudness level. Three experiments were executed on three separated days.

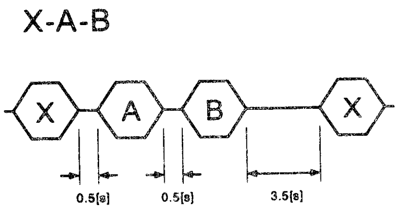


Fig.2 Intervals between stimuli.

3.3 Experiment 1

3.3.1 Method of synthesizing speech stimuli

The first experiment was designed to study the perceptual contributions of voice source and vocal tract characteristics by using different combinations of voice source parameters and formant trajectories parameters among three speakers.

3.3.2 Results

The perceptual speaker identification results are depicted in Fig.3. A total of 100 judgments (10 persons × 10 times) were done for each different A×B speaker combination. Referring to Fig.3, when MMM's vocal tract characteristics were combined with MSH's voice source characteristics, only 6% (or 6 times) of the cases is the resultant voice identified as MSH's. In other words, 94% of the time that voice was identified as MMM's voice. On the other hand, if we combined MSH's vocal tract characteristics with MMM's voice source characteristics, the resultant voice is identified as MMM's voice only 1% of the time. In the diagram a result of 0% is found in 3 of the cases. The experiment clearly shows that the vocal tract characteristics contribute much more to the perception of the speaker individuality that the voice source.

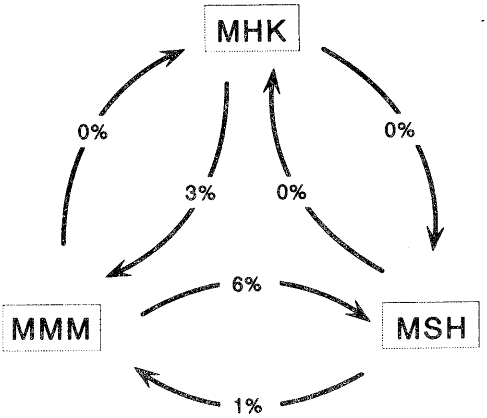


Fig.3 Indentification rates resulting from exchanging voice source parameters between three speakers.

3.4 Experiment 2

3.4.1 Method of synthesizing speech stimuli

The results of preliminary experiment show voice source characteristics having very little effect on the perception of speaker individuality. Experiments 2 and 3 were designed to test the perceptual contribution of vocal tract characteristics. In experiment 2 and 3, we use a fixed set MHK's voice source parameters to generate all stimuli in order to avoiding any voice source effect.

To further study vocal tract characteristics, Discrete Cosine Transform (DCT) was applied to separate formant trajectories of the speech signal into static and dynamic features. the formant trajectories can be represented by their DCT coefficients,

$$F_i(n) = \frac{1}{\sqrt{2}} C_i(0) + \sum_{k=1}^{N-1} C_i(k) cos\left[\frac{(2n+1)}{2N} k\pi\right], \quad (6)$$

$$0 \leq n \leq N-1,$$

$$C_i(0) = \frac{\sqrt{2}}{N} \sum_{n=0}^{N-1} F_i(n), \quad (7)$$

$$C_i(k) = \frac{2}{N} \sum_{n=0}^{N-1} F_i(n) cos\left[\frac{(2n+1)}{2N} k\pi\right], \quad (8)$$

$$1 \leq k \leq N-1,$$

where $F_i(n)$ is the i-th formant frequency at n-th frame and N is the number of analysis frames. $\frac{1}{\sqrt{2}} C_i(0)$ is the mean value of $F_i(n)$ showing the static feature of the i-th formant trajectory. The other DCT coefficients $C_i(k), 1 \leq k \leq N-1$, keep the dynamic features of the i-th formant trajectory.

We define residual error $e_i^{(K)}$ of the i-th formant trajectory as

$$e_i^{(K)} = \left[\frac{1}{N} \sum_{n=0}^{N-1} \{F_i(n) - F_i^{(K)}(n)\}^2 \right]^{1/2}, \quad (9)$$

$$F_i^{(K)}(n) = \frac{1}{2} C_i(0) + \sum_{k=1}^{K} C_i(k) cos \left[\frac{(2n+1)}{2N} k\pi \right], \quad (10)$$

$$K = 1, \cdots, N-1,$$

$$F_i^{(0)} = \frac{1}{\sqrt{2}} C_i(0). \quad (11)$$

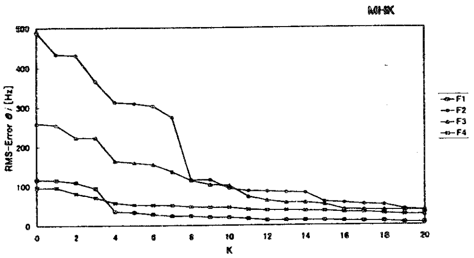


Fig.4 Residual Errors.(MHK's formant trajectories)

Figure 4 shows the pattern of decrease of the residual error of formant F1 to F4. As the number of coefficients to represent the formant trajectories is increased, the residual error is decreased. By using 21 coefficients, all the residual errors are lower than 50 Hz. It is also shown that compared to F1 and F4, F2 and F3 have more dynamic characteristics.

We synthesized the speech by using one person's static features and another person's dynamic features. The perceptual experiments were made to show the perceptual contributions of static and dynamic features of the vocal tract, to speaker individuality. An example of the formant trajectories for the test stimulus is shown in Fig.5. (a) and (c) are the original formant trajectories of MHK and MSH. The formant trajectories of test stimulus (b) contains MSH's static features and MHK's dynamic features.

3.4.2 Results

The perceptual speaker identification rates for experiment 2 are shown in Fig.6. If MSH's static characteristics are combined with MMM's dynamic characteristics, 95% of the time, the voice is identified as MSH's voice. In the reverse, the result is 100%. If we add the main part of dynamic features to the static features, the identification rate nearly reaches 100%. From the diagram, it is shown that static features carry over 89% of perceptual contribution.

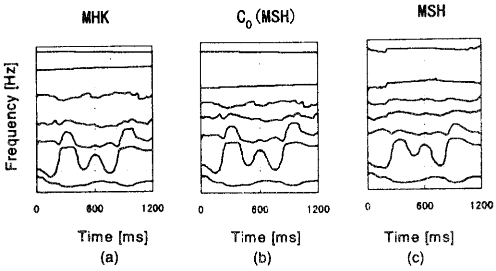


Fig.5 An example of formant trajectories while combine MHK's dynamic features with MSH's static features.

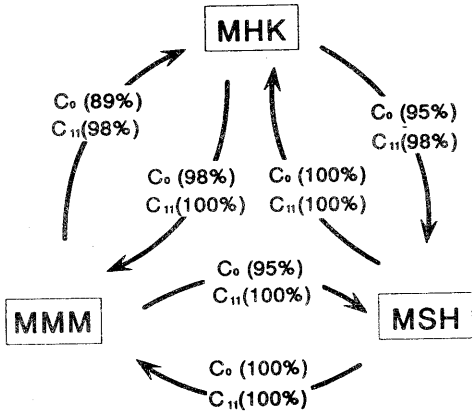


Fig.6 Indentification rates resulting from exchanging static and dynamic fatures between three speakers.

3.5 Experiment 3

3.5.1 Method of synthesizing speech stimuli

It is known that the acoustic phonetics of vowels depend on the lower formant frequency (F1, F2 and F3). In experiment 3, we exchange the lower formant trajectories (form F1 to F3) and higher formant trajectories (from F4 to F7) between three speakers.

3.5.2 Results

Figure 7 shows the results of perceptual speaker identification rates in six different combinations. In the case of MHK and MMM, it seems that high formant trajectories contribute a large part of individuality information (94%,86%), while in the case of MSH and MHK, the perceptual contributions of low formant trajectories (84%,77%) is bigger than that of high formant trajectories. In the case of MMM and MSH, both low and high formant trajectories are of nearly the same importance. The figures shown below the diagram are the distances of the average of F1-4 trajectories between the test stimulus and the original synthetic speech signal. These distances can explain above perceptual results(except the case of MSH's F1-3 and MMM's F4-7). The lower the distance, the higher the identification rate. In the case of MMM and MSH, it is rather difficult to do the identification test, since the distances are far from both MMM and MSH.

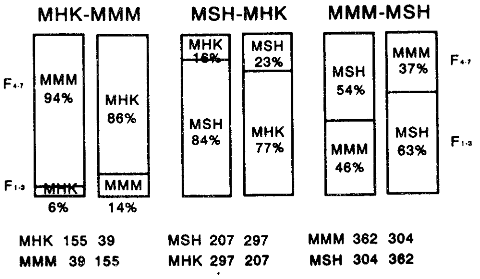


Fig.7 Indentification rates resulting from exchanging low and high formant trajectories between three speakers.

The result can also be explained from comparing the average spectrum of the test stimulus to that of two speakers. It can be seem from Fig. 8 that the general spectral envelope of the test stimulus is closer to MMM in (a), to MSH in (b), and to MSH in (c).

4 Discussion

Acoustic differences between speech signals which carry the same linguistic information come from inter-speaker acoustic variability and intra-speaker variability. Experiment 1 showed that vocal tract characteristics contribute more to the perception of speaker individuality than the voice source. From a source-filter theory point of view, both variabilities appear in voice source and vocal tract characteristics. Average pitch value seems to be one of the important features of voice source characteristics. In this study, the average pitch of three sentences uttered by three speakers are 108 Hz for MHK, 123 Hz for MMM, and 111 Hz for MSH. There is no significant difference among the three subjects. A normal speaker can change the average pitch to a little higher or lower value without any difficulty. In fact, people use different pitch values to express their intentions. The intra-speaker variability of voice source characteristics is relatively large. Consequently there is overlap between the features of voice source characteristics of different speakers (where male speakers are concerned). On the other hand, because of speaker's anatomical constraints and articulatory behavior, the acoustic variability of vocal tract characteristics is rather small. It seems from the results of the experiment, that the difference in variability between voice source characteristics and vocal tract characteristics, the later being small, results in the vocal tract characteristic contributing more to the perception of speaker individuality.

Experiment 2 and 3 were carried out to investigate the effect of the vocal tract characteristics in the absence of variation in voice source characteristics. In experiment 2, we use the DCT method to separate vocal tract characteristics into static features and dynamic features. The result clearly showed that the static characteristic of the vocal tract is a primary cue to the perception of speaker individuality. Generally, static features reflect the speaker's anatomical characteristics, while dynamic features reflect more the speaker's talking behavior. The results imply that mapping of the average characteristic of vocal tract would be an easy and practical technique for voice conversion. In the previous study, we had investigated the anatomical structures of the three subjects. The average length of their vocal tract are 17.9 cm, 15.9 cm and 16.5 cm for MHK, MMM, and MSH, respectively. Even though the average length (15.9 cm) of MMM's vocal tract is close to the vocal tract length of a female, the first three formant frequencies are shown similar to MHK's values[11]. The values of the first three formant frequencies were calculated from their measured area functions. We exchange the lower formant trajectories (form F1 to F3) and higher formant trajectories (from F4 to F7) in experiment 3. Coincidentally, we found that higher formant trajectories conveyed the speaker individuality between MHK and MMM. In the case of MSH and MHK, the perceptual contribution of low formant trajectories is bigger than that of high formant trajectories. The distances of the average of F1-4 trajectories between the test stimulus and the original synthetic speech can explain most of the results.

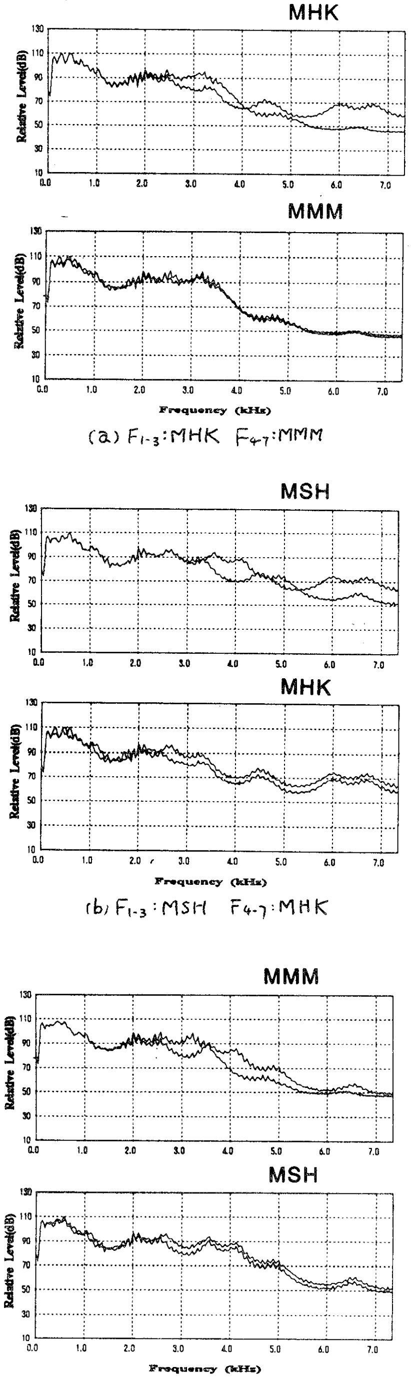
5 Conclusions

In this study, we investigated the perceptual contributions of voice source and vocal tract characteristics to speaker individuality. The results show that

(1) Vocal tract characteristics contribute more to the perception of speaker individuality than the voice source;

(2) The static (average) characteristic of the vocal tract is a primary cue to the perception of speaker individuality;

(3) Speaker individuality is conveyed primarily by the average F1-F4 value.



(a) $F_{1\text{-}3}$:MHK $F_{4\text{-}7}$:MMM

(b) $F_{1\text{-}3}$:MSH $F_{4\text{-}7}$:MHK

(C) $F_{1\text{-}3}$:MMM $F_{4\text{-}7}$:MSH

Fig.8 Comparison of average spectrum of test stimulus to that of two speaker. (a)MHK and MMM, (b)MSH and MHK, (c)MMM and MSH.

Acknowledgment

The authors would like to thank the persons who participated the experiments. The authors also thank Mr. Masahiro Matsuda for his help in the making of the figures in this paper. This work was partly supported by the Ashigin International Foundation of Ashikaga Bank, Tochigi, Japan.

References

[1] H. Kasuya and C.S. Yang, "Voice quality associated with voice source," J.Acoust. Soc. Jan. vol.(J)51, no.11, pp.869-875, 1995.

[2] J. Suzuki, "Correlation of speaker's physical features and speech," J. Acoust. Soc. Jpn. vol.(J)41, no.12, pp.895-900, 1985.

[3] K. Itoh and S. Saito, "Effects of acoustical feature parameters of speech on perceptual identification of speaker," Trans. IEICE, vol.J65-A, no.1, pp.101-108, 1982.

[4] H. Kuwabara and T. Takagi "Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method," Speech Communication, vol.10, nos5-6. pp.491-495, 1991.

[5] S. Furui and M. Akagi "Perception of voice individuality and physical correlates," Tech. Rep. Hear. Acoust. Soc. Jpn. H85-18, pp.1-8, 1985.

[6] T. kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes," J. Acoust. Soc. Jpn. vol.(E)16, no.5, pp.283-289, 1995.

[7] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," IEEE Trans. Acoust., Speech & Signal Process., vol.29, no.3, pp.342-350, 1981.

[8] F.K. Soong and A.E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," IEEE Trans. Acoust. Speech & Signal Process., vol.36, no.6, pp.631-639, 1988.

[9] W. Ding, H. Kasuya and S. Adachi, "Simultaneous estimation of vocal tract and voice source parameters based on an ARX model," IEICE Trans. Inf. & Syst., Vol.E78-D, No.6, pp.738-743, 1995.

[10] W. Zhu and H. Kasuya, "A new speech synthesis system based on the ARX speech production model," Proc. ICSLP96, Philadelphia, vol.3, pp.1413-1416, 1996.

[11] C.S. Yang and H. Kasuya, "Speaker individuality of vocal tract shapes of Japanese vowels measured by magnetic resonance images," Proc. ICSLP96, Philadelphia, vol.2, pp.949-952, 1996.