ROLES OF STATIC AND DYNAMIC FEATURES OF FORMANT TRAJECTORIES IN THE PERCEPTION OF TALK INDEDIVDUALITY

Weizhong Zhu and Hideki Kasuya
Kasuya Lab
Faculty of Engineering, Utsunomiya University
2753 Ishii-machi, Utsunomiya, 3221 Japan.
TEL & FAX: +81 28 689 6122, E-mail: zhu@klab.ishii.utsunomiya-u.ac.jp

ABSTRACT

Experiments were performed to investigate perceptual contributions of static and dynamic features of vocal tract characteristics to talker individuality. An ARX (Auto-regressive with exogenous input) speech production model was used to extract separately voice source and vocal tract parameters from a Japanese sentence, /aoiueoie/ ("Say blue top" in English). The Discrete Cosine Transform (DCT) was applied to resolve formant trajectories of the speech signal into static and dynamic components. The perceptual contributions were quantitatively studied by systematically replacing the corresponding formant components extracted from Japanese sentences uttered by three males. Results of the experiments show that the static (average) characteristic of the vocal tract is a primary cue to talker individuality.

1. INTRUDUCTION

Due to built-in properties of the speech production system, speech waves convey not only linguistic and paralinguistic information but also extra-linguistic information that carries talker's idiosyncratic features resulting from the differences between talkers in the physiological structure of the vocal apparatus and talking behaviour. A deeper understanding of talker individuality is important in various speech research areas, such as talker identification and verification, talker adaptation in speech recognition, synthesis of natural speech, and voice quality conversion.

It has been reported that spectral envelopes are more responsible for perceptual identification of talkers than the pitch or the LPC residual signals [1],[2]. It has also been reported that talker individuality in spectral envelopes mainly exists in the frequency band between 2.5 and 3.5 kHz [3] or above 22 ERB rate (2,212 Hz) [4]. Those studies, however, deal with spectral envelopes that include both voice source and vocal tract characteristics.

In this paper, an ARX speech analysis method is used to separate voice source characteristics from vocal tract characteristics [5]. The Rosenberg-Klatt (RK) model is used to simulate a glottal waveform of voiced speech. The Kalman filter algorithm is used to estimate the filter coefficients of the ARX model and then formant-antiformant parameters are obtained by solving for the roots of the coefficients. The simulated annealing method is employed as a non-linear optimization approach to estimate the voicing source parameters.

The acoustic analysis was performed on a Japanese sentence /aoiueoie/ ("Say blue top" in English) uttered by three males. In perceptual experiments, test stimuli were synthesized by replacing a part of the vocal tract parameters of one talker by that of another. Results of the experiments have shown that static (average) features of the vocal tract is a primary cue to talker individuality.

The paper is organized as follows. In section 2, we briefly describe the ARX speech production model and analysis algorithm used to extract acoustic parameters from the speech signal and the cascade formant synthesizer used to synthesize test stimuli. A method of perceptual similarity experiments is described in section 3. We show and discuss the results in section 4. Finally, we give conclusion.

2. ANALYSIS-SYNTHESIS METTHOD

2.1. ARX Speech Production Model

Speech production process is modelled as an IIR filter with an equation error $e(n)$,

$$s(n) + \sum_{i=1}^{p} a_i s(n-i) = \sum_{j=0}^{q} b_j u(n-j) + e(n), \quad (1)$$

where $s(n)$ and $u(n)$ denote a speech signal and a differentiated glottal waveform at time n, respectively. In the equation, a_i and b_j are filter coefficients, and p and q are model orders. When $e(n)$ is assumed to be white, the equation represents an ARX model. By performing the z-transform on the equation, one gets the following,

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

where $S(z)$, $U(z)$ and $E(z)$ are the z-transform of the speech signal $s(n)$, the voicing source $u(n)$, and the equation error $e(n)$, respectively. $B(z)/A(z)$ indicates the vocal tract transfer function for the voiced sound, whereas $1/A(z)$ represents the one for the unvoiced sound.

2.2. Voicing Source Model

The RK model is used to represent a differentiated glottal waveform because of its ability to adjust independently both the waveform and spectral slope as well as of its relative easiness of implementation. This model generates a rudimentary waveform $g(n)$ defined as

$$\begin{aligned} g(n) &= 2an - 3bn^2, \quad 0 \le n \le T \cdot OQ, \\ &= 0, \qquad\qquad T \cdot OQ < n < T, \\ a &= \frac{27 \cdot AV}{4 \cdot (OQ^2 \cdot T)}, b = \frac{27 \cdot AV}{4 \cdot (OQ^3 \cdot T^2)}, \end{aligned} \tag{3}$$

where T is a pitch period, AV an amplitude parameter and OQ an open quotient of the glottal open phase to the duration of a complete glottal cycle. $g(n)$ is zero in the closed phase. The differentiated glottal waveform $u(n)$ is generated by smoothing $g(n)$ with a low-pass filter where tilt of the spectral envelope is adjusted by a spectral tilt parameter TL.

2.3. Analysis Algorithm

In order to estimate the IIR filter coefficients, the filter is expanded into a time-variant system so that a Kalman filter algorithm can be used. The simulated annealing method based on the mean square equation error (MSEE) criterion, is employed for non-linear optimization to estimate the voice source parameters. The formant-antiformant parameters are obtained by solving for the roots of the coefficients [5].

2.4. Cascade Formant Synthesizer

A cascade formant synthesizer is used to synthesize the voiced and unvoiced speech [6]. The RK model is used to synthesize the voiced speech, whereas the M-series white noise is used to synthesize the unvoiced speech. The synthesizer is composed of the second-order resonator in cascade form. The system function of each resonator is expressed as

$$\begin{aligned} H(z) &= \frac{a}{1 - bz^{-1} - cz^{-2}}, \\ b &= 2\exp(-\pi B / f_s)\cos(2\pi F / f_s), \\ c &= -\exp(-2\pi B / f_s), \\ a &= 1 - b - c, \end{aligned} \tag{4}$$

and that of anti-resonator as

$$\begin{aligned} H(z) &= a' + b'z^{-1} + c'z^{-2,}, \\ a' &= 1/a, \quad b' = -b/a, \quad c' = -c/a, \end{aligned} \tag{5}$$

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

3. PERCEPTUAL SIMILARITY EXPERIMENTS

3.1. Speech Materials And Analysis

Three male adults (MHK, MMM, MSH) participated in the experiments. We chose these three subjects because we already studied their vocal tract shapes of five Japanese vowels measured by magnetic resonance images [7]. The simple Japanese sentence /aoiueoie/ ("Say blue top" in English) was recorded by a condenser microphone (SONY, C-38B) on a DAT tape in a sound proof room and was sampled at a sampling frequency of 14.7 kHz. The ARX speech analysis was performed to estimate voicing source parameters and formant trajectories. Some manual modifications were made on the sixth and seventh formant trajectories to obtain the smooth formant trajectories.

3.2. Method of Synthesizing Speech Stimuli

In the preliminary experiment, we found that voice source characteristics had very little effect on the perception of talker individuality. We focused on the role of the formant trajectories. We used a fixed set of MHK's voice source parameters to generate all the stimuli in order to avoid any voice source effect.

The 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,

$$\begin{aligned} F_i(n) &= \frac{1}{\sqrt{2}} C_i(0) + \sum_{k=1}^{N-1} C_i(k) \cos\left[\frac{(2n+1)}{2N} k\pi\right], \\ &\qquad\qquad 0 \le n \le N-1, \\ C_i(0) &= \frac{\sqrt{2}}{N} \sum_{n=0}^{N-1} F_i(n), \\ C_i(k) &= \frac{2}{N} \sum_{n=0}^{N-1} F_i(n) \left[\frac{(2n+1)}{2N} k\pi\right], \\ &\qquad\qquad 1 \le k \le N-1, \end{aligned} \tag{6}$$

where $F_i(n)$ is the i-th formant frequency at n-th frame and N is the number of analysis frames. $C_i(n)/\sqrt{2}$ is the mean value of $F_i(n)$, representing the static feature of the i-th formant trajectory. The other DCT coefficients $C_i(k)$, $1 \le k \le N-1$, convey the dynamic features of the i-th formant trajectory. We define residual error $e_i^{(K)}$ of the i-th formant trajectory as

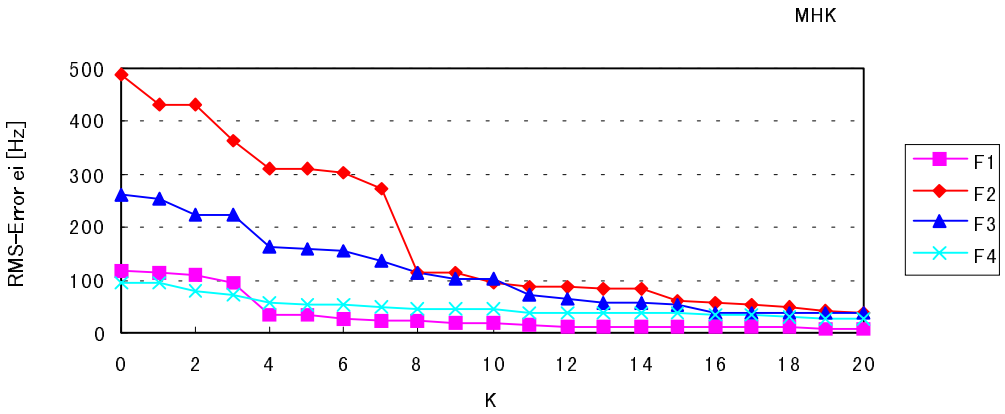

Fig. 1. Residual Errors.(MHK's formant trajectories)

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

$$F_i^{(K)}(n) = \frac{1}{2} C_i(0) + \sum_{k=1}^{K} C_i(k) \left[\frac{(2n+1)}{2N} k\pi \right], \qquad (7)$$

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

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

Figure 1 shows the residual errors of formant frequencies, F1 to F4, as functions of the order K. 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 smaller than 50 Hz. It is also shown that F2 and F3 vary more than F1 and F4,.

We synthesized speech stimuli by using one talker's dynamic features and another talker's static features.

3.3. Method Of Perceptual Judgement

An X-A-B judgement method was used, where X was a test stimulus which was synthesized by using a set of the parameters arranged from two different talkers, and A and B were synthetic speech signals generated from the original parameters of two talkers. Listeners were asked to judge whether X was A's or B's voice. Ten male listeners participated in the 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 and 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.

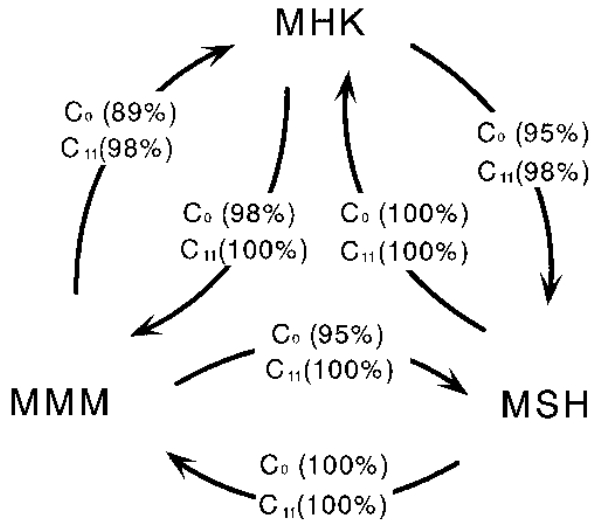

Fig. 2. Identification scores resulting from exchanging static and dynamic features between three talkers.

4. RESULTS AND DISCUSSION

Perceptual talker identification scores are depicted in Fig. 2. A total of 100 judgements (10 persons * 10 times) were made for each of the different talker combinations. Referring to Fig. 2, when MMM's dynamic features were combined with MSH's static features, 95% (or 95 times) of the cases results in MSH. In other words, only 5% of the time was identified as MMM's voice. On the other hand, if we combine MSH's dynamic features with MMM's static features, the resultant voice is identified as MMM's voice by 100% of the time. Other combinations show similar results. If we further add some of the dynamic features to the static features, the identification rate reaches nearly 100%. These show that the static features contribute much more to the perception of talker individuality than the dynamic features.

Table 1. Estimated average values of formant trajectories F1-F4.

Sentence	Subject	F1	F2	F3	F4
/aoiueoie/	MHK	435	1504	2473	3421
	MMM	426	1539	2431	3247
/aiueo/	MHK	457	1463	2451	3444
	MMM	443	1559	2428	3276
/ieoau/	MHK	435	1294	2468	3412
	MMM	439	1361	2396	3227
/aioiu/	MHK	417	1487	2350	3418
	MMM	396	1613	2353	3247
/eoiiau/	MHK	424	1563	2510	3404
	MMM	412	1610	2491	3239

In order to find general statistic properties of formant trajectories, we have selected another four sentences which are also composed of different combinations of Japanese vowels only. Table 1 presents estimated average values of formant trajectories of F1 to F4 from the five sentences uttered by MHK and MMM. From Table 1 we find that MMM has lower values of average F1,F3,F4 than MHK and a higher value of average F2 in general. We also find that the average F4 values among the five sentences are nearly the same for both MHK and MMM.

Static features of formant trajectories reflect the talker's anatomical constraints, while dynamic features reflect more the talker's speaking behaviour.

5. CONCLUSION

We investigated perceptual contributions of the vocal tract characteristics to talker individuality. The perceptual experiments show that the static (average) characteristic of the vocal tract is a primary cue to talker individuality. This implies that mapping of the average characteristic of the vocal tract would be one promising approach to talker conversion of a speech utterance.

ACKNOWLEDGEMENT

The authors would like to thank the subjects who participated in the experiments. This work was partly supported by the Ashigin International Foundation of Ashikaga Bank, Tochigi, Japan.

REFERENCES

[1] K. Itoh and S. Saito, *"Effects of acoustical feature parameters of speech on perceptual identification of speaker"*, *Trans. IEICE*, Vol. J65-A, pp. 101-108, 1982.(in Japanese)

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

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

[4] T. Kitamura and M. Akagi, *"Speaker individuality in speech spectral envelopes"*, *J. Acoustic. Soc. Jpn*. Vol(E)16, pp. 283-289, 1995.

[5] W. Ding, H. Kasuya and S. Adachi, *"Simultaneous estimation of vocal tract and voice source parameters based on an ARX model"*, *Trans. IEICE, Inf. & Syst.*, Vol. E78-D, pp. 738-743, 1995.

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

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