Analysis of Individual Characteristics in Vowel Spectral Envelopes

Renta Goto, Keita Misawa, and Yoshifumi Okada

***Abstract*—The purpose of this study is to understand frequency bands related to individual voice quality (individuality) when using the simple vowels of the Japanese language ("a", "i", "u", "e" and "o"). Differences in voice quality for individual simple vowels are determined by formant frequencies. In this study, we have attempted to emphasize the acoustic features of individuality, by masking the frequencies characterizing these simple vowels. Variance among speakers (VAS) in the spectral envelopes was used as an index for detecting individuality. In this paper, we surveyed and reported the individuality acoustic features for 10 Japanese men.**

***Index Terms*— Individual, masking, simple vowel, Variance among speakers, spectral envelope**

I. Introduction

CURRENTLY, there is an abundance of research for extracting the individuality of speakers [1], with the aim of applying this to the voice synthesis and personal identification fields. From existing studies [2]-[5], it has been reported that frequency bands containing individuality within the simple vowel spectral envelopes exist within the range of 1740 Hz to 8000 Hz.

When identifying simple vowels, the frequency of the formant is vital. The formant is the peak of the frequency spectrum, and each simple vowel has its own respective formant [6]. Typically, we are able to identify individuals to a certain extent from voice quality regardless of differences in simple vowels. This suggests that individuality is contained in frequency bands other than the formant (non-formant region). For this reason, it is considered that, by masking the formant peak frequency band (formant region), it is possible to emphasize the frequency bands containing individuality.

In this study, we mask the formants characterizing the simple vowels, and investigate the frequency bands containing the individuality characterizing the speakers. As an index to identify individuality, we use variance among speakers (VAS) that is variance in the logarithmic amplitude of each frequency point obtained from different speakers. This paper uses the voices of 10 male speakers and reports the results of investigating the frequency bands representing individuality.

Manuscript received January 12, 2017.

R. Goto is with the the Division of Information and Electronic Engineering, Muroran Institute of Technology, 27-1, Mizumoto-cho, Muroran, Hokkaido 050-8585, Japan (e-mail: renta@cbrl.csse.muroran-it.ac.jp).

K. Misawa is with the Division of Information and Electronic Engineering, Muroran Institute of Technology, 27-1, Mizumoto-cho, Muroran, Hokkaido 050-8585, Japan (e-mail: misawa@cbrl.csse.muroran-it.ac.jp).

Y. Okada is with the College of Information and Systems, Muroran Institute of Technology, 27-1, Mizumoto-cho, Muroran, Hokkaido 050-8585, Japan (corresponding author to provide phone: +81-143-5408; fax: +81-143-5408; e-mail: okada@csse.muroran-it.ac.jp).

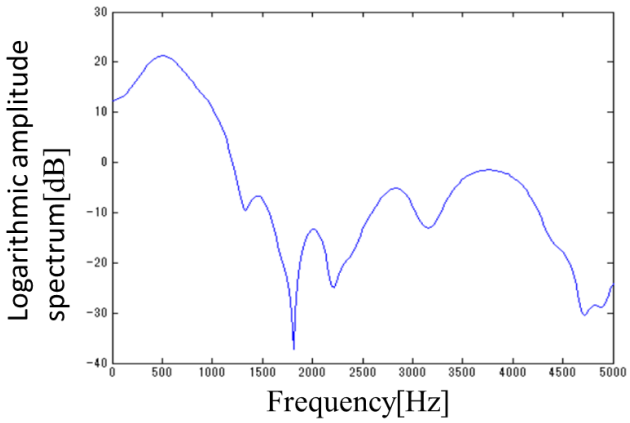

Fig. 1. Spectrum envelope

II. Spectral Envelopes

Fig. 1 gives an example of spectral envelopes. Spectral envelopes are two-dimensional graphs where the horizontal axis is frequency and the vertical axis is the logarithmic amplitude spectrum. The spectral envelope enables us to deduce the rough form of the frequency spectrum for the voice quality of the speaker.

III. Method

Fig. 2 shows the procedure for this study. In this study, through the procedure explained below, we calculate variance among speakers (VAS) from the voice quality of each subject, to detect the frequency bands most often contained in the individuality of the speaker.

A. Creation of Voice Samples

Voice samples of the Japanese simple vowels language ("a", "i", "u", "e" and "o") are collected from the subject. The subjects were ten Japanese men in their 20's. Each speaker was asked to sound out each simple vowel three times for 1 second or more. The voice samples were collected using an OLYMPUS-made linear OCM recorder (model number:LS-11). 1000 milliseconds were cut off each recorded voice sample, and these were divided into four with each sample being 250 milliseconds. In this way, for each simple vowel, 12 samples were created from each speaker

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

(120 samples for all speakers).

B. Linear Predictive Coding (LPC) analysis and Masking

The processing of this section is carried out only for masking. LPC analysis is applied to the samples cut from the recorded voice, and the formant region detected [7]. The five types of simple vowels are characterized according to the formant region and peak values for each spectral envelope. The range of each simple vowel format is set as follows.

$$f * 2^{-\frac{1}{6}} - f * 2^{\frac{1}{6}} \quad (1)$$

Here, f is the median value of the formant frequency obtained from the 120 samples of each simple vowel. In this study, based on the reference document [6], the first formant to the third formant was detected using the above-described range. The LPC coefficient and Fourier transform coefficient are set to 32 and 2048, respectively. Masking is performed by attenuating the above-described ranges using the band elimination filter for each voice sample.

C. Extraction of the Spectral Envelopes

The processing from this section is carried out both with and without masking.

First, the Fourier transform is performed after applying the window function to each sample. Next, the cepstrum is acquired by applying the inverse Fourier transform to the logarithmic amplitude spectrum. Finally, the spectral envelope is extracted by performing the Fourier transform again on the cepstrum after filter processing. The Fourier coefficient and cepstrum coefficient for spectral envelope extraction is set to 2048 and 60, respectively.

D. Calculation of VAS

VAS for the logarithmic amplitude spectrum for each frequency point in each simple vowel is calculated for the simple vowel spectral envelopes obtained in III. If the VAS is large, there is a large disparity in the frequency points between individuals. That is to say, this means that it contains individuality. VAS for each frequency point for each simple vowel is calculated per the following formula.

$$VAS = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 \quad (2)$$

Here, n is the number of samples, x_i is the logarithmic amplitude spectrum for sample i and m is the mean value of the logarithmic amplitude spectrum for each frequency point.

IV. Experiment

A. Preliminary experiment

In this study, we suppose that, rather than formant regions expressing the simple vowel itself, non-formant regions contain more individuality. Here, to verify this supposition, the following preliminary experiment is carried out for each simple vowel.

First, we calculated the logarithmic amplitude spectrum VAS at each frequency point for the 120 spectral envelopes related to each simple vowel. Following this, to investigate whether there was a significant difference in the VAS between formant regions and non-formant regions, we used the Mann–Whitney U test. If the VAS was more significant among non-formant regions, the non-formant region was considered to contain more individuality.

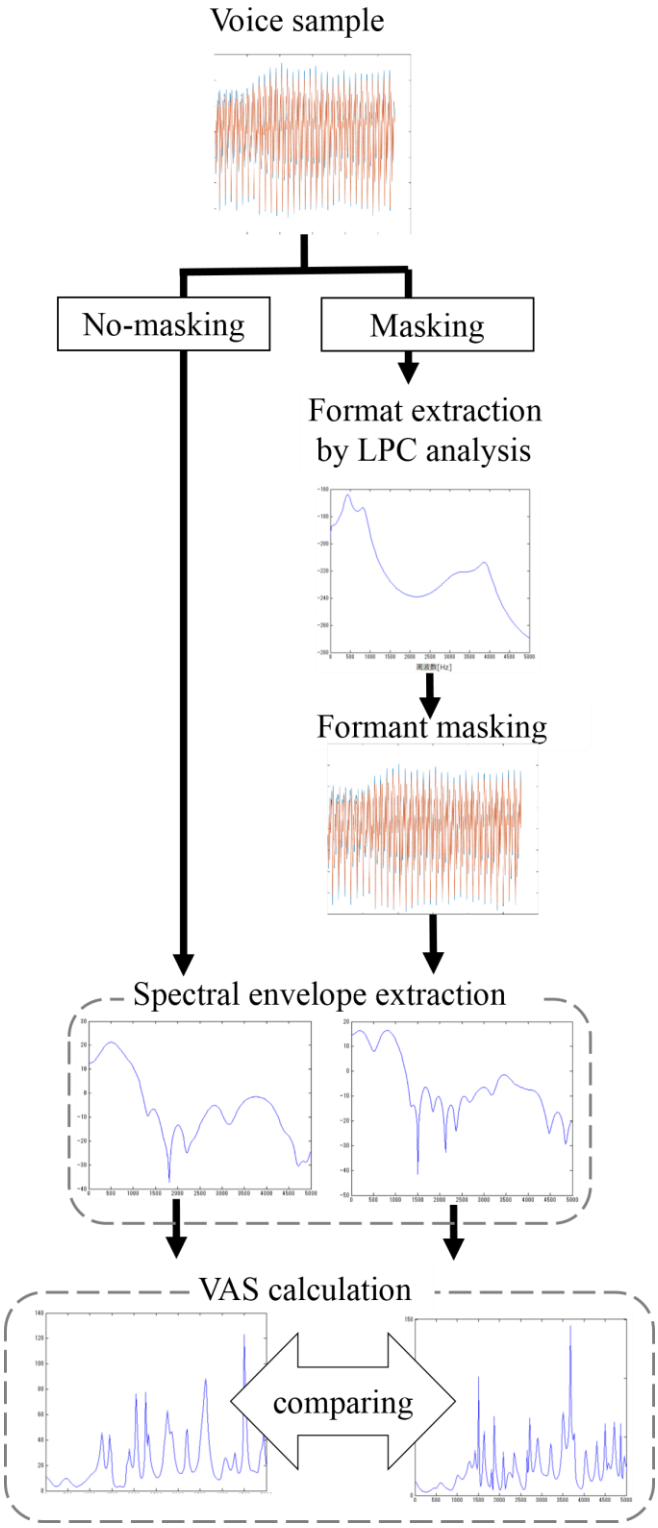

Fig. 2. Procedure of this study

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Comparison of Distribution Values for Masking and Non-masking

If the supposition in section IV is correct, by masking the formant region, it is considered that the features of other frequency bands shall clearly stand out. Here, we conducted an experiment to verify this supposition. In concrete terms, to survey whether there are any differences in VAS with or without masking for each simple vowel, the Wilcoxon signed-rank test was used. The frequency band range for mask processing was set to the same as that of the preliminary test.

Detection of Frequency Bands Containing Individuality

For each simple vowel, we used the VAS to detect the frequency bands using individuality according to the following procedures. We calculated the difference d for each frequency point between VAS when there was masking and VAS when there was no masking. When the difference d in the frequency points was bigger than $m_d + 1SD$, the frequency point is seen as including individuality. Here, m_d is the mean of difference d, and SD is the standard deviation of difference d for each frequency point.

V. Results and Discussions

A. Preliminary Experiment

Significant differences were observed in all the “a”, “i”, “u”, “e” and “o” simple vowels used in the preliminary experiment (significance standard α=0.05). This result demonstrated that there are differences in the VAS at the various frequency points between formant regions and non-formant regions. That is to say, it shows that information indicating individuality exists in the non-formant region.

B. Main Experiment

Comparison of Variance Values With and Without Masking

Fig. 3 shows the VAS for the simple vowel “a”. The broken line in the diagram shows the result without masking and the solid line is the result with masking. From this diagram, we can see that multiple frequency bands exist in which the variance increases when masking is used. In particular, there is a significant variance in the frequency bands between 2500 Hz and 4500 Hz. From this, it can be considered that the features of the simple vowel itself are counteracted by the formant masking, and the voice quality features of the respective speakers stand out. These trends were observed in the same way for all the simple vowels other than “u”.

Fig. 4 is the VAS for simple vowel “u”. From this diagram, we can see that the variance increases in many of the frequency bands. In particular, the variance values were large in the frequency bands of 4500 Hz or more. This result is greatly different to that of the other simple vowels. The cause of this is thought to be that the third formant frequency for the simple vowel “u” band intermittently exists in a wide range. The masking range was carried out in relation to a continuous specific frequency. In the future work, we will introduce a masking method for the intermittently existing formant region.

Detection of Frequency Bands Containing Individuality

Fig. 5 shows frequency bands containing specified individuality based on VAS. The horizontal axis is the frequency and the black represents the frequency bands judged to include individuality and the white represents the other frequency bands. From this diagram, we can see that with the simple vowels “a”, “i”, “e” and “o”, the frequency bands containing individuality appear concentrated within a particular range (approximately 3500 Hz to 4500 Hz). For simple vowel “u”, frequency bands containing individuality appear as the others are seen near 4000 Hz, but were not observed in any other frequencies. The cause of this is that in frequencies other than 4000 Hz, as both the mean m_d and standard deviation SD are large, the threshold is not satisfied. This is thought to be due to the existence of intermittent formant frequencies as explained in the previous section.

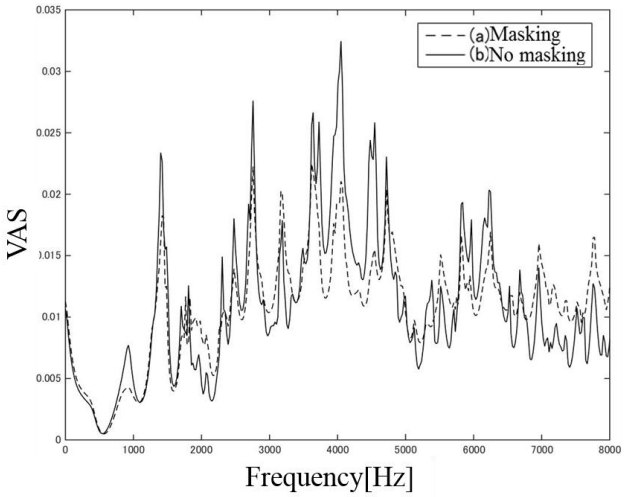

Fig. 3. VAS of the simple vowel “a”

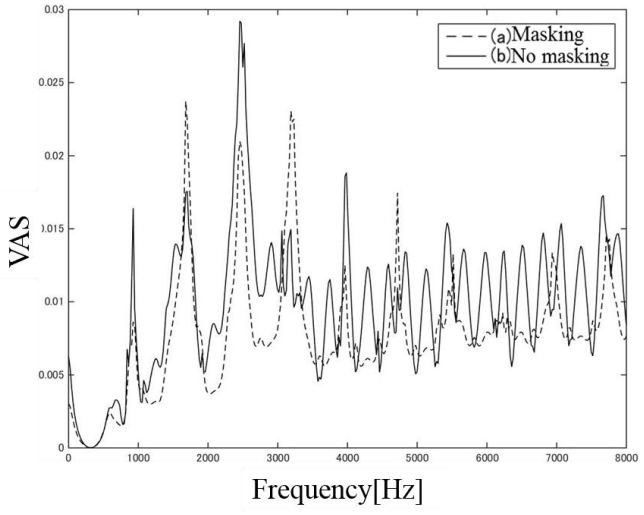

Fig. 4. VAS of the simple vowel “u”

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

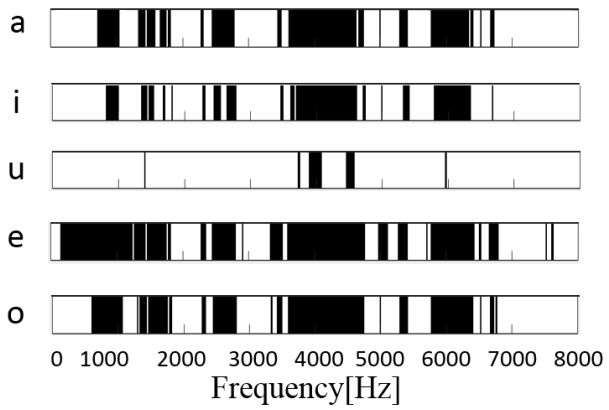

Fig. 5. Frequency is including individuality

VI. Conclusion

The objective of this study was to understand features demonstrating individuality from simple vowel spectral envelopes. This basic idea is that the acoustic features of the individuality can be emphasized by masking the formant expressing the features of the simple vowel itself. First, in the preliminary experiment, for all the “a”, “i”, “u”, “e” and “o” simple vowels, it was shown that individuality may exist in the non-formant regions. Furthermore, in the main experiment, it was shown that for simple vowels other than “u”, frequencies in which differences between speakers were significant, that is to say frequency bands that have information for distinguishing individuals are included within the range from approximately 3500 Hz to 4500 Hz.

In this study, we dealt with the simple vowels of men. In the future, we will also conduct experiments on the simple vowels of women. In addition, it is also necessary to change the range in which masking is performed, and examine the trends of variance among speakers (VAS). Moreover, the development of a speaker recognition method using information obtained from this study is a vital issue.

References

[1] Hiroya Fujisaki, “Prosody, models, and spontaneous speech,” in Sagisaka et al. eds., Computing Prosody. Springer, 1997, pp. 27–42.
[2] Tatsuya Kitamura, Masato Akagi, “Speaker individualities in each spectral envelopes,” J. Acoust. Soc. Jpn. (E), 1995, pp. 83–289.
[3] K.-P., Li, G.W. Hughes, “Talker differences as they appear in correlation matrices of continuous speech spectra,” J. Acoust. Soc. Am., vol. 55, 1974, pp. 833–873.
[4] Parham Mokhtari, Frantz Clermont, “Contributions of selected spectral regions to vowel classification accuracy,” n Proc. ICSLP’94, 1994, pp. 1923–1926.
[5] Adrian Leemann, Marie-José Kolly, Volker Dellwo, “Speaker-individuality in suprasegmental temporal features: Implications forforen sicvoice comparison,” Forensic Science International, Vol. 238, 2014, pp 59-67.
[6] Ray D. Kent, Charles Read, *The Acoustic Analysis of Speech.* Singular Publishing Group, Inc, 1992, pp.22-34.
[7] D.O'Shaughnessy, "Linear Predictive Coding", IEEE Potentials. vol. 7 no. 1, 1998, pp. 29-32.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)