**2006 IEEE International
Symposium on Signal Processing
and Information Technology**

Speaker Verification Using Narrow-band Envelope Correlation Matrices

S. Gotoh[1]*, M. Kazama*[2]*, M. Tohyama*[3]*, and Y. Yamasaki*[1]

Global Information and Telecommunication Studies
Waseda University
Shinjuku-ku, Tokyo, Japan

Abstract

We confirmed that a speaker's vocal individuality is contained in the inter-band correlations of narrow-band (1/4 or 1/8 octave bands) temporal envelopes. Two types of envelope correlation matrices (ECMs) were made for 53 speakers, using three utterances of an identical sentence (assuming a situation where a password for verification was stolen) so that any differences in the spoken contents might not greatly influence their individuality. Type-A (reference) ECMs of two of the utterances were constructed to make a speaker's individual template, and a type-B ECM was constructed using the other utterance. Speaker matching tests between the two types of ECMs, based on Gaussian mixture model (GMM) matching scores, verified the validity of the individual speakers. In particular, a speaker's voice could be verified using spoken materials through the telephone band (250 Hz- 3 kHz), a high frequency range (2- 11.3 kHz), or a wide frequency range (250 Hz- 11.3 kHz).

1. Introduction

This article discusses our investigation of a speaker's individual voice signature in the inter-band correlations of narrow-band temporal envelopes. Speaker identification or verification is an attractive application for speech technology [1-3]. Bimbot et al. [4] suggested that temporal changes in the short-term power spectrum contain a signature 4-8 kHz higher than in the telephone-band. We noticed that inter-band correlations of the narrow-band envelopes [5], which correspond to the temporal changes in the short-term power spectrum, might represent individual vocal signatures. Correlations in the frequency domain could be interpreted as a cepstral analysis, if we take the envelopes in a dB scale.

We conducted speaker-matching tests on 25 male and 28 female speakers according to the similarities in the envelope correlation matrices (ECMs) from among the speakers. The ECMs were constructed from the utterances of an identical single sentence spoken by every speaker in a wide frequency range over 6 kHz. The matching results based on GMM using the ECMs indicate that a speaker's signature could be represented by an ECM in the telephone band, as well as in the wide (250 Hz- 11.3 kHz) or high (2- 11.3 kHz) frequency ranges.

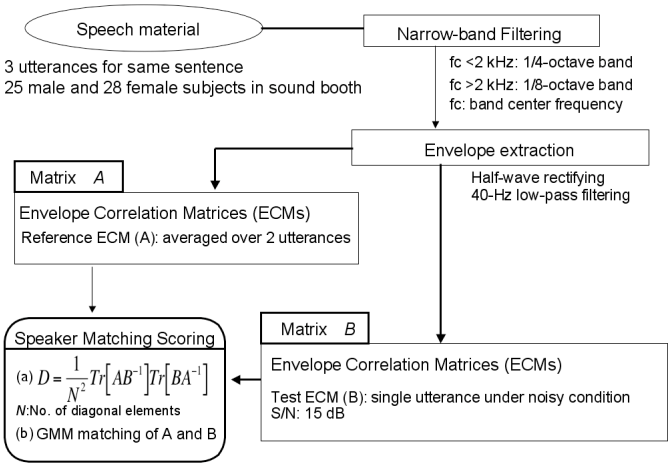

Figure 1: Speaker-to-speaker matching test

2. Narrow-band speech envelope analysis

All the subjects were native Japanese and a single identical Japanese sentence spoken by each subject was recorded in a sound booth at a sampling rate of 48 kHz, using a 16-bit A/D converter. Three takes were recorded for every subject. The sentence was easily readable, used daily, and about two seconds long when spoken.

The speech signals were divided into 21 1/8-octave bands between 2 and 11.3 kHz and divided into 12 1/4-octave bands between 250 Hz and 2 kHz using a filter bank (fourth-order Butterworth; MATLAB's signal-toolbox "butter"). The envelope of each band signal was obtained through a low-pass filter with a 40-Hz cut-off frequency after half-wave rectification. Then, we derived the inter-band correlation matrices (33×33) of the envelopes for each speaker. The

[1]Graduate School of Global Information and Telecommunication Studies, Waseda University
[2]Spatial Science Research Institute, Waseda University
[3]Global Information and Telecommunication Institute, Waseda University

envelope correlation is defined by [5]

$$\rho(i,j) = \frac{\overline{E_i(n)E_j(n)}}{\sqrt{\overline{E_i^2(n)}\ \overline{E_j^2(n)}}} \tag{1}$$

where

$$E_l(n) = E_{l_0}(n) - \overline{E_{l_0}(n)}. \tag{2}$$

The $\overline{*}$ denotes the long time average, $E_{l_0}(n)$ represents the l-th band envelope, and n is the sampled time. We derived two types of envelope-correlation matrices (ECMs) for each speaker. One (reference-ECM) is used for the speaker's signature template, which was derived from the two utterances, and the other one (test-ECM) was made for the speaker-matching tests using a single utterance that included surrounding noise.

3. Speaker matching tests according to ECM

We conducted speaker-to-speaker matching tests using the ECMs under noisy conditions (S/N: 15dB) (Fig. 1). Figure 2 illustrates the power spectrum for the environmental noise. Figure 3 shows the samples of the speakers' reference-ECMs. The results indicated that the ECMs could have individual vocal differences. We used a test-ECM (B) for every speaker, using a single utterance that was not used for reference-ECM (A). The reference ECM was made by averaging the two utterances.

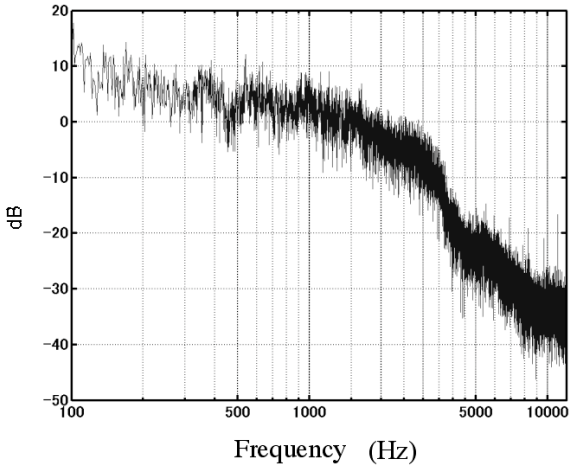

Figure 2: Environmental noise power spectrum

Scoring the similarities between the test- and reference-ECMs can be done according to the harmonic sphericity measure using [4, 6]

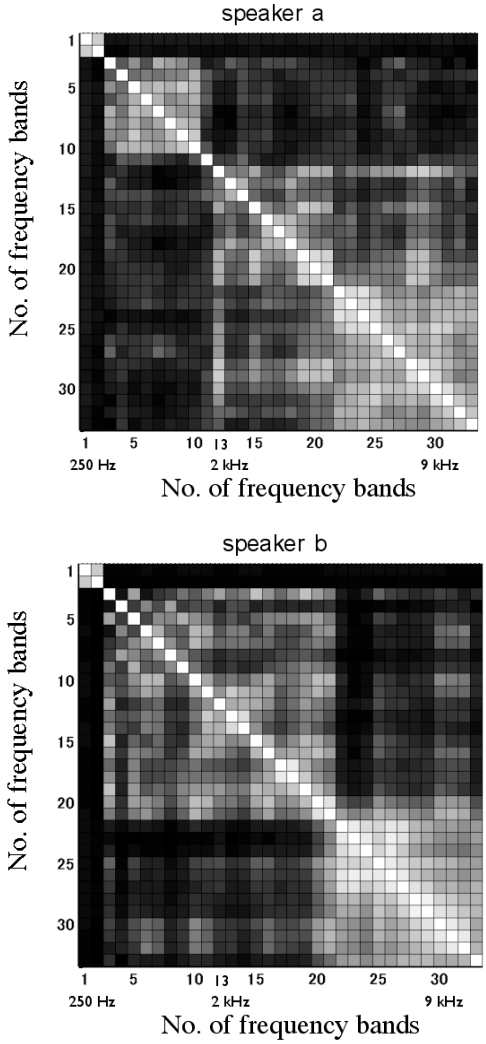

Figure 3: Samples of reference ECMs

$$D = \frac{1}{N^2}\mathrm{T_r}[AB^{-1}]\mathrm{T_r}[BA^{-1}] \tag{3}$$

where A and B denote the reference- and test-ECMs to be compared, respectively, and N is the number of diagonal elements.

We also performed GMM matching tests using the ECMs [7]. For this purpose we prepared two reference ECMs for each of the two utterances, where the second utterance's ECM was used for training purposes. The matching score for a speaker was normalized by cohort normalization [8]. The matching test outputs a log likelihood as a matching score. If we take the unnormalized verification scores as

$$score_u = \log \widehat{p}(\mathbf{O}|I) \tag{4}$$

where $\widehat{p}$ represents a likelihood estimate, **O** denotes the sequence of the column vectors of the ECM, and I is the speaker's assigned number, then the cohort normalized scores can be defined by a difference in a log likelihood

$$score_n = \log \widehat{p}(\mathbf{O}|I) - \max_k[\log \widehat{p}(\mathbf{O}|c_k(I))] \quad (5)$$

where $\log \widehat{p}(\mathbf{O}|c_k(I))$ is the log likelihood of the observation vector sequence for the model of the k-th speaker in the cohort $C(I)$ assigned to I.

Figure 4 shows the D^{-1} (1st and 3rd panels from the top) scores and the GMM matching scores (2nd and 4th panels from the top). The matching tests were performed separately for the males and females. The results show that the ECMs contain an individual speaker's information. The GMM matching scores more clearly revealed the individual vocal differences in the ECMs for the male (a) and female (b) speakers than the scores by the harmonic sphericity measure. Therefore, we can verify all the speakers in this experiment, if the GMM matching score is positive.

4. Speakers' verification according to different ECM frequency ranges

As described in the introduction, a speaker's vocal signature analysis based on spectral and temporal dynamics requires a wide frequency band (including high frequency) [4]. We prepared three types of ECMs with different frequency ranges, one with a low frequency range (250 Hz- 2 kHz), one with a high frequency range (2-11.3 kHz), and one in the telephone band (250 Hz- 3 kHz). Figure 5 shows the matching tests results. Panel (a) shows the verification scores by GMM (with cohort normalization) for the low frequency range, panel (b) represents those for the high frequency range, and panel (c) illustrates the results using the telephone band.The results indicate that the ECMs in the higher frequency range contain a speakers' individuality, not those in the low frequency range. The results also suggest that if the range includes the frequencies over 2 kHz, we are able to verify all the speakers even for the telephone band.

5. Conclusion

We confirmed that the inter-band correlations of narrow-band temporal envelopes contain a speaker's vocal signature. Two types of envelope correlation matrices (ECM) were made for 53 speakers; (a) matrices were prepared from two utterances for a speaker's individual template and (b) single-utterance-based matrices. All the spoken sentences

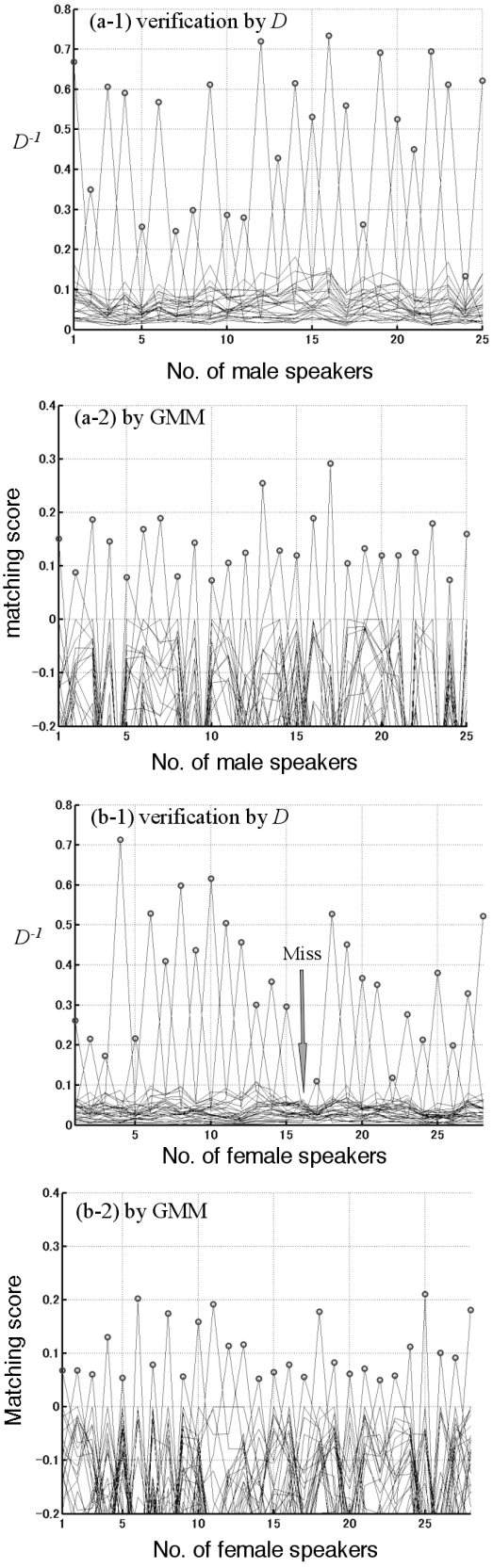

Figure 4: Speaker matching test results for wide frequency-range speech (250 Hz - 11.3 kHz) for (a) males and (b) females

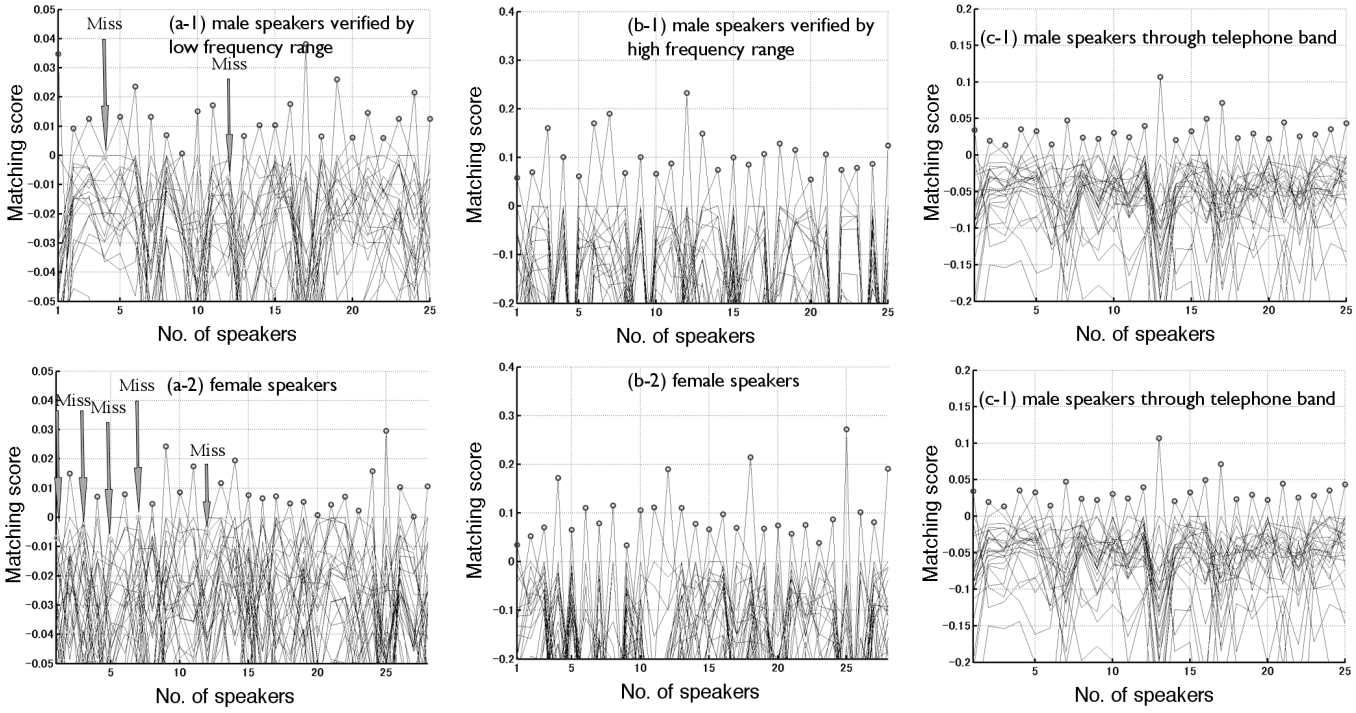

Figure 5: Frequency range and speaker verification by GMM; (a) Low frequency range (fc $<$ 2 kHz), (b) High frequency range (fc $>$ 2 kHz), and (c) Telephone band (250 Hz$<$ fc $<$ 3 kHz)

were identical. It was proven that ECMs were able to distinguish the vocal individuality of a speaker, even in the telephone band, by matching the tests between the (a) and (b) matrices. We also found that GMM matching scores are better for speaker verification than harmonic sphericity measures.

6. Acknowledgment

The authors would like to thank Mr. T. Taniguchi for his valuable discussions on GMM.

7. References

[1] T. F. Quatieri, D. A. Reynolds, and G. C. O'Leary, "Estimation of Handset Nonlinearity with Application to Speaker Recognition," IEEE SAP 8(5) pp. 567-584 (2000)

[2] M. S. Zilovic, R. P. Ramachandran, and R. J. Mammone, "Speaker Identification Based on the Use of Robust Cepstral Features Obtained from Pole-Zero Transfer Functions," IEEE SAP 6(3) pp. 260-267 (1998)

[3] Jo-Anne Bachorowski and Michael. J. Owren, "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech" J. Acoust. Soc. Am., Vol. 106 No. 2, pp. 1054-1063 (1999)

[4] F. Bimbot, I. Magrin-Chagnolleau, and L. Mathan, "Second-order statistical measures for text-independent speaker identification", Speech Communication, 17, pp. 177-192 (1995)

[5] Steven van de Par and Armin Kohrauch, "Analytical expressions for the envelope correlation of narrowband stimuli used in CMR and BMLD research", J. Acoust. Soc. Am. Vol. 103, No. 6, pp. 3605-3620 (1998)

[6] R. D. Zilka, "Text-Independent Speaker Verification Using Utterance Level Scoring and Covariance Modeling." IEEE SAP 10(6), pp. 363-370 (2002)

[7] Joseph P. Campbell, JR., g Speaker Recognition: A Tutorialh Proc. IEEE vol. 85, pp. 1437-1462, Sep (1997)

[8] A. E. Rosenberg, J. DeLong, C.-H. Lee, B-H. Juang, and F. K. Soong, gThe use of cohort normalized scores for speaker verification,h in Proc. Int. Conf. Spoken Language Processing, University of Alberta, Canada, pp. 599-602 (1992)