IEICE TRANS. INF. & SYST., VOL.E91–D, NO.3 MARCH 2008

PAPER Special Section on Robust Speech Processing in Realistic Environments

Speaker Verification in Realistic Noisy Environment in Forensic Science

Toshiaki KAMADA[†,††a)], ***Student Member***, **Nobuaki MINEMATSU**[††b)], **Takashi OSANAI**[†], ***Members***, **Hisanori MAKINAE**[†], ***and*** **Masumi TANIMOTO**[†], ***Nonmembers***

SUMMARY In forensic voice telephony speaker verification, we may be requested to identify a speaker in a very noisy environment, unlike the conditions in general research. In a noisy environment, we process speech first by clarifying it. However, the previous study of speaker verification from clarified speech did not yield satisfactory results. In this study, we experimented on speaker verification with clarification of speech in a noisy environment, and we examined the relationship between improving acoustic quality and speaker verification results. Moreover, experiments with realistic noise such as a crime prevention alarm and power supply noise was conducted, and speaker verification accuracy in a realistic environment was examined. We confirmed the validity of speaker verification with clarification of speech in a realistic noisy environment.

key words: *text-dependent speaker verification, narrow-band noise, clarification of speech, missing feature theory, forensic science*

1. Introduction

The speaker verification that we are carrying out targets voice telephony from the viewpoint of forensic science, which is field in which our research results are requested. Regarding voice samples in forensic science, there are many poor conditions compared with those in general research in a noisy environment. In cases where speech contents cannot be understood, we clarify the speech in order to understand the content. However, the research results of speaker verification with the clarification of speech are unsatisfactory. Nevertheless, speaker verification in a very bad noisy environment might actually be required. It seems that the term "the clarification of speech" is not commonly used in general research, because general research in such a noisy environment that the speech content cannot be understood is not demanded. However, there is often such a demand in forensic science. The noise reduction processing in such a situation is called the clarification of speech. Therefore, this noise reduction processing is called the clarification of speech in this paper.

The characteristics of noise cannot be examined off-line in general research on the speech or speaker recognition in a noisy environment, and a real-time result of processing is demanded. However, in forensic science, the characteristics of the noise can be adequately examined off-line, and there is little demand for real-time processing. Therefore, after the characteristics of the noise can be examined in detail, the best filter and parameters for the clarification of speech can be decided. Moreover, the data for comparison, which is usually obtained under good conditions with little noise after a suspect was arrested, can often be re-collected, which is one of the distinctive features in forensic science, although the sample of the unknown speaker, which might have been recorded in a noisy environment, cannot be collected again.

In this study, a text-dependent speaker verification experiment with the clarification of speech was conducted using speech recorded in a noisy environment that filled the condition of forensic science, and the relationship between the verification result and speech clarification was examined. We dealt with narrow-band (periodicity) noise as the object noise, and the verification results for different characteristics of the noise were compared, because a crime prevention alarm and power supply noise, which is a narrow-band noise, are a scene that most appears quite frequently in forensic science and that are demanded from the viewpoint of investigation support. From a thing of above, we deal with these two kinds of noise as real environmental noise, and we concentrate on the case of noise reduction processing for this narrow-band noise in this study. The verification experiment with a sine wave noise which has one frequency was conducted before with a crime prevention alarm and power supply noise which has multiple frequency elements. We will describe the possibility of speaker verification with speech clarification of speech including the narrow-band noise that SNR is less than 0 dB bellow.

2. Speaker Verification

2.1 Speech Corpus

In the speaker verification experiments, we used a speech of utterances by approximately 3000 Japanese adult males recorded through a telephone. The speech data of 300 people chosen from this speech corpus was used for the speaker verification experiment. The speech data shown in Table 1 were recorded at the sampling frequency of 11.025 Hz, but were downsampled from 11.025 kHz to 8 kHz before the experiment because of being recorded through the telephone.

Manuscript received July 2, 2007.
Manuscript revised September 19, 2007.

[†]The authors are with the National Research Institute of Police Science, Kashiwa-shi, 277–0882 Japan.
[††]The authors are with the Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa-shi, 277–8562 Japan.
a) E-mail: kamada@nrips.go.jp
b) E-mail: mine@gavo.t.u-tokyo.ac.jp

DOI: 10.1093/ietisy/e91–d.3.558

Table 1 Speech data.

Number of speakers	300 people
Time difference of utterances	3–4 months
Number of utterances	3 times
Repetition of utterances per day	3 times
Utterance contents (Japanese)	5 vowels (*a, i, u, e, o*) 5 words (*KURUMA, RENRAKU, BAKUDAN, GINKO, KEISATSU*)
Sampling frequency	8 kHz
Quantization	16 bit

Table 2 Analysis conditions.

Analysis window	Hamming window
Frame length	32 ms
Frame shift	16 ms
Preemphasis	1st-order adaptive
Analysis method	LPC analysis
Analysis order	12th-order
Feature	LPC cepstrum

2.2 Dynamic Time Warping

The analysis shown in Table 2 was performed, and dynamic time warping (DTW) [1] was adopted as the method of calculating the euclidean distance between the unknown speaker's data and the contrast (known speaker's) data. In the general research of speech recognition and speaker recognition, features such as mel-cepstrum and MFCC and analysis methods of model-based statistics and probability such as VQ and GMM are usually used [2]–[4]. There are few speech data and a case that construction of a statistical model is difficult, the method of DTW is effective in the field of forensic science. The parametric method as LPC sometimes is effective also than the nonparametric method still more in a noisy environment.

In this work, we used the method adopted in conventional research, not aiming at the improvement of the verification accuracy using the features and the analysis methods, and aiming the clarification of the relationship between the clarified speech and the verification accuracy, although we had already studied these issues employing the these features and analysis methods [5]–[7].

2.3 Speaker Verification Method

For the text-dependent speaker verification, the distance between data of the same speech content spoken by various speakers was calculated. The combination of speech data obtained on the same day was not used for one person. The first time of utterance and the second time, the second time and the third time were compared, and the distance distribution for one speaker (among speaker) was obtained from 300 people. The distance between one speaker and another speaker, using a different combination of data was similarly obtained. This combination was thinned out.

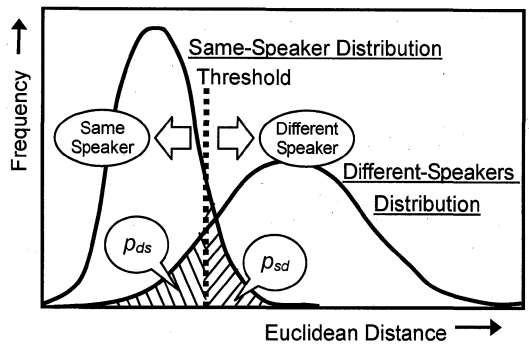

Fig. 1 Method of judgment.

Figure 1 shows pattern diagrams of the method of judgment. An equal error rate ($p_{ds} = p_{sd}$) was obtained from the same-speaker and different-speakers distance distributions for each utterance content. The threshold of judgment was calculated, and this threshold was assumed to be the criterion for judging whether utterances were spoken by the same speaker. The correct answer rate when using this criterion was assumed to be the verification rate of the speaker verification experiment.

3. Speaker Verification in Narrow-Band Noisy Environment

3.1 Speech Data

Because our purpose was to clarify the relationship between the noise power band (center frequency and bandwidth), which is a characteristic of the noise, and the clarification of speech and the speaker verification, we decided to use the sine wave with one frequency element as the narrow-band noise actually treated, although there is much noise with two or more narrow-band spectral elements.

When root mean square (RMS) of the speech signal $s(t)$ in a frame section from time t_0 to time t_N is S_{RMS} and RMS of the noise signal $n(t)$ is N_{RMS}, the speech signal in a noisy environment $x(t)$, which becomes SNR [dB], is obtained as

$$x(t) = s(t) + \frac{S_{RMS}}{10^{SNR/20} N_{RMS}} n(t)\,, \tag{1}$$

$$S_{RMS} = \sqrt{\frac{1}{N} \sum_{t=t_0}^{t_N} s^2(t)}\,, \tag{2}$$

$$N_{RMS} = \sqrt{\frac{1}{N} \sum_{t=t_0}^{t_N} n^2(t)}\,. \tag{3}$$

The obtained speech data was assumed to be the unknown speaker's data in a noisy environment. Because we assume it is different from a place where a speaker uttered and a place where the speech was recorded and where a noise was gained, we think that consideration for the Lombard effect is not necessary.

3.2 Band Elimination Filter

We used a band elimination filter (BEF) to clarify speech

with sine wave noise. In this study, we used the Kaiser filter with sox (sound exchange) [8] because the Kaiser filter is conventionally used as the FIR filter.

When the eliminated bandwidth of BEF is f_D [Hz] for the f sine wave noise, the elimination band of BEF is between $(f - f_D/2)$ and $(f + f_D/2)$. The gain of each cutoff frequency is −6 dB and the attenuation rate is adjusted to the width of the Kaiser window appropriately so as to eliminate the noise.

3.3 Speaker Verification Experiments Based on Missing Feature Theory

The unknown speaker's data was speech in an environment with 2 kHz sine wave noise, and the contrast data was original speech data basing on the telephonic band and a sampling frequency. SNR was set to 10, 0, and −10 dB, the speech clarification of the unknown speaker's data was carried out with BEF, and the speaker verification experiments with the contrast data were run. Table 3 shows the verification condition of the unknown speaker's data in the experiment. We focus on the verification result of the variable eliminated bandwidth of BEF, and the experiment under the condition of not using BEF (this is the 'Noise' condition) for the unknown speaker's data was run for comparison. Moreover, an experiment under the condition that noise was not superimposed onto the unknown speaker's data (this is the 'Clear' condition) was carried out to examine the influence of BEF processing on the verification rate. It is not our aim to bring the verification rate after speech clarification close to 100%, but in this study, it is improved to the verification accuracy achieved under the 'Clear' condition, that is, the condition without noise.

We proposed to use the Missing Feature Theory for the improvement of the verification accuracy [9], [10]. We examined how the verification accuracy was changed upon processing with BEF (missing feature mask (MFM) processing) the contrast speaker's data in a manner similar to the processing of the unknown speaker's data. In MFM processing of the contrast data, the sine wave noise present in the unknown speaker's data was not added to the contrast data. As the 'Noise' condition for MFM, noise was added to the contrast data such that SNR became the same as that of the unknown speaker's data.

3.4 Experimental Results

Figure 2 shows the verification results. The improvement of the verification rate upon using MFM processing of contrast data is shown, and it was confirmed to improve greatly, particularly when the eliminated bandwidth was 1 kHz under the both the condition with BEF processing and without BEF processing. When the eliminated bandwidth is about 600 Hz or less, the verification accuracy close to the target value under the condition without the noise is obtained because it is less affected by BEF processing. We found that the verification accuracy is the most affected by the condition of applying a MFM in the contrast data.

Next, the unknown speaker's data was speech in an environment with 1, 2 and 3 kHz sine wave noise, and the contrast data was processed with BEF. SNR was set to −10 dB. The other experimental condition is same in Sect. 3.3. Figure 3 shows the verification results. When the eliminated bandwidth is about 600 Hz or less, the verification accuracy did not appear to be affected so much by the characteristic of sine wave noise.

Table 3 Verification experiment conditions.

Noise	no BEF processing
200 Hz – 2 kHz	BEF processing (f_D = 200 Hz – 2 kHz)
Clear	SNR = ∞ (target value)

4. Speaker Verification in Realistic Noisy Environment: Crime Prevention Alarm

4.1 Speech Data

In Sect. 3, we described speaker verification in the case of a simulated noisy environment. However, in this section, we present speaker verification in a realistic noisy environment with a crime prevention alarm. The alarm was a narrow-band noise that had a frequency element of about 3,200 Hz. This frequency element could be removed by BEF processing with the 400 Hz bandwidth. The sound spectrograph of the alarm alone and the alarm after BEF processing are shown in Fig. 4. Figure 4 shows that the noise component can be removed by the BEF. The spectrum of the alarm is shown in Fig. 5. We prepared the speech data in the realistic noisy environment with SNR set to 10, 0, and −10 dB.

4.2 Experimental Results of Speaker Verification in Realistic Noisy Environment: Crime Prevention Alarm

The speaker verification experiments with the clarification of speech were conducted under the verification condition shown in Table 4. The methodology of the verification experiments is the same as described in Sect. 3. In these experiments, the condition that alarm noise is add to the contrast data was also examined because this alarm noise comprises not only the narrow-band noise but also the wide-band noise component. Figure 6 shows the experimental results. The same kind of noise component is made to function as the MFM that is added to the contrast data before BEF processing.

4.3 Discussion

These results indicate that the verification rate is improved by adding the alarm noise to contrast data before BEF processing, but not as much as to the target value. The difference from the target value is 10% or more under the 10-dB-SNR condition. A high verification rate is not obtained, al-

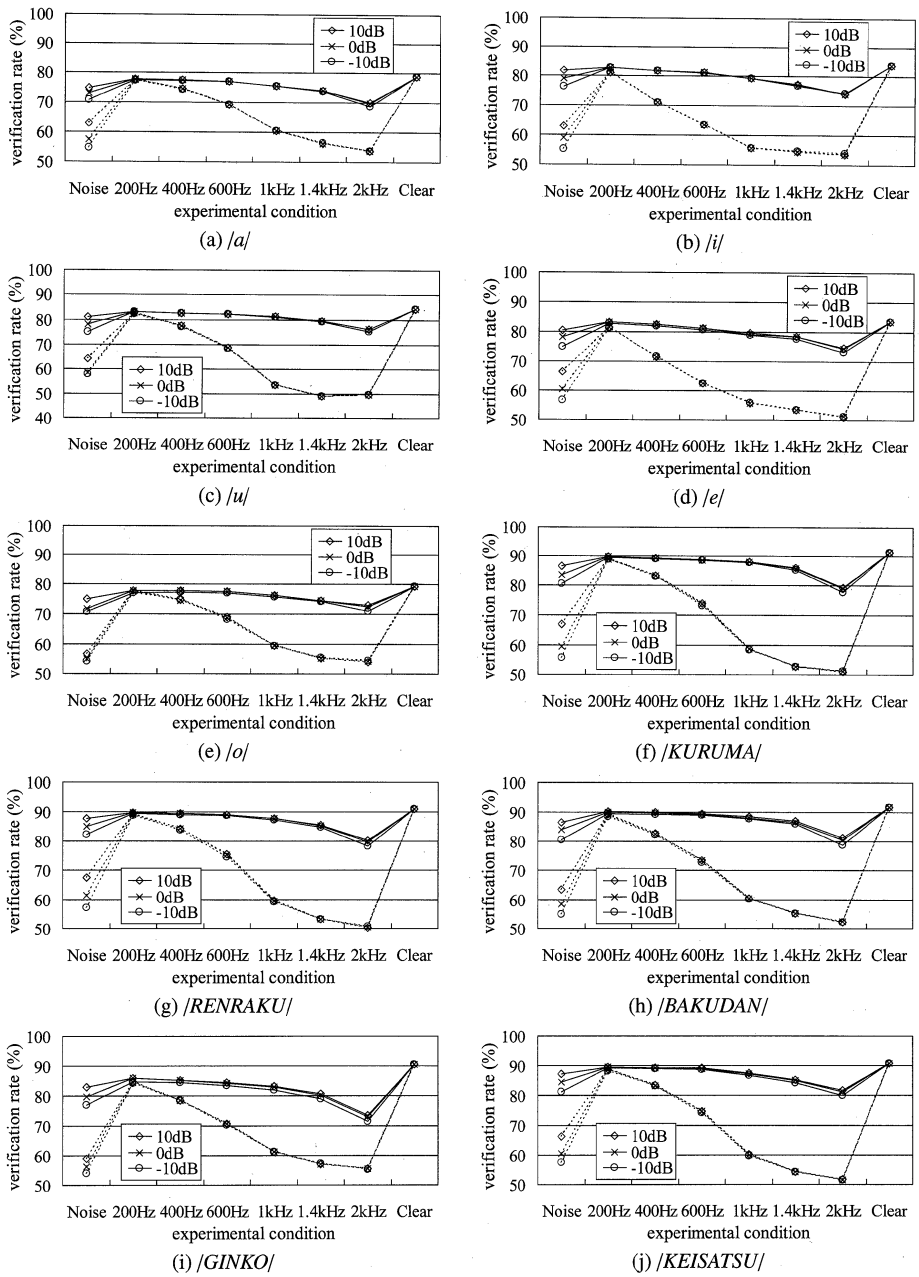

Fig. 2 Results of verification experiment: (i) MFM processing of contrast data, shown by the solid line; (ii) no MFM processing of contrast data, shown by the dotted line.

though Fig. 4 shows that alarm noise reduction is successful. It is clear that because the wide-band noise is not removed, a higher verification rate is no obtained. In this study, we do not consider the wide-band noise in the clarification of speech or noise reduction.

It is found that speaker verification by this method with wide-band-noise processing is very difficult, from the existing work [11], [12]. If the wide-band noise is not included in the crime prevention alarm noise, it appears to be certain that a verification accuracy near the target value can be obtained, on the basis of the result in Sect. 3 and the success of alarm noise reduction.

As a consequence of the above results, it has been found that to obtain high verification accuracy in speaker verification in a realistic noisy environment, it is important to examine the characteristics of the noise for successful noise reduction.

5. Speaker Verification in Realistic Noisy Environment: Power Supply Noise

5.1 Speech Data

In this section, we presented speaker verification in a realistic noisy environment with power supply noise of 50 Hz, in order to examine the relationship between the clarification

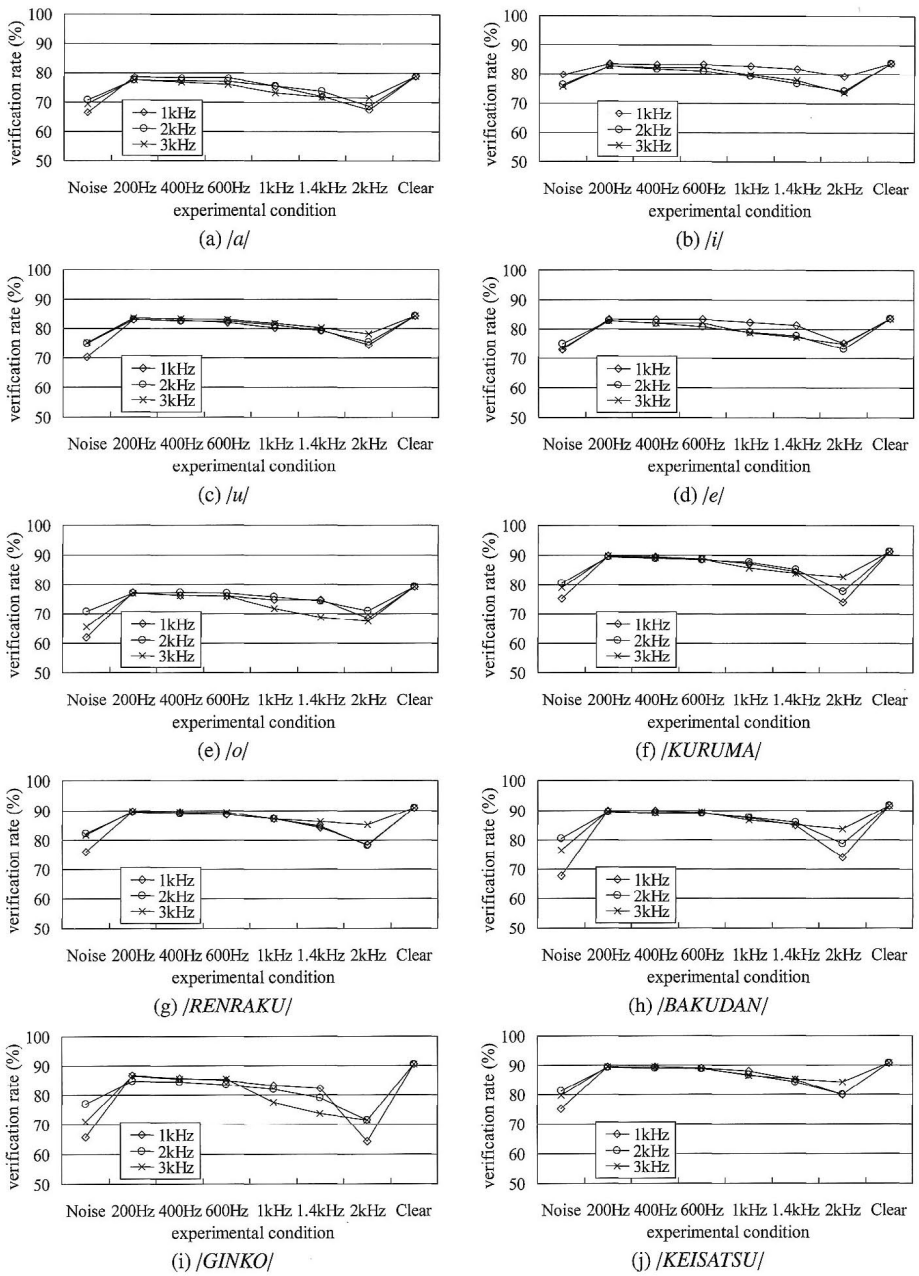

Fig. 3 Results of verification experiment with MFM processing of contrast data.

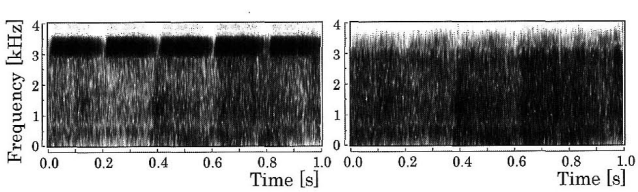

Fig. 4 Sound spectrograph of crime prevention alarm (left) and alarm with BEF (right).

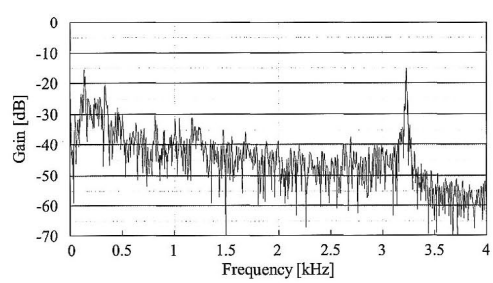

Fig. 5 Spectrum of crime prevention alarm.

of speech and the verification accuracy in a realistic environment. We prepared the speech data in a noisy environment with SNR set to 10, 0, and −10 dB. A sample sound spectrograph of the speech with power supply noise is shown in Fig. 7.

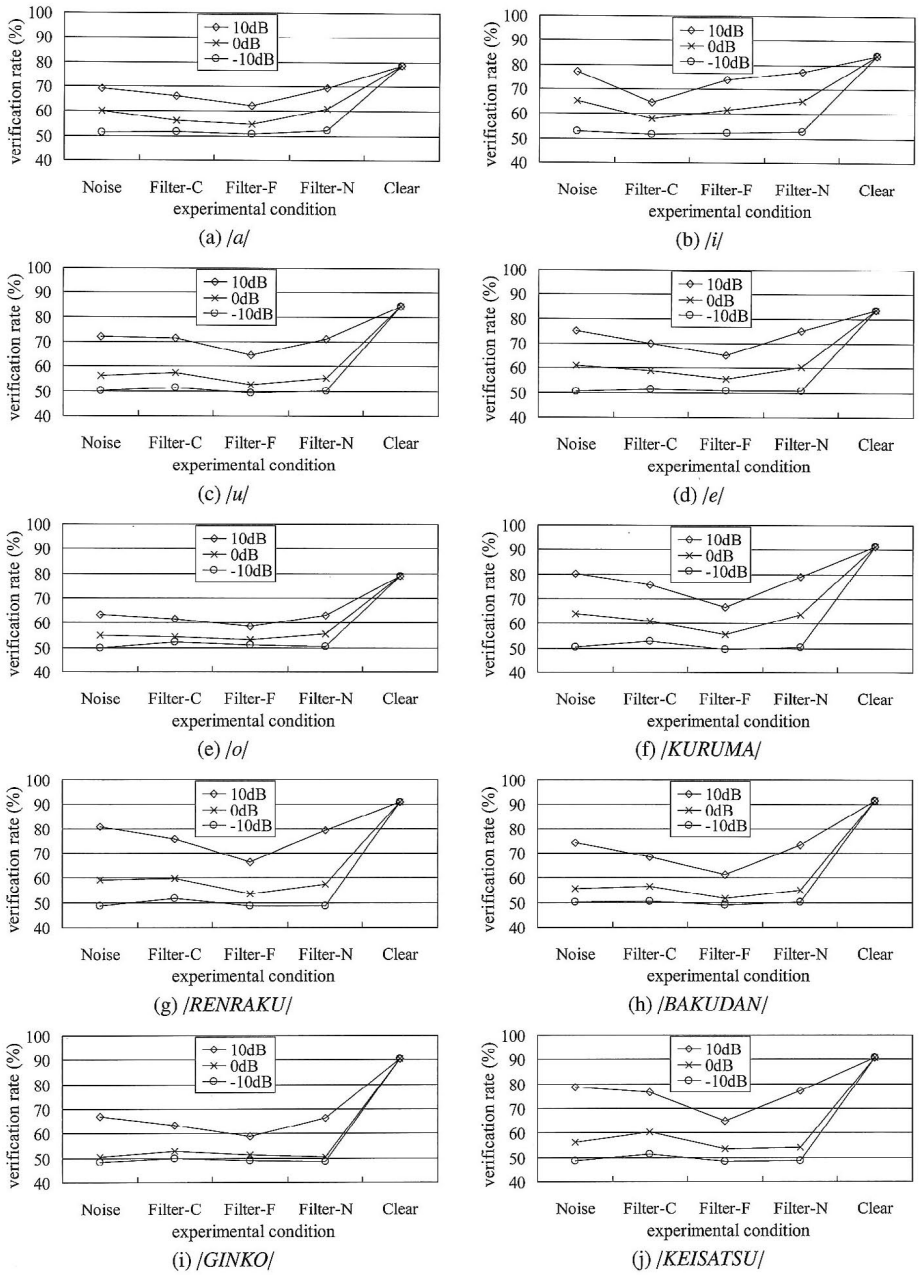

Fig. 6 Results of verification experiment for realistic noisy environment: crime prevention alarm.

Table 4 Verification experiment conditions for realistic environment: crime prevention alarm.

condition	unknown speaker's data	contrast data
Noise	no processing	with added noise
Filter-C	BEF processing	no processing
Filter-F	BEF processing	BEF processing
Filter-N	BEF processing	with added noise & BEF processing
Clear	SNR = ∞ (target value)	SNR = ∞ (target value)

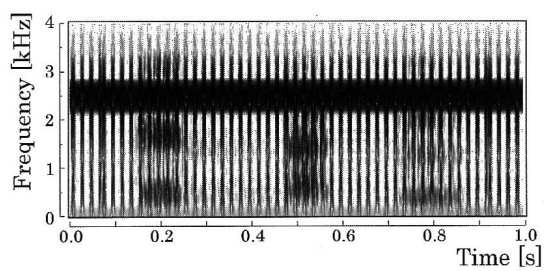

Fig. 7 Sound spectrograph of speech (*/RENRAKU/*) with power supply noise.

5.2 Comb Filter

Power supply noise is a narrow-band noise with a 50 Hz harmonic component because it is not a pure sound like the sine wave noise, although power supply noise has the frequency of 50 Hz. The frequency comb filter, not the band elimination filter, is effective for the clarification of speech with such a noise characteristic. The effect of speech clarification improves with the use of a frequency filter that eliminates the fundamental frequency and harmonic components. We can design the comb filter comparatively easily. When the period τ for the signal $x(t)$ at time t is given, the comb filter that reduces the noise with the fundamental and harmonic components can be described as

$$y(t) = x(t) - x(t - \tau) \tag{4}$$

and frequency response of the comb filter is shown in Fig. 8. Using the comb filter, we applied the clarification of speech to an unknown speaker's data in a realistic noisy environment with power supply noise.

5.3 Experimental Results of Speaker Verification in Realistic Noisy Environment: Power Supply Noise

The speaker verification experiments with the clarification of speech were performed under conditions shown in Table 5. Figure 9 shows the experimental results. The experimental results show that the use of the comb filter for the clarification of speech is often effective in improving the verification accuracy when the same comb filter (missing feature mask) is applied to the contrast data as the comb filter applied to the unknown speaker's data. The decrease in the verification accuracy was suppressed from about 5 to about 10% per word, and it was shown that speaker verification with the clarification of speech including narrow-band noise was very effective for the realistic environment because the verification rate per word exceeded roughly 80% when SNR was −10 dB.

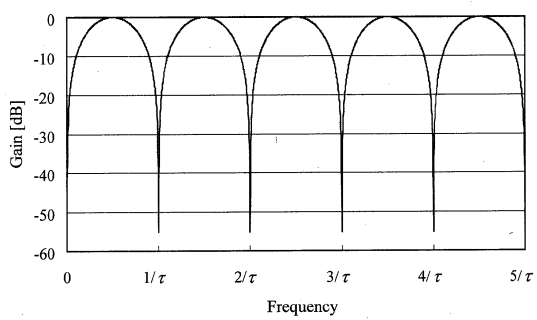

Fig. 8 Freqency response of comb filter.

Table 5 Verification experiment conditions for realistic environment: power supply noise.

condition	unknown speaker's data	contrast data
Noise	no processing	with added noise
Filter-C	comb filter processing	no processing (no MFM processing)
Filter-F	comb filter processing	comb filter processing (MFM processing)
Clear	SNR = ∞ (target value)	SNR = ∞ (target value)

6. Conclusions

As a result of forensic speaker verification experiments in noisy environments and experiments on the relationship between the clarification of speech and the verification accuracy, we have found that the effect of the clarification of speech with narrow-band noise on the verification accuracy is significant, and the verification accuracy is greatly ameliorated by clarifying the speech.

The improvement of the verification accuracy upon the clarification of speech with narrow-band noise was also confirmed for a realistic environment by the experiment in which power supply noise was used. Moreover, we found that the clarification of speech for the improvement of the verification accuracy is useful for the situation of a highly noisy environment of SNR of 0 dB to −10 dB.

However, it is difficult to obtain high verification accuracy by the clarification of speech when wide-band noise is included, and it is necessary to examine a speaker verification method for the case that wide-band noise is included. When a noise component remains after noise reduction, the method for adding the same kind of noise to the contrast data is effective for speaker verification.

In the future, we will discuss the application of the results of this study to forensic work. We will also further investigate effective methods of the clarification of speech for speaker verification in noisy environments, and the features and verification methods effective in improving the verification accuracy.

Acknowledgements

The authors would like to express sincere thanks to Mr. Masahiro Suwa, Criminal Investigation Laboratory, Saitama Prefectural Police, H.Q., who provided some realistic noisy environment data.

References

[1] H. Sakoe and S. Chiba, "Recognition of continuously spoken words based on time-normalization by dynamic programming," J. Acoust. Soc. Jpn. (E), vol.27, no.9, pp.483–490, Sept. 1971.
[2] S. So and K.K. Paliwal, "Multi-frame GMM-based block quantization of line spectral frequencies for wideband speech coding," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Process., vol.1, pp.121–124, Philadelphia, USA, March 2005.
[3] M. Nishida and T. Kawahara, "Speaker model selection based on the bayesian information criterion applied to unsupervised speaker indexing," IEEE Trans. Speech Audio Process., vol.13, no.4, pp.583–592, July 2005.
[4] S. Kuroiwa, Y. Umeda, S. Tsuge, and F. Ren, "Nonparametric speaker recognition method using earth mover's distance," IEICE

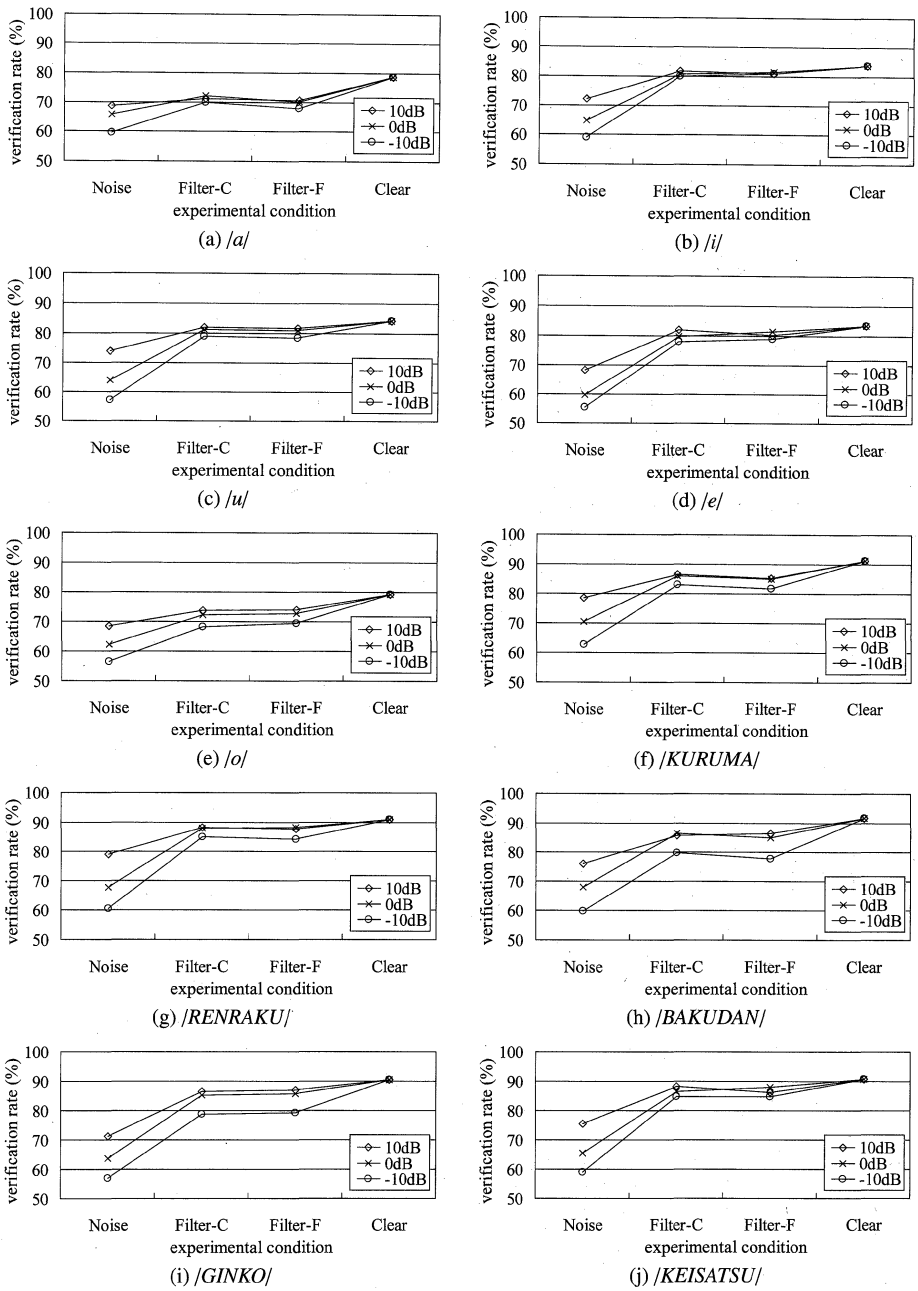

Fig. 9 Results of verification experiment for realistic noisy environment: power supply noise.

Trans. Inf. & Syst., vol.E89-D, no.3, pp.1074–1081, March 2006.

[5] H. Noda and T. Osanai, "On the relation between the number of speakers and the reliability of recognition rate in speaker recognition," IEICE Trans. Fundamentals (Japanese Edition), vol.J73-A, no.4, pp.717–724, April 1990.

[6] H. Makinae, T. Osanai, T. Kamada, and M. Tanimoto, "Investigating feature property for speaker recognition using the MFCC with consideration of frequency band," Proc. 2006 Autumn Meeting of the Acoustical Society of Japan, pp.75–76, Sept. 2006.

[7] T. Osanai, K. Ozeki, T. Kamada, H. Makinae, and M. Tanimoto, "Effects of feature parameter transformation in text-independent speaker verification based on cross vector quantization distortion," Proc. 2006 Autumn Meeting of the Acoustical Society of Japan, pp.57–58, Sept. 2006.

[8] http://sox.sourceforge.net/

[9] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Process., vol.1, pp.121–124, Seattle, USA, May 1998.

[10] S. Kuroiwa, S. Tsuge, and F. Ren, "Lost speech reconstruction method using speech recognition based on missing feature theory and HMM-based speech synthesis," Proc. INTERSPEECH 2006, pp.1105–1108, Pittsburgh, USA, Sept. 2006.

[11] T. Kamada, T. Osanai, H. Makinae, and M. Tanimoto, "Speaker verification in very bad noise environment," Proc. 2005 Autumn Meeting of the Acoustical Society of Japan, pp.129–130, Sept. 2005.

[12] T. Kamada, N. Minematsu, T. Osanai, H. Makinae, and M. Tanimoto, "Speaker verification in noisy environment," IEICE Technical Report, SP2006-178, March 2007.

Toshiaki Kamada received the B.E. degree in electrical engineering from the University of Tokyo, Tokyo, Japan, in 1995. He has been with the National Research Institute of Police Science, Chiba, Japan, since 1995. His research interests include speaker recognition and speech processing in forensic field. He is a member of ASJ and the Japanese Association of Forensic Science and Technology (JAFST).

Nobuaki Minematsu received the Ph.D. degree in electronic engineering in 1995 from the University of Tokyo. In 1995, he was an assistant researcher at Department of Information and Computer Science of Toyohashi University of Technology and in 2000, he was an associate professor at Graduate School of Engineering of the University of Tokyo. Since 2004, he has been an associate professor at Graduate School of Frontier Sciences of the University of Tokyo. From 2002 to 2003, he was a visiting researcher at Kungl Tekniska Högskolan in Sweden. He has wide interest in speech from science to engineering, including phonetics, phonology, speech perception, speech analysis, speech recognition, speech synthesis, and speech application. He is a member of ASJ, Information Processing Society of Japan, Japanese Society for Artificial Intelligence, Phonetic Society of Japan, International Speech Communication Association, the International Phonetic Association, the Computer Assisted Language Instruction Consortium, and the European Association for Computer Assisted Language Learning.

Takashi Osanai received the B.E., M.E. and Ph.D. degrees in electro-communications from University of Electro-Communications, Tokyo, Japan, in 1985, 1987 and 2007, respectively. Since 1987, he has been with the National Research Institute of Police Science, Chiba, Japan, where he is currently a chief of Third Information Science Section. In 2001, he has been a visiting fellow at the Phonetics Laboratory in the Department of Linguistics at the Australian National University, Canberra, Australia. His current research interests include speaker recognition in forensic field. He is a member of ASJ and JAFST.

Hisanori Makinae received the B.E. and M.E. degrees in Electrical and Communication Engineering from Tohoku University, Miyagi, Japan, in 1998 and 2000, respectively. Since 2005, he has been with the National Research Institute of Police Science, Chiba, Japan, where he is currently a researcher of Third Information Science Section. His research interests include speaker recognition in forensic field. He is a member of ASJ and JAFST.

Masumi Tanimoto received the B.E. degree from Shibaura Institute of Technology in 1970. He has been with the National Research Institute of Police Science, Chiba, Japan, since 1970 and is currently a director of Fourth Forensic Science Department since 2004. His major research interests include speech processing and signal processing. He is a member of ASJ and JAFST.