Speaker Verification with Non-Audible Murmur Segments

Mariko Kojima†, *Tomoko Matsui*‡, *Hiromichi Kawanami*†, *Hiroshi Saruwatari*†, *Kiyohiro Shikano*†

†Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan
‡The Institute of Statistical Mathematics, Tokyo, Japan
email: mariko-k@is.naist.jp

Abstract

We propose a speaker verification method using non-audible murmur (NAM) segments, which are different from normal speech and hard for other people to catch them. To use NAM, we therefore take a text-dependent verification strategy in which each user utters her/his own keyword phrase and utilize not only speaker-specific but also keyword-specific acoustic information. We expect this strategy to yield a relatively high performance. NAM segments, which consist of multiple short-term feature vectors, are used as input vectors to capture keyword-specific acoustic information well. To handle segments with a large number of dimensions, we use the support vector machine (SVM). In experiments using NAM data of 19 male and 10 female speakers recorded in three different sessions, we achieved equal error rates of 0.04% (male) and 1.1% (female) when using 145-ms-long NAM segments. These rates are half or less those obtained with 25-ms-long input vectors.

Index Terms: speaker verification, non-audible murmur, segments.

1. Introduction

Biometric authentication is increasingly being used in security control. It should be extremely difficult to impersonate biometrical information such as fingerprints, irises, and voices. Moreover, such information cannot be lost by the user. In particular, voice authentication (or speaker verification) generates less psychological resistance in users than the other authentication methods and can easily be performed over cellular phones. Many studies of this technology have been reported[1][2], and demand for it is expected to increase. However, its performance is still insufficient and is usually much lower than that of fingerprint and iris authentications. Another problem with voice authentication is that even though a text-dependent approach using a keyword phrase for each user is expected to provide high performance, this approach is not practical because of the opportunities for attacks involving interception and playback of live utterances.

NAM offers a new style of speech input[3]. It is hard for other people to catch NAM and it is recorded using a special microphone placed on the surface of the body. NAM data actually includes murmurs, some body vibrations, and a smaller number of external noises. Using NAM instead of normal speech lets us safely take the text-dependent approach using keyword phrases, and it should provide effective and noise robust authentication.

To date, we have investigated speaker verification with NAM using keyword utterances recorded in one session[4]. The performances of Gaussian-mixture-model-based (GMM-based) and SVM-based methods were compared using a short-term (25-ms) feature vector as an input vector, and we found that the SVM-based method performed as well or better than the GMM-based one. However, keyword-specific acoustic information was not specially utilized in the previous work. In speech recognition with neural networks (NNs), Waibel et al.[5] reported that a concatenation of several short-term feature vectors captured the text-specific acoustic information more efficiently than individual short-term feature vectors and it was possible to improve the performance by using the concatenation as input data to NNs.

In this paper, we propose speaker verification using NAM segments, which consist of several short-term feature vectors, as input data so as to make good use of keyword-specific acoustic features. Since the segments are represented as vectors with a high number of dimensions, we introduce SVM[8] to deal with it. In SVM, the kernel function is used to alleviate the curse-of-dimensionality problem. We evaluated our method using NAM data recorded in three different sessions to study the robustness against session-to-session variations in NAM data.

2. Speaker verification using NAM

In this section, we introduce the NAM data and explain our SVM-based method.

2.1. NAM

NAM is produced in a voiceless utterance action and is uttered when one grumbles to oneself not intending to be heard by others, says prayers, or makes silent wishes. One only moves the speech organ while breathing, without vocal cord vibration or glottis narrowing. Figure 1 compares spectrograms and waveforms of normal speech and NAM. The utterance contents were the same in both cases: "mada seishiki ni kimatta wake deha nainode (in Japanese)." We can see that NAM includes the main information under 4 kHz while the information in higher frequency bands cannot be observed.

Breath-induced vibration of the vocal tract is transmitted as NAM through the body directly to a capacitor microphone worn on the surface of the skin below the mastoid bone. Figure 2 shows the NAM microphone and its attachment method.

2.2. SVM-based method

Figure 3 shows the procedure of our method. In training, an SVM is trained for each speaker. The SVM is a binary classifier and the training data for each speaker is divided into two sets for positive (+1) and negative (-1) classes. The +1 class data consists of keyword utterances of a registered speaker and the -1 class data consists of non-keyword utterances of the speaker and utterances of other speakers. An input vector is a concatenation of n short-

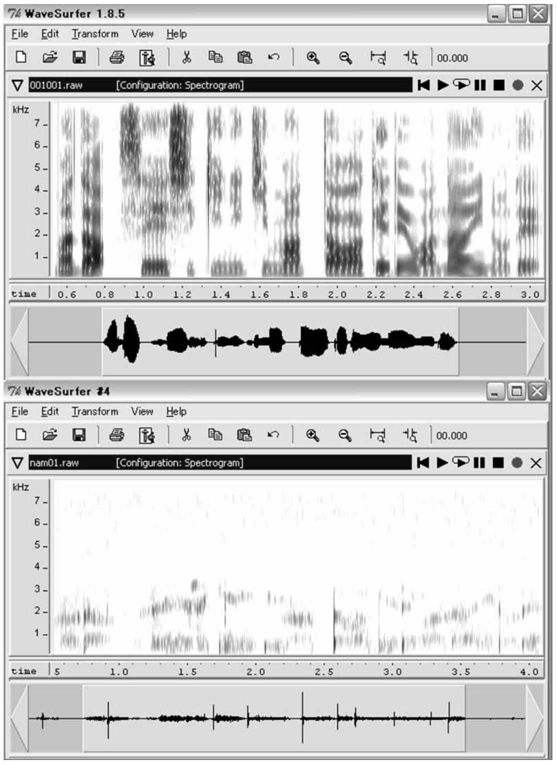
WaveSurfer 1.8.5
File Edit Transform View Help
00.000
001001.raw [Configuration: Spectrogram]
kHz
7
6
5
4
3
2
1
time 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0
WaveSurfer #4
File Edit Transform View Help
00.000
nam01.raw [Configuration: Spectrogram]
kHz
7
6
5
4
3
2
1
time 5 1.0 1.5 2.0 2.5 3.0 3.5 4.0

Figure 1: Comparison of spectrograms and waveforms of normal speech (top) and NAM (bottom).

term feature vectors extracted from the training utterances. The concatenation is assumed to represent keyword-specific acoustic features well. In testing as in training, concatenations of n short-term feature vectors are made for each utterance and used as input vectors. SVM gives a confidence index for each input vector, which is called the 'margin'. The confidence index averaged over the test utterance is compared with a threshold to judge speaker identity.

3. Experiments

We conducted speaker verification experiments using keyword utterances and confirmed the effectiveness of NAM segment input. We also evaluated the robustness against session-to-session variations by comparing the performances of our SVM-based method with the conventional GMM-based method.

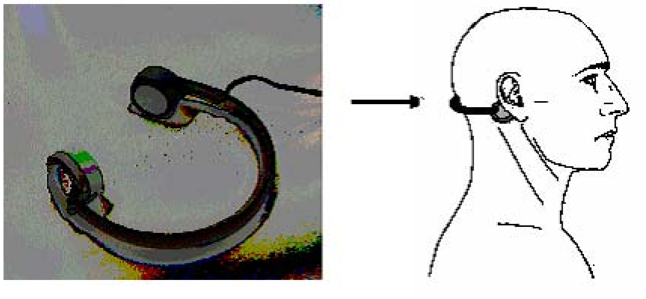

Figure 2: NAM microphone and its attachment method.

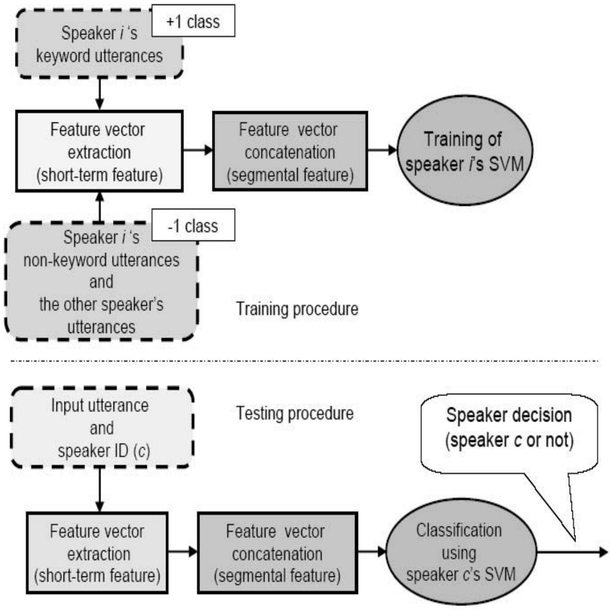
+1 class
Speaker i 's keyword utterances
Feature vector extraction (short-term feature)
Feature vector concatenation (segmental feature)
Training of speaker i's SVM
-1 class
Speaker i 's non-keyword utterances and the other speaker's utterances
Training procedure
Input utterance and speaker ID (c)
Testing procedure
Speaker decision (speaker c or not)
Feature vector extraction (short-term feature)
Feature vector concatenation (segmental feature)
Classification using speaker c's SVM

Figure 3: Training and testing procedures of the SVM-based method.

3.1. Data description and experimental conditions

The NAM data was collected for 19 male and 10 female speakers who uttered keyword phrases. Each keyword phrase was a concatenation of two place names of Japanese prefectures (e.g., "Tokyo-Saitama" and " Kyoto-Nara"). The utterances were recorded at the sampling rate of 16 kHz in three sessions over a six-month period. The interval between sessions was three months. In each session, each speaker uttered her/his own keyword 16 times and uttered 30 keywords of other speakers twice. A Mel frequency cepstral coefficient (MFCC) vector of 31 components, consisting of 15 MFCCs plus their first derivatives and the first derivative of the normalized log energy, was derived once every 10 ms over a 25-ms Hamming-windowed speech segment. Cepstrum mean normalization was applied.

The training data set of each speaker was composed of the data uttered by speakers of the same gender in two sessions. The data for each session consisted of 10 keywords for the +1 class and 30 non-keyword utterances of the speaker and utterances of the other speakers for the -1 class (in detail, 190 utterances of the other male speakers when the speaker was male and 90 utterances of the other female speakers when the speaker was female).

In testing, we used keyword utterances uttered in a session that was different from the training sessions. The test data set for each speaker consisted of 6 keyword utterances of the speaker and non-keyword utterances of the other speakers (in detail, 114 non-keyword utterances of the other male speakers when the speaker was male and 54 non-keyword utterances of the other female speakers when the speaker was female). The former utterances should be accepted and the latter ones should be rejected as false statements. The threshold for speaker decision was speaker dependent and set a posteriori. Thus, all evaluations were gender-dependent.

For SVM, we used SVM^{light} which is a toolkit provided by Cornell University [7]. The polynomial kernel function (1) was used.

$$k(x, y) = (x^t y + 1)^s \quad (1)$$

The power was chosen to be $s = 7$. To enable us to perform effective computation with 64-bit precision, the data was so scaled that all the elements of feature vectors lay in the interval [-0.5, 0.5].

In the GMM-based method, we used HTK(ver3.2), which is a toolkit provided by Cambridge University[8]. A GMM was made for each speaker using the keyword utterances of the speaker. Then, we made a universal-background GMM using all keyword utterances of all speakers to normalize the likelihood values of the speaker GMM[9]. Both GMMs were 32-Gaussian-mixture diagonal covariance models, because those models showed the best performance in preliminary experiments using 32- and 64-Gaussian-mixture models. The collective log-likelihood ratio between the speaker and the universal-background GMMs was compared with a threshold to judge speaker identity.

3.2. Use of NAM segments

Figures 4 and 5 show the detection error tradeoff (DET) curves piled and averaged over male and female speakers with NAM segments with lengths of 25 ms (one short-term feature vector), 45 ms (three vectors), 85 ms (7 vectors), and 145 ms (13 vectors).

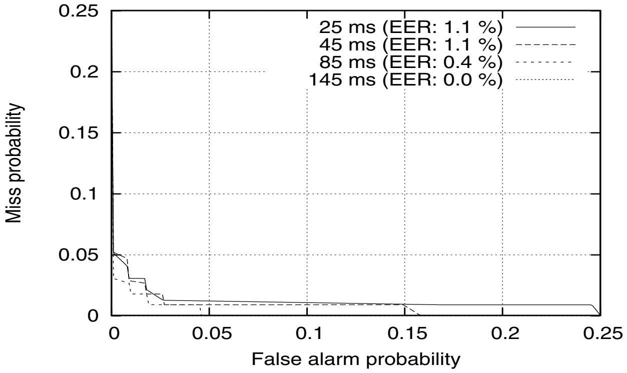

Figure 4: DET curves piled and averaged over male speakers.

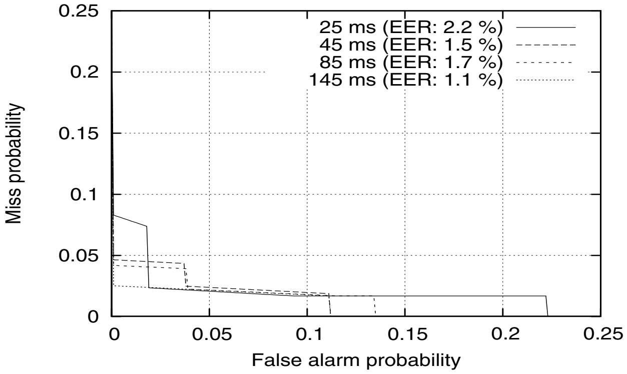

Figure 5: DET curves piled and averaged over female speakers.

As the segment length became longer, the verification performance became better. The best result was obtained with 145-ms-long segments, for which the mean values of the equal error rates (EERs) were 0.04% for male speakers and 1.1% for female speakers. These values are roughly 1/20 and 1/2 of those for male and female speakers, respectively, with 25-ms-long segments. These results indicate that 145-ms-long segments can efficiently represent both speaker- and keyword-specific acoustic features and that the performance is higher for text-dependent verification using keywords.

3.3. Training over multiple sessions

Table 1 compares the EERs for the SVM- and GMM-based methods when training data recorded in two sessions and one session, respectively, was used. As input vectors, 25-ms-long short-term feature vectors were used.

Table 1: EERs (%) for SVM- and GMM-based methods with training data recorded in one and two sessions (M: EER for male speakers, F: EER for female speakers)

Method	1 session	2 sessions
SVM	6.9 (M), 8.9 (F)	1.1 (M), 2.2 (F)
GMM	8.8 (M), 18.0 (F)	3.9 (M), 5.2 (F)

For the SVM-based method, the EERs were greatly reduced by using training data recorded in two sessions. When training data recorded in one session was used, the SVM-based method performed as well as or slightly better than the GMM-based method. This result indicates that for the SVM-based method using NAM, session-independent speaker-specific acoustic features can be effectively captured by using training data recorded in multiple sessions.

4. Discussion

4.1. Effect of the first derivatives of MFCCs

In the experiments, NAM segments were created by concatenating several short-term feature vectors consisting of 15 MFCCs plus their first derivatives and the first derivative of the normalized log energy. However, one may consider that the segments included information about the first derivatives of MFCCs (ΔMFCCs) computed in terms of five successive MFCC vectors. Therefore, we conducted experiments using a NAM segment of a concatenation of several feature vectors consisting of only 15 MFCCs and examined the effect of ΔMFCCs.

Table 2 lists the EERs with and without ΔMFCCs for NAM segments of various lengths. With ΔMFCCs, the numbers of dimensions of NAM segment vectors with lengths of 25, 45, 85, and 145 ms were 31, 93, 217, and 403 (=31 dim. x13 vectors), respectively. Without ΔMFCCs, the numbers of dimensions of NAM segment vectors with lengths of 45, 85, and 145 ms were 45, 105, and 195 (=15 dim. x13 vectors), respectively. Although the EERs with ΔMFCCs were slightly better than those without ΔMFCCs, the difference in the EERs was rather small. On the other hand, when ΔMFCCs are used, the cost of SVM calculation is almost double. In limited computational environments, it may not be necessary to involve ΔMFCCs in NAM segments.

4.2. Effect of keywords

In speaker verification using NAM data, keyword utterances can be utilized safely because NAM is inaudible,. Therefore, our method was evaluated using non-keyword utterances uttered by

Table 2: EERs (%) with and without ΔMFCCs (M: EER for male speakers, F: EER for female speakers)

Segment length	with ΔMFCCs	without ΔMFCCs
25 ms	1.1 (M), 2.2 (F)	——
45 ms	1.1 (M), 1.5 (F)	1.5 (M), 2.9 (F)
85 ms	0.4 (M), 1.7 (F)	0.7 (M), 1.1 (F)
145 ms	0.0 (M), 1.1 (F)	0.0 (M), 1.1 (F)

other speakers as false utterances ("basic case"). To examine the effect of using keywords, we conducted experiments using the following sets of false utterances.

basic case:	non-keyword utterances uttered by other speakers
case A:	keyword utterances uttered by other speakers [impersonation]
case B:	non-keyword utterances uttered by the (claimed) speaker [incorrect keywords]

Figure 6 shows the EERs in the basic case and in cases A and B for male and female speakers. The EERs in the basic case were much smaller than those in cases A and B for various NAM segment lengths. As the length became longer, the EERs in case A increased while those in case B decreased. These results confirm the effectiveness of the strategy in which keywords are available. While longer segments are effective for distinguishing different keyword utterances, the discriminative power on the same keyword utterances by different speakers decreases. If case A is of prime importance, segments of medium length from 45 to 85 ms should be selected as input vectors. If case B is of prime importance, longer (e.g., 145-ms) segments are more suitable.

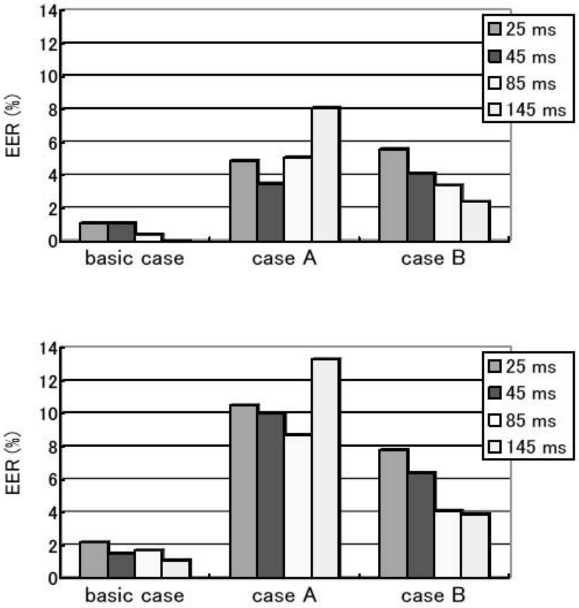

Figure 6: Male (top) and female (bottom) speakers' EERs (%) in several cases for false utterances: basic case with non-keyword utterances of other speakers, case A with keyword utterances of other speakers, and case B with non-keyword utterances of the speaker.

5. Conclusions

We investigated speaker verification using NAM segments, which enable us to take a text-dependent verification strategy using a keyword phrase. The proposed method with long (145-ms) segments was found to be effective, especially when using training data uttered in two sessions, and reduced the EERs to half or less those obtained with the method using short (25-ms) segments. We discussed the effect of ΔMFCCs and showed that long segments can represent almost the same information in ΔMFCCs and that NAM segments can be constructed with only MFCCs in environments where computational resources are limited. Moreover, the effectiveness of keywords was demonstrated for several cases of false utterances. If impersonation is of prime importance, medium-length segments (from 45 to 85 ms) should be selected as input vectors. If incorrect keywords are of prime importance, longer segments (e.g., 145 ms) are more suitable.

We plan to conduct experiments using data recorded in a larger numbers of sessions and evaluate our method in practice by using data uttered by impostors who were not used to train the speaker SVM. Since NAM data includes internal body sounds such as heart beats, which contain person-specific information, we will investigate new effective features of the internal sounds.

6. References

[1] http://www.nist.gov/speech/tests/spk/index.htm, NIST Speaker Recognition Evaluations.

[2] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology," In Proc. International Conference on Acoustics, Speech, and Signal Processing in Orlando, FL, IEEE, pp. IV: 4072-4075, 2002.

[3] Y. Nakajima, H. Kashioka, N. Campbell, K. Shikano, "Non-Audible Murmur (NAM) Recognition," IEICE Trans. Information and Systems, Vol. E89-D, No. 1, pp. 1-8, 2006.

[4] M. Kojima, H. Kawanami, H. Saruwatari, T. Matsui and K. Shikano, "Speaker Verification using Non-Audible Murmur," In Proc. Symposium on Cryptography and Information Security (SCIS1906), 2D1-4, 2006 (in Japanese).

[5] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang, "Phoneme Recognition Using Time-Delay Neural Networks," IEEE Trans. ASSP, Vol. 37, No. 03, pp. 328-339, 1989.

[6] V. N. Vapnik, "The Nature of Statistical Learning Theory," Springer, 1995.

[7] Thorsten Joachims: SVM^{light} Support Vector Machine, Version 6.01, http://www.cs.cornell.edu/People/tj/svm_light/index.html, Cornell University, Department of Computer Science, 2004.

[8] http://htk.eng.cam.ac.uk/, The Hidden Markov Model Toolkit (HTK).

[9] T. Matsui and S. Furui, "A similarity normalization method for speaker verification based on a posteriori probability," ESCA Tutorial and Research Workshop on Automatic Speaker Recognition Identification Verification, pp. 59-62, 1994.