



# Study on Speaker Verification with Non-Audible Murmur Segments

Hideki Okamoto<sup>1</sup>, Mariko Kojima<sup>1</sup>, Tomoko Matsui<sup>2</sup>,  
Hiromichi Kawanami<sup>1</sup>, Hiroshi Saruwatari<sup>1</sup>, and Kiyohiro Shikano<sup>1</sup>

<sup>1</sup>Graduate School of Information Science, Nara Institute of Science and Technology  
8916-5 Takayama-cho Ikoma-shi Nara 630-0192, Japan

<sup>2</sup>The Institute of Statistical Mathematics, 4-6-7 Minami-azabu Minato-ku Tokyo 106-8569  
hideki-o@is.naist.jp

## Abstract

We investigated a speaker verification method that uses non-audible murmur (NAM) segments using newly collected data and obtained several findings that will be useful when speaker verification systems are made in practice. NAM is recorded using a special microphone placed on the surface of the body, so it includes almost no external noise and is hard for other people to hear. By utilizing these properties, we have already reported a text-dependent method using NAM segments that can use a keyword phrase safely. This paper extends the examination with newly collected data consisting of NAM uttered by 18 male and 9 female imposter speakers and by 18 male and 10 female customer speakers. Experiments with various numbers of training utterances and sessions show that it is effective to use data recorded in multiple sessions. We also investigated the minimum number of training utterances needed in our method.

**Index Terms:** Speaker recognition, speaker verification, non-audible murmur, support vector machine

## 1. Introduction

Biometric authentication covers various technologies that measure and analyze human physical and behavioral characteristics for authentication purposes. Examples of physical characteristics include fingerprints, eye retinas and irises, facial patterns, and hand measurements, while examples of mainly behavioral characteristics include signatures, gait, and typing patterns. Biometric authentication has become widely used recently because it is difficult for an imposter to impersonate another person and biometric data cannot be forgotten[1]. However, if a person's biometric data is stolen or duplicated, an imposter could easily use it to access the protected system. Thus, the system usually invalidates compromised data, which makes it hard for the genuine person to continue using the system. This is a serious problem. Human physical and behavioral characteristics are difficult to change in practice, so it is hard for an imposter to supply such biometric data to the system.

In biometric authentication using voice with a keyword phrase[2], the problem of data theft can be addressed by changing the phrase. Voice authentication services can be less of a mental burden to users because utterances are familiar every day actions. Moreover, the services do not need special equipment except for a microphone and can be deployed on mobile networks. However, in voice authentication, there is the problem 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.

To solve this problem, in[3], we proposed a method using non-audible murmur (NAM) segments, which is new style of speech input[4]. NAM is hard for other people to catch 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 few 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. As input data, we used NAM segments, which consist of several short-term feature vectors, so as to make good use of keyword-specific acoustic features. Since the segments were represented by vectors with a high number of dimensions, we introduced a support vector machine (SVM)[5] in which the curse-of-dimensionality problem is alleviated by utilizing the kernel function. The performance was evaluated in experiments using NAM data uttered by 18 male and 10 female speakers. Verification by gender was performed using one speaker as the customer and the other speakers as imposters, rotating through all speakers and then averaging the results. However, in practice, imposters will not always be one of the customers. Therefore, we collected new data uttered by 18 male and 9 female speakers as imposter data in addition to new data uttered by the customer speakers in another session. In this paper, we report a full examination of the method using the new data.

Moreover, in [3], we compared the performance with training data recorded in one and two sessions to study the robustness against session-to-session variations in NAM data. While 10 training utterances were used for each speaker in the case of training data recorded in one session, 20 training utterances were used in the case of training data recorded in two sessions. Although we showed that the performance with training data recorded in two sessions was higher than that in one session, we did not identify the cause of the difference. Two possible factors can be considered: one depends on the number of training utterances and the other on the number of sessions. We conducted additional experiments with a fixed number of training utterances and with various numbers of sessions to investigate the effects. In practice, it is important to reduce the burden on the customer in the registration process. We also found the minimum required number of training utterances through experiments with several different numbers of training utterances.

## 2. Speaker verification using NAM

### 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. In NAM, the main information is below 4 kHz and information in higher frequency bands is not observed. Breath-induced vibration of the vocal tract is transmitted as NAM through the body directly to a condenser micro-

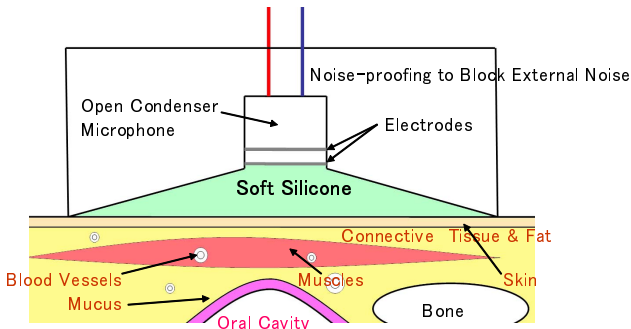
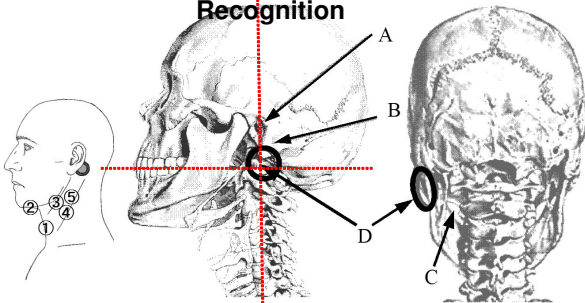


Figure 1: Cutaway of NAM microphone.

### The Best Sensing Position for NAM Recognition



A: Earhole B: Mastoid Process C: Opening of the Mouth  
D: The Best Sensing Position for NAM Recognition

Figure 2: Optimum attachment position of NAM microphone.

phone worn on the surface of the skin below the mastoid bone, as shown in Figures 1 and 2.

## 2.2. SVM-based method

The procedure of our method is shown in Figure 3. In training, an SVM is trained for each speaker. The SVM is a binary classifier[6] 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 customer 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-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 newly collected imposter data and customer data and investigated the effectiveness of NAM segment input. Moreover, in order to make clarify the effect of using data collected over multiple sessions and to investigate the amount of training data required, we examined the performance with various numbers of training utterances and sessions.

### 3.1. Data description and experimental conditions

The data is summarized in Table 1. We used keyword phrases uttered by 18 male and 10 female speakers in four sessions over

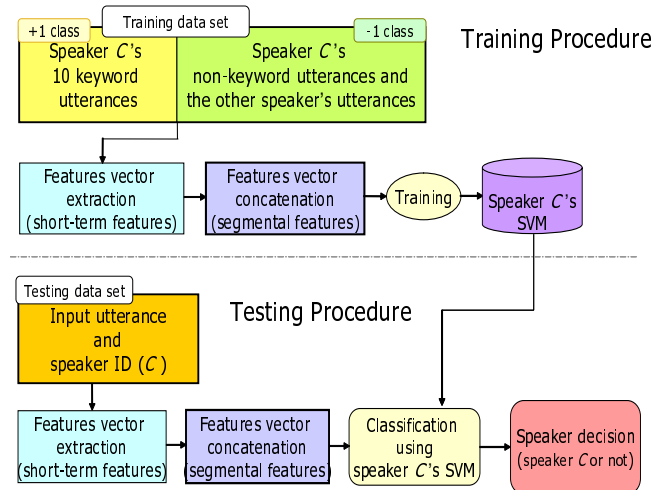


Figure 3: Training and testing procedures of the SVM-based method.

a 9-month period as customer data, while we used keyword phrases uttered by different 18 male and 9 female speakers in a different session as imposter data. The interval between sessions was more than three months. Each keyword phrase was a concatenation of two place names (Japanese prefectures, e.g., "Tokyo-Saitama" and "Kyoto-Nara"). In each session, each customer/imposter uttered her/his own keyword 16 times and uttered 29 keywords of other customers/imposters twice. An MFCC vector of 31 components, consisting of 15-dimensional MFCCs plus  $\Delta$ MFCCs and  $\Delta$ power, was derived for 10 ms over a 25-ms Hamming-windowed speech segment. NAM segments were created by concatenating several feature vectors consisting of only MFCCs because the segments can include information about the first derivatives of MFCCs ( $\Delta$ MFCCs) computed in terms of five successive MFCC vectors. Cepstrum mean normalization was applied.

The training data set of each customer speaker was composed of the data uttered by customer speakers of the same gender in three sessions. The data for each session consisted of 10 keyword utterances for the +1 class and 15 non-keyword utterances of the speaker (randomly selected from 30 keywords) and utterances of the other customer speakers for the -1 class (in detail, 170 utterances of the other male customer speakers when the speaker was male and 90 utterances of the other female customer speakers when the speaker was female).

In testing, we used keyword utterances uttered by each customer speaker in a session that was different from the training sessions and imposter utterances. The test data set basically consisted of 6 keyword utterances of the customer speaker and imposter utterances (in detail, 108 utterances of male imposter speakers when the customer speaker was male and 54 utterances of female imposter speakers when the customer speaker was female). We call this test data set the "basic case".

In the basic case, the method was evaluated using non-keyword utterances uttered by imposters as false utterances. In practical conditions, we sometimes need to assume keyword utterances uttered by other speakers (impersonation: "case A") or non-keyword utterances uttered by the customer speaker (incorrect keywords: "case B") as false utterances. In case A, we assigned the customer keyword utterances uttered by the other speakers to the -1 class, and in case B, we assigned the non-keyword utterances uttered by the customer speaker to the -1 class.

The threshold for speaker decision was speaker-dependent and set a posteriori. For experiments, we used  $SVM^{light}$ [6]

Table 1: Data description.

Customers	Numbers	18 males and 10 females
	Recording	4 sessions (Jun, Sep, Dec 2005, Feb 2006)
	Contents	30 keyword utterances (e.g., “Kyoto-Nara” and “Tokyo-Saitama”)
Imposters	Numbers	18 males and 9 females (different from customers)
	Recording	1 session (Oct 2006)
	Contents	30 keyword utterances (different from those of customers, e.g., “Otaru-Hakodate”)
Number of utterances per speaker		the person keyword utterance $\times$ 16 repetitions, 29 others’ keyword utterances $\times$ 2 repetitions
Analysis conditions		window-length: 25 ms, shift: 10 ms, sampling rate: 16 kHz
Parameters		15-dimensional MFCC, 15-dimensional $\Delta$ MFCC, $\Delta$ power

which is a toolkit provided by Cornell University. The polynomial kernel function (1) was

$$k(x, y) = (x^t y + 1)^7. \quad (1)$$

To enable us to perform effective computation with 64-bit precision, the data was scaled so that all the elements of features vectors lay in the interval  $[-0.5, 0.5]$ .

### 3.2. Results using imposter data

Table 2 lists the equal error rates (EERs) in the basic case averaged over male/female speakers with NAM segments with lengths of 25 ms (31-dimensional vector; MFCC+ $\Delta$ MFCC+ $\Delta$ power), 45 ms (45-dimensional vector; 3 MFCC vector concatenation), 85 ms (85-dimensional vector; 7 MFCC vector concatenation), and 145 ms (195-dimensional vector; 13 MFCC vector concatenation).

For male speakers, the best result was obtained with 145-ms-long segments and the EER was 0.6%. For female speakers, the best result was obtained with 25-ms-long segments and the EER was 0.2%. For male speakers, the EERs decreased when longer segments were used. For female speakers, although the EER for 145-ms-long segments was five times higher than that with 25-ms-long segments, even for 145-ms-long segments, the EER was around 1%. When averaged over male and female speakers, the EER for 85-ms-long segments was the lowest. This indicates that, in practice, we should use 85-ms-long segments for the basic case.

Table 2: EERs (%) for the basic case.

Segment length	25 ms	45 ms	85 ms	145 ms
Male	1.9	2.8	1.0	0.6
Female	0.2	0.4	0.6	1.1
Average	1.1	1.6	0.8	0.9

EERs by gender for cases A (impersonation) and B (incorrect keywords) are shown in Figure 4. For case A, the EERs for all segment lengths were much higher than those for the basic case. For male speakers, the best result for case A was also obtained with 85- and 145-ms-long segments, but the EER, 5.5%, was 10 times higher than that for the basic case. For female speakers, the best result for case A was obtained with 45- and 85-ms-long segments, but the EER, 8.3%, was 45 times higher than the lowest EER for 25-ms-long segments in the basic case. When averaged over male and female speakers, the EER for 85-ms-long segments was the lowest. This indicates that in practice, we should also use 85-ms-long segments for case A.

For case B, the EERs were rather low, especially for 145-ms-long segments. The EER for male speakers was 0.4% and that for female speakers was 1.0%. As the segment length increased, the EERs decreased. This indicates that longer (e.g., 145-ms) segments are more suitable for case B.

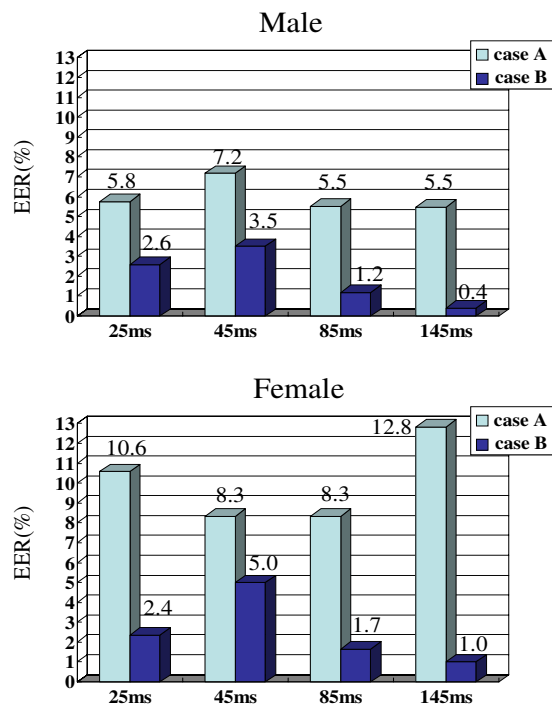


Figure 4: EERs for cases A and B.

### 3.3. Effect of numbers of training sessions

We previously showed that higher performance was obtained using training data recorded in multiple sessions[3]. In [3], for each speaker and each session, 10 keyword utterances were used for training. When we used training data recorded in  $N$  sessions, we used  $N \times 10$  keyword utterances. Therefore, there are two possible factors that might lead to the higher performance: one depends on the effect of more training utterances and the other depends on the effect of more sessions.

To identify the factor, we conducted experiments with a fixed total number of 10 keyword utterances recorded in various numbers of training sessions, as listed in Table 3. The “1-session train” means that 10 keyword utterances recorded in one training session in June 2005 were used, “2-session train” means that 5 keyword utterances recorded in June and September 2005 were randomly selected and used, and “3-session train” means that 4 keyword utterances recorded in June 2005 and 3 keyword utterances recorded in September and December 2005 were randomly selected and used. In the experiments discussed in this section, for each customer, we used the other customer speakers as imposters and used data recorded in February 2006 for testing. Here we used 85-ms-long segments.

Table 3: Numbers of keyword utterances for training.

Training session	1-session train (10,0,0)	2-session train (5,5,0)	3-session train (4,3,3)
Jun 2005	10	5	4
Sep 2005	0	5	3
Dec 2005	0	0	3
Total	10	10	10

Table 4 lists the EERs by gender for 1-, 2-, and 3-session training. As the number of sessions increased, the EERs decreased. This indicates that the higher number of sessions has a large effect and that our method can effectively capture session-independent speaker-specific acoustic features in NAM segments. Our future work includes investigating the number of sessions; that is, we need to clarify how many sessions are sufficient.

Table 4: Comparison of EERs (%) using various numbers of sessions.

Test cases	1-session train (10,0,0)		2-session train (5,5,0)		3-session train (4,3,3)	
	male	female	male	female	male	female
Basic case	9.2	12.0	2.3	11.1	<b>0.9</b>	<b>0.0</b>
Case A	18.7	25.1	10.2	23.1	<b>7.2</b>	<b>6.2</b>
Case B	9.8	6.9	3.9	5.7	<b>4.4</b>	<b>1.1</b>

### 3.4. Use of fewer training utterances

To find the minimum number of training utterances needed, we conducted experiments with fewer keyword utterances. Table 5 lists three cases of 3-session training with fewer keyword utterances. In “3-session-train (1,1,1)”, we reduced the number of keyword utterances from each session to one. In the experiments discussed in this section, for each customer, we used the other customer speakers as imposters and used data recorded in February 2006 for testing. The lengths of input segments were all 85-ms long.

Table 5: Fewer keyword utterances in 3-session training.

Training session	3-session train (3,3,3)	3-session train (2,2,2)	3-session train (1,1,1)
Jun 2005	3	2	1
Sep 2005	3	2	1
Dec 2005	3	2	1
Total	9	6	3

Figure 5 shows the ERRs by gender. EERs with two or more keyword utterances for each session were almost the same and relatively small. This indicates that, in practice, if we have two keyword utterances per customer and session, our method can achieve a relatively high performance.

## 4. Conclusions

We evaluated our speaker verification method using NAM segments with newly collected imposter and customer data and obtained several important findings that will be useful for making practical systems based on our method. First, we should use

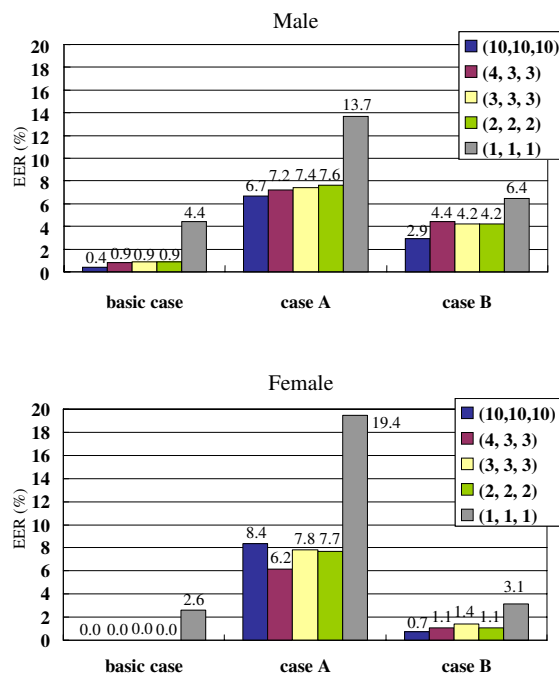


Figure 5: Comparison of EERs (%) with fewer keyword utterances.

85-ms-long segments for the basic case and for case A (impersonation); for case B (incorrect keywords), longer segments are suitable. Second, to obtain higher performance, it is important to use data recorded in multiple sessions. Third, only two keyword utterances per speaker and session are necessary, so our method can alleviate the burden on the customer in the registration process.

We plan to conduct experiments using a larger database and investigate *a priori* threshold settings for verification.

## 5. References

- [1] “Biometric Systems: Technology, Design and Performance Evaluation,” J. Wayman *et al.* (Eds.), Springer, 2004.
- [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] M. Kojima, T. Matsui, H. Kawanami, H. Saruwatari, and K. Shikano, “Speaker Verification with Non-Audible Murmur Segments,” the 9th International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP), pp. 2114-2117, Sept.2006
- [4] Y. Nakajima, H. Kashioka, N. Campbell, and K. Shikano, “Non-Audible Murmur (NAM) Recognition,” IEICE Trans. Information and Systems, Vol. E89-D, No. 1, pp. 1-8, 2006.
- [5] V. N. Vapnik, “The Nature of Statistical Learning Theory,” Springer, 1995.
- [6] H. Joachims: *SVM<sup>light</sup>* Support Vector Machine, Version 6.01, [http://www.cs.cornell.edu/People/tj/svm\\_light/index.html](http://www.cs.cornell.edu/People/tj/svm_light/index.html), Cornell University, Department of Computer Science, 2004.