interspeech 2008 incorporating SST 08

Speaker Verification with Non-Audible Murmur Segments by Combining Global Alignment Kernel and Penalized Logistic Regression Machine

Hideki Okamoto[1], *Tomoko Matsui*[2], *Hiromichi Kawanami*[1], *Hiroshi Saruwatari*[1], *and Kiyohiro Shikano*[1]

[1] Nara Institute of Science and Technology, Graduate School of Information Science
[2] The Institute of Statistical Mathematics
{hideki-o,kawanami,sawatari,shikano}@is.naist.jp, tmatsui@ism.ac.jp

10.21437/Interspeech.2008-398

Abstract

We investigate a novel method for speaker verification with non-audible murmur (NAM) segments. NAM is recorded using a special microphone placed on the neck and is hard for other people to hear. We have already reported a method based on a support vector machine (SVM) using NAM segments to use a keyword phrase effectively. To further exploit keyword-specific features, we introduce a global alignment (GA) kernel and penalized logistic regression machine (PLRM). In the experiments using NAM from 55 speakers, our method achieved an error reduction rate of roughly 60% compared with the SVM-based method using a polynomial kernel.

Index Terms: speaker verification, non-audible murmur, global alignment kernel, penalized logistic regression machine

1. Introduction

Biometric authentication has become widely used recently because it is difficult for an imposter to impersonate another person and biometric data cannot be forgotten. For biometric authentication using voice [1], the 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 [2,3], we proposed a method using non-audible murmur (NAM) segments, which consist of several short-term feature vectors, so as to make good use of keyword-specific acoustic features. NAM is hard for other people to catch and it is recorded using a special microphone placed on the surface of the neck skin below an ear [4]. NAM data actually includes murmurs and some body vibrations. In [4], a practical input interface for the recognition of NAM has been investigated and it is expected that several information services with NAM on mobile networks can be developed in future. Using NAM instead of normal speech lets us safely take the text-dependent approach using keyword phrases, and it should provide effective authentication on the input interface with NAM. Since the NAM segments are represented by vectors with a large number of dimensions, we utilized a support vector machine (SVM), in which the curse-of-dimensionality problem is alleviated by utilizing a kernel function. In experiments using NAM data uttered by 28 registered speakers and 27 imposter speakers, we obtained an equal error rate of 1.5% on average.

In this paper, we report introducing the global alignment (GA) kernel to better capture keyword-specific acoustic features. While standard kernels such as Gaussian and polynomial kernels are vector kernels, the GA kernel is a vector sequence kernel and constructed using similarities based on dynamic time warping (DTW) scores [5]. The GA kernel can effectively handle time series with variable lengths and local dependencies between neighboring states of the time series.

Moreover, we introduce a penalized logistic regression machine (PLRM) [6-7] instead of SVM to obtain higher verification performance. In speaker identification experiments [8], the PLRM-based method was compared with methods based on SVM and on a Gaussian mixture model (GMM) and found to be as good or better. Furthermore, PLRM provides a probabilistic estimate, while SVM gives a confidence index, which is termed the 'margin'. The probabilistic outputs of PLRM can be handled more easily than SVM outputs to set a prior threshold in practice.

2. GA kernel and PLRM-based method using NAM

In this section, we explain the components of our method—NAM, GA kernel, and PLRM—and then discuss the whole procedure.

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. NAM is recorded using a special microphone placed on the surface of the neck skin below an ear (below the mastoid bone) as shown in Fig. 1. It means that almost no external noise is included and it is hard for other people to hear. Breath-induced vibration of the vocal tract is transmitted as NAM through the body directly to a condenser microphone. Figure 2 illustrates the cross section of the NAM microphone and human body around the vocal tract.

In NAM, the main information is below 4 kHz and information in higher frequency bands is not observed as shown in Fig.3.

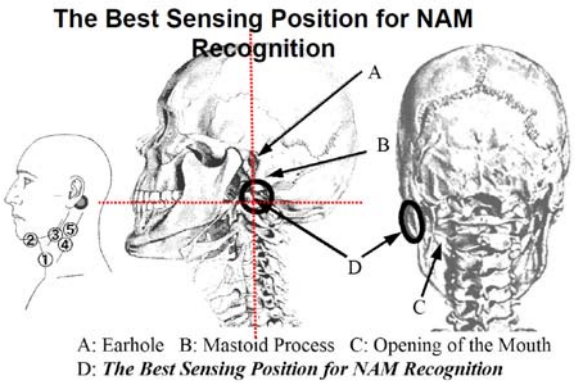

Fig. 1. The sensing position of a NAM microphone.

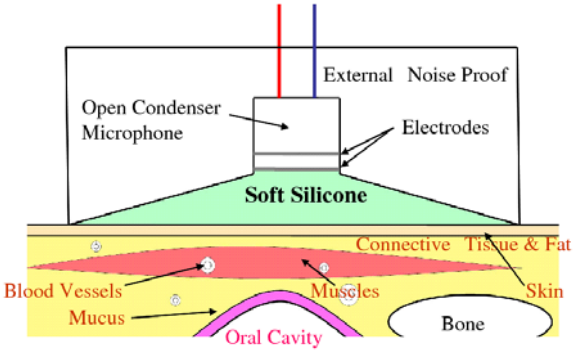

Fig. 2. The cross section of a NAM microphone and human body.

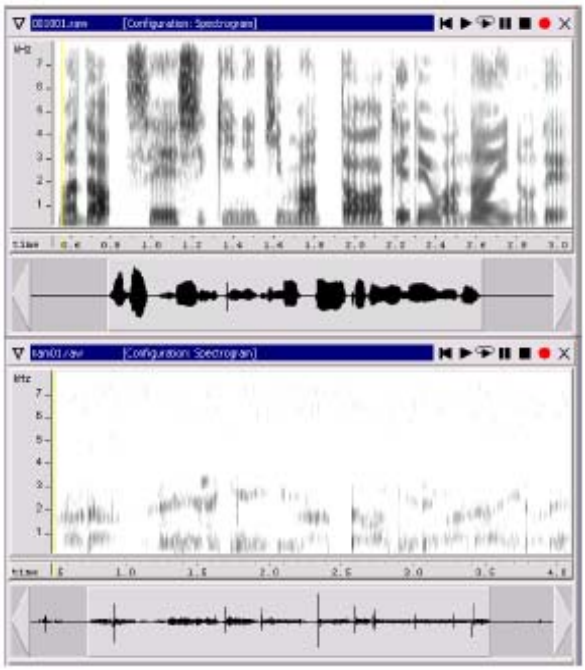

Fig. 3. Comparison of Spectrogram and waveform of normal speech (top) and NAM (bottom).

2.2. Global alignment kernel

The GA kernel [5] elaborates on the DTW family of distances by considering the same set of elementary operations, namely substitutions and repetitions of tokens, to map one sequence onto another. Associating a certain score with each of these operations, DTW algorithms use dynamic programming techniques to compute an optimal sequence of operations with a high overall score. For the GA kernel, the score spanned by all possible alignments is instead considered and a smoothed version of their maximum is taken. Its effectiveness was confirmed through comparison with the conventional HMM-based method (HMM: hidden Markov method) in isolated-word speech recognition experiments.

Now let us consider the alignment π of length $|\pi| = p$ between two sequences $\mathbf{x}$ and $\mathbf{y}$ and a pair of increasing p-tuples (π_1, π_2) such that

$$1 = \pi_1(1) \le \cdots \le \pi_1(p) = n,$$

$$1 = \pi_2(1) \le \cdots \le \pi_2(p) = m,$$

with unitary increments and no simultaneous repetitions. We write the DTW score $S(\pi)$ as

$$S(\pi) = \sum_{i=1}^{|\pi|} \varphi(x_{\pi_1(i)}, y_{\pi_2(i)}) \tag{1}$$

where φ is an arbitrary conditionally positive-definite kernel. The GA kernel is then defined as

$$K(\mathbf{x},\mathbf{y}) = \sum_{\pi \in A(\mathbf{x},\mathbf{y})} e^{S(\pi)} = \sum_{\pi \in A(\mathbf{x},\mathbf{y})} e^{\sum_{i=1}^{|\pi|} \varphi(x_{\pi_1(i)}, y_{\pi_2(i)})} = \sum_{\pi \in A(\mathbf{x},\mathbf{y})} \prod_{i=1}^{|\pi|} k(x_{\pi_1(i)}, y_{\pi_2(i)}) \tag{2}$$

where $A(\mathbf{x}, \mathbf{y})$ is the set of all possible alignments between $\mathbf{x}$ and $\mathbf{y}$ and $k = e^{\varphi}$.

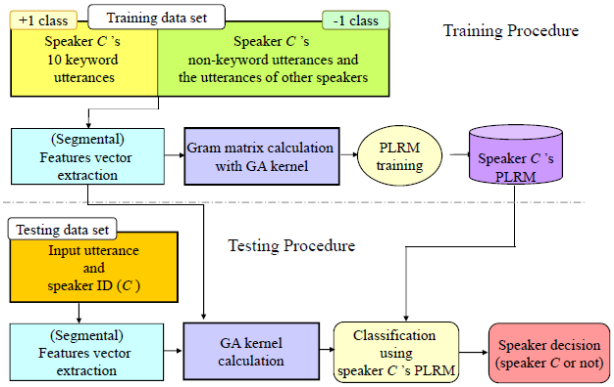

Fig. 4. Training and testing procedures.

2.3. Penalized logistic regression machine

In PLRM [6,7], the posterior probability of a class y given an observation x is modeled as

$$P(y \mid x) = \frac{\exp(f(x,y))}{\sum_{y' \in Y} \exp(f(x,y'))}, \tag{3}$$

where

$$f(x,y) = \sum_n \alpha_n(y) k(x_n, x). \tag{4}$$

Here, $\{x_n\}$ is a set of training observations, $k(\cdot,\cdot)$ is a kernel, and $\alpha_n(y)$ are the kernel product weights, which are the free parameters of the model and subject to optimization. It should be evident from these formulations that $P(y|x)$ is a real number between zero and one.

Assume that we have a collection of N feature vectors x_n and corresponding labels y_n. Let A be a $|Y| \times N$ matrix containing all of the kernel product weights $\{\alpha_n(y)\}$. The weights are optimized according to the negative log likelihood of the training labels:

$$L(A) = -\sum_n \log P(y_n \mid x_n) \quad (5)$$

$$\hat{A} = \arg\min_A L(A) . \quad (6)$$

Although we do not present the details here, the gradient and Hessian of the total loss $L(A)$ with respect to A have some very nice properties that enable us to use conjugate gradient descent methods relatively efficiently. To avoid overtraining and subsequent poor generalization, one usually adds a regularization term to the total loss.

2.4. Speaker verification procedure

The procedure of our method is shown in Fig. 4. In training, a PLRM is trained for each speaker. Although PLRM is a multi-class classifier, we use it here as a binary classifier. 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. Concatenations of n short-term feature vectors extracted from the training data are made for each utterance and used as an input vector sequence to calculate the Gram matrix with the GA kernel. The concatenation was originally assumed to represent keyword-specific acoustic features well. By utilizing the GA kernel, we further capture the keyword-specific acoustic features. The PLRM parameters are estimated using the Gram matrix.

In testing, as in training, concatenations of n short-term feature vectors are made for the input utterance, and the values of the GA kernel function between the input utterance and all training utterances are calculated. The values are given to the PLRM of the claimed speaker, and the probabilistic estimate of the speaker is obtained. This estimate is compared with a threshold to judge the speaker's identity.

3. Experiments

We compared the performances of our GA kernel+PLRM-based method and the polynomial kernel+SVM method in speaker verification experiments.

3.1. Data description and experimental conditions

We used keyword phrases uttered by 18 male and 10 female speakers in four sessions over a 9-month period (Jun. 2005, Sep. 2005, Dec. 2005 and Feb. 2006) as customer data, while we used keyword phrases uttered by a different set of 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 his/her own keyword 16 times and uttered 29 keywords of other customers/imposters twice. An MFCC (Mel frequency cepstral coefficient) vector of 31 components, consisting of 15-dimensional MFCCs plus ΔMFCCs and Δpower, 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

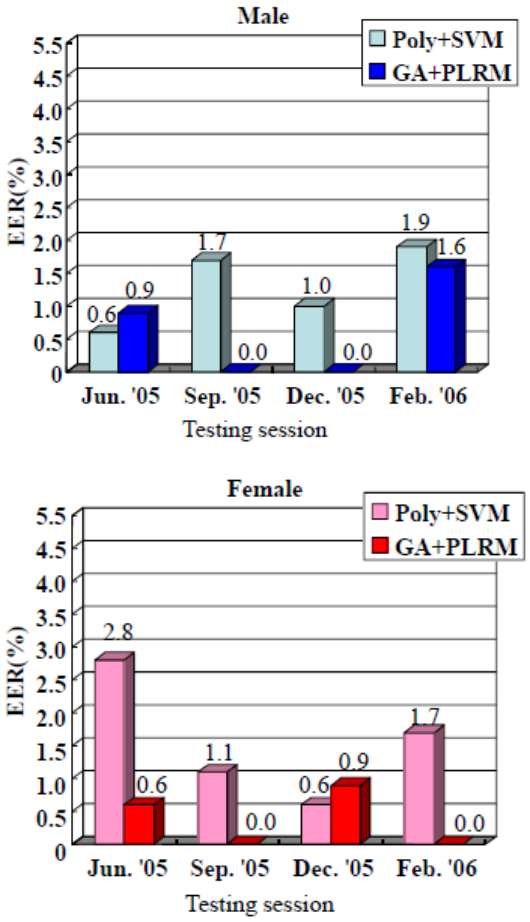

Fig. 5. Comparison of the performance for each session between the GA kernel+PLRM method and the polynomial.

of MFCCs (ΔMFCCs) computed in terms of five successive MFCC vectors. Cepstrum mean normalization was applied.

The training dataset 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 dataset 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 dataset the "basic case".

The threshold for speaker decision was speaker-dependent and set a posteriori to equalize the false acceptance and false rejection rates. For the GA kernel function, φ was defined as the Gaussian kernel with parameter $\sigma = 1$. For the polynomial kernel function, the power was chosen to be 7. For SVM, we use SVM^{light}, which is a toolkit provided by Cornell University [9].

Table 1. Comparison of equal error rates between the GA kernel+PLRM method and the polynomial kernel+ SVM method when using 25- and 85-ms-long NAM segments (basic case).

		25 ms	85 ms
Male	Polynomial kernel+SVM	2.2	1.3
	GA kernel+PLRM	**0.9**	**0.6**
Female	Polynomial kernel+SVM	1.2	1.6
	GA kernel+PLRM	**0.7**	**0.4**

3.2. Results

Table 1 lists the equal error rates (EERs) in the basic case for NAM segments with lengths of 25 ms (31-dimensional vector; MFCC+ΔMFCC+Δpower) and 85 ms (85-dimensional vector; 7 MFCC vector concatenations). In [3], we compared the performances with 45-ms, 85-ms and 145-ms-long segments and found that 85-ms-long segments were practical. Therefore we selected to use 85-ms-long segments here. For both male and female speakers, our GA kernel+PLRM method outperformed the polynomial kernel+SVM method: the averaged error reduction rate was 59%. Moreover, the size of the Gram matrix for the GA kernel was roughly the reciprocal of 250×250 (roughly 250 vectors in each utterance) times the size of the Gram matrix for the polynomial kernel and the PLRM training was fast. These results indicate that our GA kernel+PLRM method is effective and can capture keyword-specific acoustic features very well.

Figure 5 compares the EERs for each session for testing between our GA kernel+PLRM method and the polynomial kernel+SVM method. For half of the sessions, the EERs were 0.0, so our method is efficient.

4. Discussion

In practical conditions, we sometimes need to assume that keyword utterances uttered by other speakers (impersonation case) or non-keyword utterances uttered by the customer speaker (incorrect keyword case) are false utterances. For testing, in the impersonation case, we assigned the customer keyword utterances uttered by the other speakers to the -1 class, and in the incorrect keyword case, we assigned the non-keyword utterances uttered by the customer speaker to the -1 class.

Table 2 lists the EERs for the impersonation and incorrect keyword cases, respectively. While the EERs of our GA kernel+PLRM method were lower than those of the polynomial kernel+SVM method for the incorrect keyword case, the results were opposite for the impersonation case. For both cases, the EERs for 85-ms-long NAM segments were lower than those for 25-ms-long ones. It can be considered that since the GA kernel captures keyword-specific acoustic features very well, it has trouble dealing with the impersonation case. However, in a real situation, the keyword of the customer cannot be stolen because NAM is not captured by others. Moreover, speaker-specific features are well represented in NAM segments.

5. Conclusions

We investigated speaker verification using NAM segments based on a combination of the GA kernel and PLRM. Our method was found to be effective especially for the basic and incorrect keyword cases and reduced the error rates by more than half. Keyword-specific acoustic features are well represented on the Gram matrix of the GA kernel and the NAM segments. In future, we plan to investigate a priori threshold settings for verification.

Table 2. Comparison of equal error rates between the GA kernel+PLRM method and the polynomial kernel+SVM method when using 25 and 85-ms-long NAM segments.

Impersonation case			
		25 ms	85 ms
Male	Polynomial kernel+SVM	6.9	7.0
	GA kernel+PLRM	**13.9**	**9.5**
Female	Polynomial kernel+SVM	14.6	11.3
	GA kernel+PLRM	**29.0**	**18.4**
Incorrect keyword case			
		25 ms	85 ms
Male	Polynomial kernel+SVM	2.9	2.3
	GA kernel+PLRM	**1.3**	**0.7**
Female	Polynomial kernel+SVM	2.7	2.6
	GA kernel+PLRM	**2.0**	**0.5**

6. References

[1] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology," *in Proc. ICASSP*, pp. 4072-4075, 2002.

[2] M. Kojima, T. Matsui, H. Kawanami, H. Saruwatari and K. Shikano, "Speaker Verification with Non-Audible Murmur Segments," *in Proc. Interspeech*, pp. 2114-2117, 2006.

[3] H. Okamoto, M. Kojima, T. Matsui, H. Kawanami, H. Saruwatari, and K. Shikano, "Study on Speaker Verification with Non-Audible Murmur Segments," *in Proc. Interspeech*, pp. 2017-2020, 2007.

[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] M. Cuturi, J. P. Vert, O. Birkenes, and T. Matsui, "A Kernel for Time Series Based on Global Alignments," *in Proc. ICASSP*, pp. 413-416, 2007.

[6] K. Tanabe, "Penalized Logistic Regression Machines: New methods for statistical prediction 1," *ISM Cooperative Research Report* 143, pp. 163-194, 2001.

[7] K. Tanabe, "Penalized Logistic Regression Machines: New methods for statistical prediction 2," *in Proc. IBIS*, Tokyo, pp. 71-76, 2001.

[8] T. Matsui and K. Tanabe, "Comparative Study of Speaker Identification Methods: dPLRM, SVM and GMM," *IEICE Trans. Information and Systems*, vol. E89-D, no. 3, pp. 1066-1073, 2006.

[9] T. Joachims, SVM^{light}, http://svmlight.joachims.org/