ACCELERATION OF SEQUENCE KERNEL COMPUTATION FOR REAL-TIME SPEAKER IDENTIFICATION

Makoto Yamada, [1]*Masashi Sugiyama,* [2]*Gordon Wichern, and* [3] *Tomoko Matsui*

[1]Department of Computer Science, Tokyo Institute of Technology
[2]Arts, Media, and Engineering, Arizona State University
[3]Department of Statistical Modeling, Institute of Statistical Mathematics

e-mail: makoto.05@engalumni.colostate.edu

ABSTRACT

The sequence kernel has been shown to be a promising kernel function for learning from sequential data such as speech and DNA. However, it is not scalable to massive datasets due to its high computational cost. In this paper, we propose a method of approximating the sequence kernel that is shown to be computationally very efficient. More specifically, we formulate the problem of approximating the sequence kernel as the problem of obtaining a *pre-image* in a reproducing kernel Hilbert space. The effectiveness of the proposed approximation is demonstrated in text-independent speaker identification experiments with 10 male speakers—our approach provides significant reduction in computation time with limited performance degradation. Based on the proposed method, we develop a real-time kernel-based speaker identification system using Virtual Studio Technology (VST).

Index Terms— Sequence kernel, k-means algorithm, pre-image, Virtual Studio Technology (VST)

1. INTRODUCTION

Automatic speaker identification is a crucial user interface technology and has applications in various areas, e.g., pin-code-based security systems for mobile devices, conference systems [1], and robotics. In these applications, speaker identification is expected to work in real-time. Thus, the time response, or time spent on identification should be minimized.

Kernel methods such as the support vector machine (SVM) [2] and kernel logistic regression (KLR) [3] are successful approaches in speaker identification, given that the kernel functions are designed appropriately. Recently, a *mean operator sequence kernel* (MOSK) has been introduced for speaker identification [4], which utilizes a sequence of frame-level features for capturing long-term structure in phones, syllables, words, and entire utterances. MOSK measures the similarity between two sequences by computing the inner product between the means of the sequences *implicitly* in the feature space. The MOSK based speaker verification system was shown to significantly outperform other methods such as the Gaussian mixture model (GMM) and the SVM with finite-dimensional kernels.

Although MOSK performs well in the speaker verification task, its computational complexity limits its use in applications where real time processing is required. Specifically, MOSK requires NN' vector kernel computations for measuring the similarity between two data sequences of length N and N', respectively. The goal of this paper is to develop a computationally efficient alternative to the MOSK for real time speaker identification. The first step in our approach is to approximate the MOSK using k-means clustering. Then, we formulate the problem of approximating the sequence kernel as the problem of obtaining a *pre-image* in a reproducing kernel Hilbert space (RKHS) [2]. A pre-image is a vector in the input space mapped to the target feature vector in the RKHS.

The practical effectiveness of the proposed method is investigated in text-independent speaker identification experiments with 10 male speakers. Results demonstrate that the proposed method provides significant reduction in computation time while speaker identification accuracy is only moderately degraded. Furthermore, using the pre-image approximation we develop a real-time speaker identification system using Virtual Studio Technology (VST).

2. PROBLEM FORMULATION

In this section, we formulate the speaker identification problem based on the kernel logistic regression (KLR) model.

2.1. Kernel-based Text-independent Speaker Identification

An utterance sample X pronounced by a speaker is expressed as a set of N *mel-frequency cepstrum coefficient (MFCC)* [5] vectors of dimension d:

$$\mathrm{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{d \times N}.$$

For training, we are given n labeled utterance samples:

$$\mathcal{Z} = \{(\mathrm{X}_i, y_i)\}_{i=1}^n,$$

where $y_i \in \{1, \ldots, K\}$ denotes the index of the speaker who pronounced X_i. The goal of speaker identification is to predict the speaker index of a test utterance sample X based on the training samples. We predict the speaker index c of the test sample X following *Bayes decision rule*:

$$\max_c p(y = c \,|\, \mathrm{X}).$$

For approximating the class-posterior probability, we use

$$p(y = c \,|\, \mathrm{X}; \mathrm{V}) = \frac{\exp f_{\boldsymbol{v}_c}(\mathrm{X})}{\sum_{l=1}^K \exp f_{\boldsymbol{v}_l}(\mathrm{X})},$$

where $\mathrm{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K]^\top \in \mathbb{R}^{K \times n}$ is the parameter, $^\top$ denotes the transpose, and $f_{\boldsymbol{v}_l}$ is a discriminant function corresponding to speaker l. This form is known as the *softmax* function and widely used in multiclass logistic regression. We use the following kernel regression model as the discriminant function $f_{\boldsymbol{v}_l}$:

$$f_{\boldsymbol{v}_l}(\mathrm{X}) = \sum_{i=1}^n v_{l,i} \mathcal{K}(\mathrm{X}, \mathrm{X}_i) \quad l = 1, \ldots, K,$$

where $\boldsymbol{v}_l = (v_{l,1}, \ldots, v_{l,n})^\top \in \mathbb{R}^n$ are parameters corresponding to speaker l and $\mathcal{K}(\mathrm{X}, \mathrm{X}')$ is a kernel function.

We employ maximum likelihood estimation for learning the parameter V. The negative log-likelihood function $\mathcal{P}^{\log}(\mathrm{V}; \mathcal{Z})$ for the kernel logistic regression model is given by

$$\mathcal{P}^{\log}(\mathrm{V}; \mathcal{Z}) = -\sum_{i=1}^n \log P(y_i \,|\, \mathrm{X}_i; \mathrm{V}),$$

where $\mathrm{K} = [\mathcal{K}(\mathrm{X}_i, \mathrm{X}_j)]_{i,j=1}^n$ is the kernel Gram matrix. $\mathcal{P}^{\log}(\mathrm{V}; \mathcal{Z})$ is a convex function with respect to V and therefore its unique minimizer can be obtained using, e.g., the Newton method [6].

2.2. Mean Operator Sequence Kernel [4]

The performance of KLR depends on the choice of the kernel function. In this paper, we use the *mean operator sequence kernel* (MOSK) [4] as the kernel function since it allows us to handle feature sequences of different length. For sequences of d-dimensional feature vectors of length N and N',

$$\mathrm{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{d \times N},$$
$$\mathrm{X}' = [\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{N'}] \in \mathbb{R}^{d \times N'},$$

MOSK is defined as

$$\begin{aligned} \mathcal{K}(\mathrm{X}, \mathrm{X}') &= \frac{1}{N} \sum_{p=1}^N \boldsymbol{\phi}(\boldsymbol{x}_p)^\top \frac{1}{N'} \sum_{p'=1}^{N'} \boldsymbol{\phi}(\boldsymbol{x}'_{p'}), \\ &= \frac{1}{NN'} \sum_{p=1}^N \sum_{p'=1}^{N'} k(\boldsymbol{x}_p, \boldsymbol{x}'_{p'}), \end{aligned}$$

where

$$k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}')$$

is a 'base' vector kernel function.

MOSK requires NN' vector kernel computations for calculating the similarity between utterances X and X′. Therefore, the MOSK computation is not suited for real-time application when NN' is very large.

3. APPROXIMATION OF MOSK

In this section, we provide an approximation method of the MOSK computation. Below, we focus on the Gaussian kernel as the base kernel function:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right).$$

3.1. Approximating Mean Operator Sequence Kernel by Parts

For $D \ll N$, let us divide the samples $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ into D clusters $\{\mathcal{C}_1, \ldots, \mathcal{C}_D\}$ such that

$$\begin{aligned} \mathcal{C}_i \cap \mathcal{C}_j &= \emptyset \quad \text{for } i \neq j, \\ \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_D &= \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}. \end{aligned}$$

We may use the k-means clustering algorithm for this purpose. Then, $\frac{1}{N}\sum_{p=1}^N \phi(\boldsymbol{x}_p)$ can be expressed as

$$\begin{aligned} \frac{1}{N} \sum_{p=1}^N \phi(\boldsymbol{x}_p) &= \frac{1}{N} \left\{ \sum_{\boldsymbol{x} \in \mathcal{C}_1} \phi(\boldsymbol{x}) + \cdots + \sum_{\boldsymbol{x} \in \mathcal{C}_D} \phi(\boldsymbol{x}) \right\}. \\ &= \frac{\pi_1}{N_1} \sum_{\boldsymbol{x} \in \mathcal{C}_1} \phi(\boldsymbol{x}) + \cdots + \frac{\pi_D}{N_D} \sum_{\boldsymbol{x} \in \mathcal{C}_D} \phi(\boldsymbol{x}), \end{aligned} \quad (1)$$

where N_i is the number of samples in cluster $\mathcal{C}_i$ and $\pi_i = N_i/N$.

If we can approximate the mean $\frac{1}{N_i}\sum_{\boldsymbol{x} \in \mathcal{C}_i} \phi(\boldsymbol{x})$ by a single point $\phi(\boldsymbol{\mu}_i)$, the computational cost of the mean in the feature space is reduced from $\mathcal{O}(N)$ to $\mathcal{O}(D)$. To obtain a good approximation point $\boldsymbol{\mu}_i$, we minimize the following criterion:

$$J_i(\boldsymbol{\mu}_i) = \|\phi(\boldsymbol{\mu}_i) - \frac{1}{N_i} \sum_{\boldsymbol{x} \in \mathcal{C}_i} \phi(\boldsymbol{x})\|^2.$$

This is often called the *pre-image* problem in the context of kernel methods [2]. For the Gaussian kernel, the above criterion can be written as

$$J_i(\boldsymbol{\mu}_i) = 1 - \frac{2}{N_i} \sum_{\boldsymbol{x} \in \mathcal{C}_i} k(\boldsymbol{\mu}_i, \boldsymbol{x}) + \frac{1}{N_i^2} \sum_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{C}_i} k(\boldsymbol{x}, \boldsymbol{x}'), \quad (2)$$

where we used

$$k(\boldsymbol{\mu}_i, \boldsymbol{\mu}_i) = \exp\left(-\frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_i\|^2}{2\sigma^2}\right) = 1.$$

Taking the derivative of Eq.(2) with respect to $\boldsymbol{\mu}$, we have

$$\frac{\partial J_i(\boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}_i} = \frac{\partial}{\partial \boldsymbol{\mu}_i}\left[-\frac{2}{N_i}\sum_{\boldsymbol{x}\in\mathcal{C}_i}\exp\left(-\frac{\|\boldsymbol{\mu}_i-\boldsymbol{x}\|^2}{2\sigma^2}\right)\right]$$
$$= \frac{1}{\sigma^2 N_i}\sum_{\boldsymbol{x}\in\mathcal{C}_i}\exp\left(-\frac{\|\boldsymbol{\mu}_i-\boldsymbol{x}\|^2}{2\sigma^2}\right)(\boldsymbol{\mu}_i-\boldsymbol{x}). \quad (3)$$

Equating Eq.(3) to zero, we have:

$$\widehat{\boldsymbol{\mu}}_i = \frac{\sum_{\boldsymbol{x}\in\mathcal{C}_i}\exp\left(-\frac{\|\boldsymbol{\mu}_i-\boldsymbol{x}\|^2}{2\sigma^2}\right)\boldsymbol{x}}{\sum_{\boldsymbol{x}'\in\mathcal{C}_i}\exp\left(-\frac{\|\boldsymbol{\mu}_i-\boldsymbol{x}'\|^2}{2\sigma^2}\right)}. \quad (4)$$

We use Eq.(4) as a re-estimation formula, i.e., $\widehat{\boldsymbol{\mu}}_i$ is updated by Eq.(4) with $\boldsymbol{\mu}_i$ in the right-hand side replaced by the current estimate $\widehat{\boldsymbol{\mu}}_i$ and this is repeated until convergence.

Then Eq.(1) yields

$$\frac{1}{N}\sum_{p=1}^{N}\phi(\boldsymbol{x}_p) \approx \sum_{i=1}^{D}\pi_i\phi(\widehat{\boldsymbol{\mu}}_i). \quad (5)$$

Based on Eq.(5), MOSK can be approximated by

$$\mathcal{K}(\mathrm{X},\mathrm{X}') \approx \sum_{i=1}^{D}\pi_i\phi(\widehat{\boldsymbol{\mu}}_i)^\top\sum_{i'=1}^{D'}\pi'_{i'}\phi(\widehat{\boldsymbol{\mu}}_{i'})$$
$$= \sum_{i=1}^{D}\sum_{i'=1}^{D'}\pi_i\pi'_{i'}k(\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\mu}}_{i'}). \quad (6)$$

Following the k-means clustering algorithm, we call the proposed method the *k-means operator sequence kernel* (k-MOSK). The number of vectorial kernel computations in the original MOSK is NN', while that in k-MOSK is DD'. Thus k-MOSK would be computationally much more efficient than MOSK given that D and D' are much smaller than N and N'. It is clear that k-MOSK satisfies positive definiteness; thus it is a valid kernel function.

The computation of the k-means clustering algorithm for every utterance in the test phase is expensive. So we compute the kernel between a training sample X and a test sample $\mathrm{X}' = \{x'_1,\ldots,x'_{N'}\}$ as

$$\mathcal{K}(\mathrm{X},\mathrm{X}') = \frac{1}{N'}\sum_{i=1}^{D}\sum_{p=1}^{N'}\pi_i k(\widehat{\boldsymbol{\mu}}_i,\boldsymbol{x}'_p). \quad (7)$$

4. EXPERIMENTS

In this section, we compare the performance of MOSK and k-MOSK with different numbers of clusters D in a speaker identification task.

Table 1. Training sentences and test words (in Japanese, written using the Hepburn system of Romanization).

	Contents
Training sentences:	1. seno takasawa hyakunanajusseNchi hodode mega ookiku yaya futotteiru 2. oogoeo dashisugite kasuregoeni natte shimau 3. tashizaN hikizaNwa dekinakutemo eha kakeru
Testing words:	1. mouichido 2. torikaeshi 3. teisei 4. horyuu 5. shoukai

4.1. System and Data Acquisition

The data for training and testing were collected from 10 male speakers, where each speaker uttered several different words as listed in Table 1.

The duration of an utterance for each training sentence was approximately four seconds. Thus, the total duration of utterances over three training sentences was approximately 12 seconds per speaker. For testing purposes, we use utterances of 5 words recorded in three sessions over six months with no time overlap to the training session. Thus the total number of test words was 150 (10 speakers $\times$ 5 words $\times$ 3 sessions).

A feature vector of 26 dimensions, consisting of 12 MFCCs, normalized log energy, and their first derivatives, is derived once every 10ms over a 25.6ms Hamming-windowed speech segment. We divide each training utterance into 300ms disjoint segments, each of which corresponds to a set of features of size 26×30. On the other hand, for testing, we use the whole utterance of each word consisting of approximately 1000ms duration for computing MOSK and k-MOSK since each word is treated as a single test sample.

4.2. Results

We evaluate the proposed k-MOSK with the several different numbers of clusters D. The Gaussian width σ in the base Gaussian kernel is chosen from

$$\{8, 10, 12, 14, 16\}$$

by 10-fold *cross-validation* (CV). In our preliminary experiments, we observed that the 10-fold CV scores tend to be heavily affected by the random split of the training samples. We conjecture that this is due to non-i.i.d. nature of the MFCC features, which is different from the theoretical assumptions of CV. In order to obtain reliable experimental results, we repeat the CV procedure 50 times with different random data splits and use the average score for model selection.

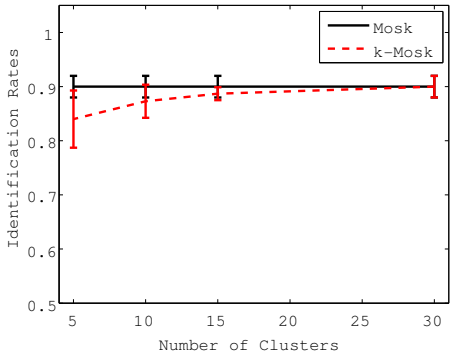

Fig. 1. Speaker identification rates obtained using 30, 15, 10, and 5 clusters, with selected kernel widths of 12, 14, 14, and 16, respectively.

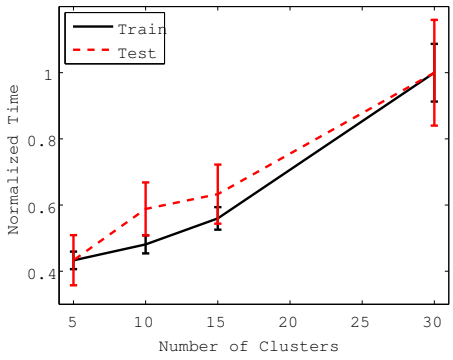

Fig. 2. The normalized computation time of MOSK and k-MOSK in training and testing using a standard personal computer with QuadCore 2.0GHz processor and 2GB memory.

Figure 1 depicts the speaker identification rates for the test words using MOSK and k-MOSK with different numbers of clusters D. In Figure 2, we plot the computation time of MOSK and k-MOSK in training and testing using a standard personal computer with a QuadCore 2.0GHz processor and 2GB memory. The computation time for MOSK is normalized to one. These results demonstrate that k-MOSK is computationally more efficient than the original MOSK with mild degradation in identification accuracy.

Based on k-MOSK, we have developed a real-time kernel-based speaker identification system using a Virtual Studio Technology (VST) plugin (see Figure 3). A demo movie is available at *http://dsp.syuriken.jp/demo/sid.html*.

5. CONCLUSION

The mean operator sequence kernel (MOSK) is a useful kernel function in speaker identification, but tends to be computationally inefficient. In this paper, we provided an approximation scheme based on the pre-image in a reproducing kernel Hilbert space. Through numerical experiments, the proposed method was shown to be useful in text-independent speaker identification when combined with kernel logistic regression and cross validation. In the future, we plan to implement the proposed speaker identification system in small devices such as digital signal processors (DSP) for robotics, conference systems, and human-computer interfaces.

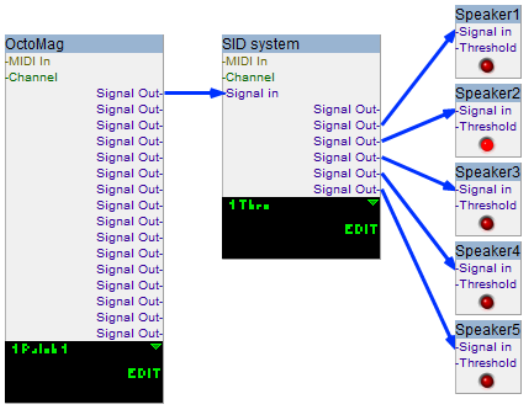

Fig. 3. Five-speaker identification system implemented with the VST plugin, where OctoMag is the waveplayer and the SID system is the kernel-based speaker identification module. Each LED lights when the corresponding speaker is speaking.

6. REFERENCES

[1] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, 2006.

[2] B. Schölkopf and A. J. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002.

[3] M. Yamada, M. Sugiyama, and T. Matsui, "Semi-supervised speaker identification under covariate shift," *Signal Processing*, 2009, to appear.

[4] J. Mariethoz and S. Bengio, "A kernel trick for sequences applied to text-independent speaker verification systems," *Pattern Recognition*, vol. 40, no. 8, pp. 2315–2324, 2007.

[5] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[6] K. Tanabe, "Penalized logistic regression machines: New methods for statistical prediction 1," Institute of Statistical Mathematics, Tech. Rep. 143, 2001.