Applied Acoustics 112 (2016) 79–83

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

# Speech enhancement method based on low-rank approximation in a reproducing kernel Hilbert space

Yanping Zhao [a], Robert Caiming Qiu [b], Xiaohui Zhao [a,*], Bo Wang [a]

[a] College of Communication Engineering, Jilin University, Changchun 130012, China
[b] Department of Electrical and Computer Engineering, Tennessee Technological University, Cookeville 38505, USA

ARTICLE INFO

*Article history:*
Received 23 October 2015
Received in revised form 6 May 2016
Accepted 9 May 2016
Available online 21 May 2016

*Keywords:*
Speech enhancement
Kernel function
Low-rank approximation
Error bound

ABSTRACT

Speech signal is corrupted unavoidably by noisy environment in subway, factory, and restaurant or speech from other speakers in speech communication. Speech enhancement methods have been widely studied to minimize noise influence in different linear transform domain, such as discrete Fourier transform domain, Karhunen-Loeve transform domain or discrete cosine transform domain. Kernel method as a nonlinear transform has received a lot of interest recently and is commonly used in many applications including audio signal processing. However this kind of method typically suffers from the computational complexity. In this paper, we propose a speech enhancement algorithm using low-rank approximation in a reproducing kernel Hilbert space to reduce storage space and running time with very little performance loss in the enhanced speech. We also analyze the root mean squared error bound between the enhanced vectors obtained by the approximation kernel matrix and the full kernel matrix. Simulations show that the proposed method can improve the computation speed of the algorithm with the approximate performance compared with that of the full kernel matrix.



## 1. Introduction

Speech enhancement reduces noise influence under the condition of minimum speech distortion, and plays a key role in mobile communication, automatic speech recognition, listening device and hearing aid. Over the past decades, many speech enhancement algorithms are proposed in time domain and transform domain. Since the energy of speech signal does not distribute in the all transform coefficients, many speech enhancement algorithms in the transform domain can easily filtered off the noise coefficients [1]. Moreover, most of them process signal by a linear transform, such as discrete Fourier transform (DFT) [2,3], Karhunen-Loeve transform (KLT) [4] or discrete cosine transform (DCT)[1]. DFT is a popular transform, however most of DFT-based speech enhancement algorithms only utilize amplitude spectrum, i.e., the methods ignore phase information. KLT is widely used in subspace speech enhancement with high computation load. Additionally, DCT is a real transform, which is active for its excellent energy compaction property.

Kernel method as a nonlinear transform has been applied successfully in machine learning, including classification, regression and clustering [5,6]. The main theme of this method is that it transforms the original nonlinear task to a linear task in a high dimensional feature space by a reproducing kernel. Kernel method can achieve nonlinear properties and usually obtain better performance compared with linear transform. Currently, there are some applications in audio signal processing in a reproducing kernel Hilbert space (RKHS). For example, a nonlinear filtering is presented for speaker verification in [7]. Aiming at speech recognition, a sparse auditory feature extraction method is proposed in [8]. Additionally, a framework is proposed for speech enhancement by considering the tradeoff between noise reduction and speech distortion in which the relationship between statistical learning and noise reduction is discussed in [9]. However, computing the kernel matrix leads to the algorithm in [9] with at least $O(l^2)$ storage space and $O(l^2p)$ running time when the size of observation matrix is $p \times l$. Moreover, the calculation of the inverse of a kernel matrix is the main computational burden with $O(l^3)$ running time. Aiming at this problem, we study the low-rank approximation of the kernel matrix to reduce the algorithm complexity.

In this study, based on low-rank approximation in a RKHS, we propose a speech enhancement method to reduce the computational complexity and obtain approximate performance compared with that of the full kernel matrix. Moreover, the root mean squared error bound is presented between the enhanced vectors

* Corresponding author.
*E-mail addresses:* zhaoyp@jlu.edu.cn (Y. Zhao), rqiu@tntech.edu (R.C. Qiu), xhzhao@jlu.edu.cn (X. Zhao), roywang@jlu.edu.cn (B. Wang).

http://dx.doi.org/10.1016/j.apacoust.2016.05.008
0003-682X/

obtained by the approximation kernel matrix and the full kernel matrix. In order to reduce the algorithm complexity, the main idea of this paper is to use the decomposition matrix to obtain the estimation of the clean speech instead of directly using the approximation of the kernel matrix.

The rest of this paper is organized as follows. Section 2 introduces kernel method and speech enhancement in a RKHS. Section 3 details our proposed speech enhancement based on low-rank approximation in a RKHS and discusses the error bound. In Section 4 we give speech enhancement experiments. Conclusion is given in Section 5.

## 2. Background statement

### 2.1. Kernel method

Kernel-based methods are one class of statistical learning algorithms which have successfully processed the nonlinear structure data [5,6,10]. In assumption of the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q \in \mathcal{X} \subset \mathbb{R}^p$, using a non-linear mapping function $\varphi: \mathcal{X} \to \mathcal{F}, \mathbf{x}$ can be mapped from the input space $\mathcal{X}$ into a higher dimensional feature space $\mathcal{F}$. In fact the dimension of the feature space $\mathcal{F}$ is so high that $\varphi(\mathbf{x})$ cannot be directly computed. Fortunately $\varphi(\mathbf{x})$ is not required in real applications. In kernel methods, the function $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$ is defined by $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \langle \varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j) \rangle$, thus we can use $\mathbf{K}$ instead of dot product to carry out the implicit mapping from the input space $\mathcal{X}$ into the feature space $\mathcal{F}$. $\mathbf{K}$ can be calculated by choosing a kernel function, such as the common polynomial kernel function

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \left(a + \langle \mathbf{x}_i, \mathbf{x}_j \rangle\right)^d, \tag{1}$$

where $a$ is the normalization parameter and $d$ is the nonlinearity degree parameter, or Gaussian kernel function

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{v}\right), \tag{2}$$

where $v$ is a constant.

### 2.2. Speech enhancement in a RKHS

Using the representer theorem, the novel speech enhancement method in a RKHS is proposed, in which noise reduction and speech distortion are well controlled by the regularization parameter and the function approximation explained from the viewpoint of prediction and spectral filtering [9]. The noisy speech signal is written as

$$y_i = f(\mathbf{x}_i) + n_i, \tag{3}$$

where $y_i$ and $n_i$ are the noisy speech signal and the noise signal respectively, and $|y_i| \leqslant 1$. $f(\mathbf{x}_i)$ is a target function that represents the clean speech signal and can be learned from the observation set $S = \{(\mathbf{x}_i, y_i),\ i = 0, \ldots, l-1\}$. Here, $l$ denotes the frame length, and $\mathbf{x}_i \in \mathbb{R}^p$. $\mathbf{X}_l$ is constructed by $y_i$ as follows

$$\mathbf{X}_l = [\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{l-1}] = \begin{bmatrix} y_0 & y_1 & \cdots & y_{l-1} \\ y_{-1} & y_0 & \cdots & y_{l-2} \\ \vdots & \vdots & \cdots & \vdots \\ y_{1-p} & y_{2-p} & \cdots & y_{l-p} \end{bmatrix}. \tag{4}$$

In $\mathbf{X}_l$, the elements with the index less than zero are the loop-shifted value of each frame.

The target function $f(\cdot)$ can be found by minimizing the objective function

$$\begin{aligned} \hat{f}_K &= \arg\min_f H(f) \\ H(f) &= \frac{1}{l}\sum_{i=0}^{l-1} \|y_i - f(\mathbf{x}_i)\|_2^2 + \alpha \|f\|_K^2, \end{aligned} \tag{5}$$

where the subscript $K$ of $\hat{f}_K$ denotes that the full kernel matrix is used in the algorithm, $\|f\|_K^2$ is the norm of the function in a RKHS, $\alpha$ represents the regularization parameter for a tradeoff between speech distortion and noise reduction. Based on the representer theorem, the approximation of $f_K(\mathbf{x})$ is given by

$$\hat{f}_K(\mathbf{x}) = \sum_{i=0}^{l-1} c_i \mathbf{K}(\mathbf{x}, \mathbf{x}_i), \tag{6}$$

where $\mathbf{K}$ is the $l \times l$ kernel matrix whose $(i,j)$th element is given by the kernel function $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$ and $c$ is the coefficient. The objective function $H(f)$ can be rewritten as

$$H(f) = \frac{1}{l}(\mathbf{y} - \mathbf{K}\mathbf{c})^T(\mathbf{y} - \mathbf{K}\mathbf{c}) + \alpha \mathbf{c}^T \mathbf{K}\mathbf{c}, \tag{7}$$

where $\mathbf{y} = [y_0, \ldots, y_{l-1}]^T$. This is a convex differentiable function and the minimization can be obtained by setting the derivative of $H(f)$ with respect to $\mathbf{c}$ to zero. Thus the coefficient vector is given by

$$\mathbf{c} = (\mathbf{K} + \alpha l \mathbf{I})^{-1}\mathbf{y}, \tag{8}$$

where $\mathbf{I}$ is an appropriately-sized identity matrix. The estimation of the clean speech vector is obtained by

$$\hat{\mathbf{f}}_K = \mathbf{K}(\mathbf{K} + \alpha l \mathbf{I})^{-1}\mathbf{y}. \tag{9}$$

Although the use of reproducing kernels avoids directly computing the inner-product of the mapping function in the high-dimensional space, it still requires $O(l^2)$ storage space and $O(l^2 p)$ running time of $\mathbf{K}$. Moreover, the inverse calculation of $(\mathbf{K} + \alpha l \mathbf{I})$ needs $O(l^3)$ running time and it is the most computational intensive part [11]. If $l$ is large, the complexity of the algorithm will increase obviously.

## 3. Speech enhancement based on low-rank approximation in a RKHS

### 3.1. Statement of the proposed method

Since $\mathbf{K}$ is a symmetric positive semidefinite matrix (SPSD), it can be well approximated from a random subset of its columns [12]. Many low-rank approximation methods, such as Nyström-based method, column sampling, incomplete Cholesky decomposition and kernel matching pursuit [13], have been proposed in the literatures to improve computational efficiency and avoid the explicit storage with very little accuracy loss. In particular the Nyström low-rank approximation method is an efficient technique for solving large-scale learning applications [14].

The Nyström method constructs the low-rank approximation matrix using a subset of the columns of $\mathbf{K}$. By randomly and uniformly sampling $m$ columns from $\mathbf{K}$ without replacement, we construct $\mathbf{K}_{lm}$. Consequently $\mathbf{K}$ can be rearranged on the basis of the samples such that

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{mm} & \mathbf{K}_{21}^T \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix} \ \textit{and}\ \mathbf{K}_{lm} = \begin{bmatrix} \mathbf{K}_{mm} \\ \mathbf{K}_{21} \end{bmatrix}. \tag{10}$$

The approximation of $\mathbf{K}$ by the Nyström method is [15,16]

$$\widetilde{\mathbf{K}} = \mathbf{K}_{lm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{lm}^T, \tag{11}$$