A LINEAR ESTIMATOR BASED ON SUBSPACE DECOMPOSITION USING WAVELET TRANSFORM FOR SPEECH ENHANCEMENT

Chung-Hsien Yang and Jhing-Fa Wang
Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan 701, R.O.C.

ABSTRACT

In this paper, we perform the speech enhancement based on approximate Karhunen-Loeve transform. The signal is represented by using wavelet packet based on a basis search algorithm. The eigenvectors is evaluated from these bases. Then a linear estimator based on these eigenvectors is constructed and used to perform noise reduction. We evaluate the performance of this method by using the Aurora-2 database. It consist of connected digit utterances with various background noise. The spectrograms of enhanced speech are shown and the SNR improvement is also calculated. The experimental show that this method achieves satisfactory enhancement of speech.

1. INTRODUCTION

The subject of enhancing speech degraded by uncorrelated additive noise has been widely studied. Mittal and Phanm propose [1] a Karhunen-Loeve transform (KLT) based approach for speech enhancement. The basic principle is to decompose the vector space of the noisy speech into a speech-plus-noise subspace and a noise subspace. Enhancement is performed by removing the noise subspace and estimating the clean speech from the speech-plus-noise subspace [2]. The decomposition of the noisy speech is performed by KLT. KLT achieves the optimum orthogonal transform under the criteria of mean square error and represents a signal with uncorrelated coefficients because its basis are the eigenvectors of the covariance matrix. However, the computational complexity of finding these eigenvectors which requires diagonalizing a $n \times n$ matrix is $O(n^3)$.

An approximate KLT based on wavelet packet was introduced in [3]. The concept is to construct a library of orthonormal wavelet packet which is organized as a binary tree. Then, the best basis select algorithm according to a criterion is performed to travel the binary tree. The selected basis vectors are sorted in decreasing order. First $v$ basis vectors will be applied to KLT. Since $v < n$, the computational load is reduced.

In this paper, we used the approach in [1] to perform the noise reduction. For the decorrelation of covariance matrix, an approximate KLT algorithm based on wavelet instead of KLT is used. The selected approximate eigenvectors constitute a linear estimator under a criterion. Then, the estimation of clean speech is achieved by multiplying this matrix.

2. THE APPROXIMATE KLT

In this section, we will describe the components used in the approximate KLT respectively.

2.1 Wavelet Packet

The wavelet function considered here are Lemarie-Meyer's wavelets. They are orthonormal and possess good quality as well as time localization [4]. Walter and Zhang proposed [5] two simple closed-form expressions of Lemarie-Meyer's wavelets. One of them is summarized as follows

$$\psi(t + \frac{1}{2}) = \frac{1}{\pi (4 \beta t^2 - 1)} \sin \pi (1 + \beta)t - 4 \beta t \cos(1 - \beta)t,$$

$$\varphi(t) = \frac{\sin \pi (1 - \beta) t + 4 \beta t \cos \pi (1 + \beta)t}{\pi (1 - (4 \beta t^2))}.$$

The scaling function $\varphi(t)$ satisfies the two-scale difference equation

$$\varphi(t) = \sqrt{2} \sum_k h_k \varphi(2t - k).$$

Then, the wavelet function $\psi(t)$ is related to the scaling function via

$$\psi(t) = \sqrt{2} \sum_k g_k \varphi(2t - k).$$

These $\{h_k\}$ and $\{g_k\}$ are a pair of quadrature mirror filters that are related through

$$g_k = (-1)^{k+1} h_{-(k+1)}.$$ 

Plot of $\varphi(t)$ and $\psi(t)$ are shown in Fig. 1. The closed form of $h_k$ was also given in [5]

$$h[n] = \frac{\sqrt{2}}{\pi n (1 - 2 \beta n^2)} (\sin \pi (1 - \beta) \frac{n}{2} + 2 \beta n \cos \pi (1 + \beta) \frac{n}{2}).$$

The filter coefficients play a very crucial role to construct wavelet packet and the detail can be found in [6]. Usually, the wavelet packet is presented in a binary tree and each node of the binary tree is labeled by $(j, p)$ where $j$ is the depth of the node in.
2.2 The cost function for search algorithm

For a discrete signal \( f \) of size \( N \) defined on interval \([0,1]\), our goal is to choose the basis which leads to the best estimate \( \tilde{f} \) among a collection of orthonormal bases \( \{ B^p = \{ W^p \} | p \in P \} \), where \( p \) is a partition of the interval \([0,1]\) [7]. The best basis meets the criterion of minimal cost function. The cost function is defined as follows

\[
Cost(f, B^p) = \sum_{i=1}^{N} C(k, f, W^p_i, \gamma^2)
\]

where \( C(x) = -x \log x \).

Finding this minimum by comparison of the cost of all wavelet packet would require \( \mathcal{O}(N2^{N/2}) \). Coifman and Wickerhauser proposed a dynamic programming algorithm with \( \mathcal{O}(N \log N) \). This rough idea can be made precise as well as generalized to all libraries with a tree structure [8].

2.3 Finding the approximate KL-basis

Suppose there are \( N \) random vectors \( \{ X_n \in \mathbb{R}^d | n = 1,2,..,N \} \). The following steps will find the approximate KL-basis [3].

- Expanding \( N \) vectors into a complete wavelet packet coefficients;
- Calculating the variance at each node and search the variance tree for a best basis;
- Sorting the best basis vector in decreasing order;

Since we expect \( m<<d \), the reduction of computational load will be considered.

3. THE LINEAR ESTIMATOR

We used the linear estimator in [1] to perform noise reduction. Let \( z, y, \) and \( w \) be \( K \)-dimensional vectors and denote the noisy speech, clean speech and noise respectively. Since clean speech and noise are independent, we estimate the covariance matrix of clean speech \( R_y \) by

\[
\tilde{R}_y = \tilde{R}_z - \tilde{R}_w.
\]

Let \( \tilde{R}_y = U_y \Lambda_y U_y^\dagger \) be the eigenvalue decomposition of \( \tilde{R}_y \). Let \( M \) be the number of eigenvalue of \( \tilde{R}_y \) greater than zero. Let \( U_y = [U_1, U_2] \), where \( U_1 \) denotes the \( K \times M \) matrix of eigenvectors with positive eigenvalues

\[
U_1 = \{ u_{jk} | \lambda_j(k) > 0 \}.
\]

Let \( z_T = U_y^\dagger z = U_y^\dagger y + U_y^\dagger w = y_T + w_T \). The covariance matrix \( R_{w_T} \) of \( w_T \) is \( U_{y_T}^\dagger R_w U_{y_T} \). Let \( \sigma_{kk} \) be the \( k \)th diagonal element of \( R_{w_T} \). The estimate \( \tilde{y} \) is now obtained as

\[
\tilde{y} = H \tilde{z} \quad H = U_y Q U_y^\dagger
\]

where \( Q \) is a diagonal matrix.

\[
Q = diag(q_{kk}), \quad q_{kk} = \begin{cases} \alpha_k^{1/2}, & k = 1,2,...,M \\ 0, & \text{otherwise} \end{cases}
\]

and

\[
\alpha_k = \begin{cases} \exp(-\frac{\sigma^2_{w_T}(k)}{\lambda_j(k)}), & k = 1,2,...,M \\ \lambda_j(k), & \text{otherwise} \end{cases}
\]

where \( v \) is a predetermined constant.

The underlying assumption of this estimator is that the noise is stationary and also the covariance matrix \( R_w \) of the noise vector \( w \) is known. This covariance matrix may be estimated from the noise frame (silent).
4. EXPERIMENTAL RESULTS

In this section, Experiments results are presented with using the Aurora-2 database [9]. It consists of the connected digit utterances and addition of different background noise over a range of signal to noise ratios. Four different background noises were taken from the Aurora-2 database: airport noise, street noise, restaurant noise, and subway noise. The speech spectrograms of the test signals and enhanced signals are presented in Figs. 3-12. The scale of horizontal axis is time and the scale of vertical axis is frequency (0–4000 Hz).

The SNR of a speech signal is defined as

$$SNR = 10 \log_{10} \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} y_i^2(k)}{\sum_{i=1}^{N} \sum_{k=1}^{K} (z_i(k) - y_i(k))^2}$$

where $N$ is the number of frames in the given speech signal, $z_i(k)$ denotes the $k$th sample of the noisy speech or the enhanced speech, and $y_i(k)$ denotes $k$th sample of the clean speech. The amount of noise reduction is generally measured with the SNR improvement. For SNR 0, the SNR of enhanced speech under airport noise, street noise, restaurant noise, subway noise, and white noise are 1.14, 1.44, 1.20, 1.67, and 3.02 respectively. For SNR -5, the SNR of enhanced speech under airport noise, street noise, restaurant noise, subway noise, and white noise are 1.13, 0.48, 0.23, 0.16, and 2.32 respectively.

5. CONCLUSIONS

In this paper, an approach of enhancing speech based on approximate KLT using wavelet has been presented. This method is applied to Aurora-2 database to evaluate the performance. For four different background noises, the spectrograms and SNR
improvement of the enhanced speech is presented. The result of
spectrograms shows no musical residual noise and the SNR
improvement indicates noise is also reduced. The future work will
focus on increasing intelligibility of the enhanced speech.

6. REFERENCES

stationary random processes," IEEE Trans. Signal Processing,
best signal representation," IEEE Trans. Information Theory,
page.