SUBSPACE TRACKING FOR SPEECH ENHANCEMENT IN CAR NOISE ENVIRONMENTS

Jhing-Fa Wang[1], *Chung-Hsien Yang*[2] *and Kai-Hsing Chang*[3]

Department of Electrical Engineering, National Cheng-Kung University,
Tainan, Taiwan 701, R.O.C.

wangjf@csie.ncku.edu.tw[1] chyang@icwang.ee.ncku.edu.tw[2] casey019@ms55.hinet.net[3]

ABSTRACT

A signal subspace speech enhancement based on subspace tracking algorithm is presented. The proposed method incorporates a perceptual filterbank which is derived from psycho-acoustic model for subband processing. The experiments were performed using the TAICAR in-car noisy speech database. Subjective and objective tests show that our method outperforms other existing signal subspace methods.

1. INTRODUCTION

Speech enhancement attempts to improve perceptual aspects of voice communication systems when the signal is corrupted by noise (e.g., overall quality, intelligibility for human or speech recognizers). The improvement is in the sense of minimizing the effects of the noise on the system performance. Ephraim and Van-Trees proposed a signal subspace speech enhancement system [1]. The system decomposes the input signal into signal components and noise components. To improve speech quality, the noise components are discarded. Then, an estimation of the clean speech is made for the signal components. The decomposition has been done using Karhunen-Loeve transform (KLT). Subspace approaches have been successfully applied in the area of speech enhancement. However, it has a drawback of the high computational complexity for KLT.

In this paper, we propose a speech enhancement technique based on subspace methods. The subspace decomposition is achieved by using a subspace tracking algorithm [2]. The tracking method is a normalized least-mean-square (NLMS) adaptive filter. It has a computational complexity of linear order and is suitable for real time applications. To reduce signal distortion while applying the subspace tracking, a perceptual filterbank is used. The perceptual filterbank approximate the critical bands of psycho-acoustic model. To evaluate the proposed system, TAICAR database is used (**Tai**wan in-**CAR** speech database). The experimental results show that this system has good performance in car noisy environments.

The organization of the paper is as follows. Section 2 describes the overall system. It consists of perceptual filterbank, signal subspace speech enhancement and subspace tracking algorithm. In section 3, some experimental results are discussed. Section 4 draws conclusions.

2. THE PROPOSED SYSTEM

The proposed speech enhancement system based on subspace tracking is shown in Fig. 1. At the start of system processing, input signal is divided into subband time series by the analysis filterbank. Following subband analysis, the vector of subband signal is presented to the subspace tracking block to extract eigenvectors. Then, the gain adaptation is performed to estimate the clean speech. To reconstruct the enhanced full-band speech, the subband synthesizer is applied to the gain-modified vector of subband signal.

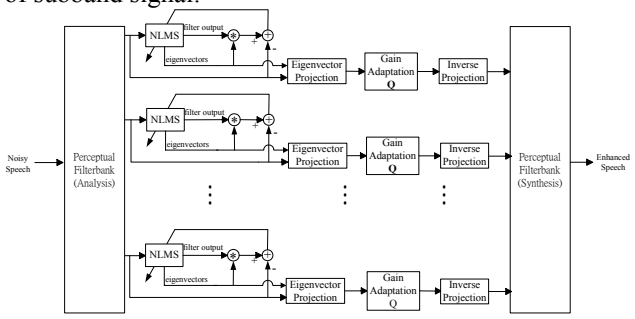


Fig. 1. Block diagram of the proposed system.

2.1. Perceptual Filterbank

The perceptual filterbank is obtained by adjusting the decomposition tree structure of the conventional wavelet packet transform in order to approximate the critical bands of the psycho-acoustic model as close as possible [3]. The primary reason for embedding the psycho-acoustic model in the filterbank is that humans are capable of detecting the desired speech in a noisy environment without prior knowledge of the noise [4]. One class of critical band scales is called Bark scale. The Bark scale z can be approximately expressed in terms of the linear frequency by

$$z(f) = 13\arctan(7.6\times10^{-4} f) + 3.5\arctan(1.33\times10^{-4} f)^2 \quad (1)$$

where f is the linear frequency in Hertz. The corresponding critical bandwidth (CBW) of the center frequencies can be expressed by

$$\mathrm{CBW}(f_c) = 25 + 75(1 + 1.4 \times 10^{-6} f_c^{\,2})^{0.69} \qquad (2)$$

where f_c is the center frequency (unit: Hertz). Theoretically, the range of human auditory frequency spreads from 20 to 20000 Hz and covers approximately 25 Barks. In this paper, the underlying sampling rate was chosen to be 8 kHz, yielding a bandwidth of 4 kHz. Within this bandwidth, there are approximately 17 critical bands as listed in Table I [5].

Table I: The characteristics of critical bands under 4 kHz

Critical Band Number	Center Frequency (Hz)	CBW	Lower Cutoff frequency (Hz)	Upper Cutoff Frequency (Hz)
1	50	-	-	100
2	150	100	100	200
3	250	100	200	300
4	350	100	300	400
5	450	110	400	510
6	570	120	510	630
7	700	140	630	770
8	840	150	770	920
9	1000	160	920	1080
10	1170	190	1080	1270
11	1370	210	1270	1480
12	1600	240	1480	1720
13	1850	280	1720	2000
14	2150	320	2000	2320
15	2500	380	2320	2700
16	2900	450	2700	3150
17	3400	550	3150	3700

According to the specifications of center frequencies, CBW, lower and upper cutoff frequencies given in Table I, the tree structure of the wavelet packet transform can be constructed as shown in Fig. 2(a). The corresponding frequency bandwidth of the wavelet packet tree is shown in Fig. 2(b). It contains 16 decomposition cells with 5 decomposition stages to approximate these 17 critical bands which are corresponding to wavelet packet coefficient sets $w_{j,m}$, where j=3, 4, 5, m=1, …, 17.

2.2. Signal Subspace Speech Enhancement

The model used in the subspace approach assumes that the noise signal is additive and uncorrelated with the speech signal,

$$\mathbf{y} = \mathbf{x} + \mathbf{n}, \qquad (3)$$

where **y**, **x**, and **n** are K-dimensional vectors and denote the noisy speech, clean speech and white noise respectively. Let $\hat{\mathbf{x}} = \mathbf{H}\mathbf{y}$ be a linear estimation of **y**, where **H** is a $K \times K$ matrix. The error signal is given by

$$\varepsilon = \hat{\mathbf{x}} - \mathbf{x} = (\mathbf{H} - \mathbf{I})\mathbf{x} + \mathbf{H}\mathbf{n} = \varepsilon_{\mathbf{x}} + \varepsilon_{\mathbf{n}}, \qquad (4)$$

where $\varepsilon_{\mathbf{x}}$ represents the signal distortion and $\varepsilon_{\mathbf{n}}$ represents the residual distortion [1]. Denoting the signal distortion energy by $\bar{\varepsilon}_{\mathbf{x}} = tr(E\{\varepsilon_{\mathbf{x}}\varepsilon_{\mathbf{x}}^T\})$ and the eigenvector matrix of covariance matrix of **x** by **U**=[$\mathbf{u}_1$, $\mathbf{u}_2$, …, $\mathbf{u}_K$], the spectral domain constrained (SDC) estimator **H** is obtained by

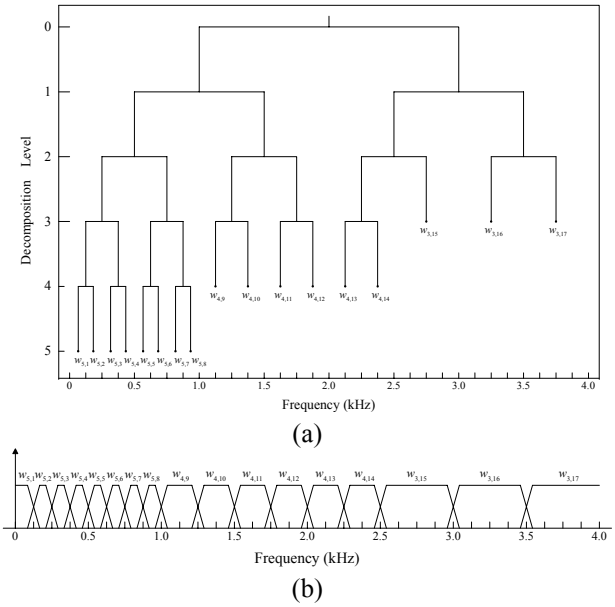


Fig. 2: (a) Tree structure of the perceptual filters. (b) The frequency bandwidths for the perceptual filters.

$$\mathbf{H} = \arg\min_{\mathbf{H}} \varepsilon_{\mathbf{x}}^2, \qquad (5)$$

with constraint

$$E\{\left|\mathbf{u}_k^T \varepsilon_{\mathbf{n}}\right|^2\} \le \alpha_k \sigma^2 \qquad k = 1,..., M,$$
$$E\{\left|\mathbf{u}_k^T \varepsilon_{\mathbf{n}}\right|^2\} = 0 \qquad k = M+1,..., K, \qquad (6)$$

where σ^2 and M are noise variance and the dimension of signal subspace respectively. Let the kth eigenvalue $\lambda_{\mathbf{x}}(k)$ associated with the kth eigenvector $\mathbf{u}_k$ be arranged in descending order for k=1, …, K. The estimator is given by

$$\mathbf{H} = \mathbf{U}\mathbf{Q}\mathbf{U}^T, \qquad (7)$$

where **Q** is a diagonal matrix with the kth diagonal element q_k given as the generalized Wiener filter of the form

$$q_k = \begin{cases} \alpha_k^{\,1/2} & k = 1,...M \\ 0 & k = M+1,...K, \end{cases} \qquad (8)$$

and

$$\alpha_k = \exp\{-\sigma^2 / \lambda_{\mathbf{x}}(k)\}. \qquad (9)$$

2.3. Subspace Tracking Algorithm

As it was discussed in Section 2.2, the estimator requires an accurate estimation of eigenvalues and eigenvectors of clean speech covariance matrix. Since noises in each subband are assumed to be white and uncorrelated with the clean speech, it is clear that the eigenvectors of noisy speech are the same as clean speech's. Hence, it may perform the eigen-decomposition of noisy speech for finding **U** and $\lambda_{\mathbf{x}}(k)$ can be obtained by subtracting the eigenvalues of noises from the eigenvalues of noisy signal.

We adopt the subspace tracking method [2] for extracting the eigenvectors. This is an adaptive method that tracks eigenvectors of covariance matrix using a normalized least-mean-square (NLMS) type algorithm. It is summarized in Table II.

Table II: Subspace tracking algorithm

Initialize : $d_i(0) = 0,\ \beta = 0.95$

$\mathbf{U}(0) = [\mathbf{u}_1(0) \mid \mathbf{u}_2(0) \mid \ldots \mid \mathbf{u}_k(0)] = \mathbf{I}_k$

For $n = 1, 2, \ldots$ do

$\mathbf{y}_1(n) = \mathbf{y}(n)$

For $i = 1, 2, \ldots, k$ do

$v_i(n) = \mathbf{u}_i^T(n-1)\mathbf{y}_i(n)$

$d_i(n) = \beta d_i(n-1) + |v_i(n)|^2$

$\mathbf{e}_i(n) = \mathbf{y}_i(n) - \mathbf{u}_i(n-1)v_i(n)$

$\mathbf{u}_i(n) = \mathbf{u}_i(n-1) + \mathbf{e}_i(n)\frac{v_i(n)}{d_i(n)}$

$\mathbf{y}_{i+1}(n) = \mathbf{y}_i(n) - \mathbf{u}_i(n)v_i(n)$

end

end

Output :

$\mathbf{U}(n) = [\mathbf{u}_1(n) \mid \mathbf{u}_2(n) \mid \ldots \mid \mathbf{u}_k(n)]$

3. EXPERIMENTAL RESULS

The evaluations are based on TAICAR speech database. The TAICAR database is briefly introduced below.

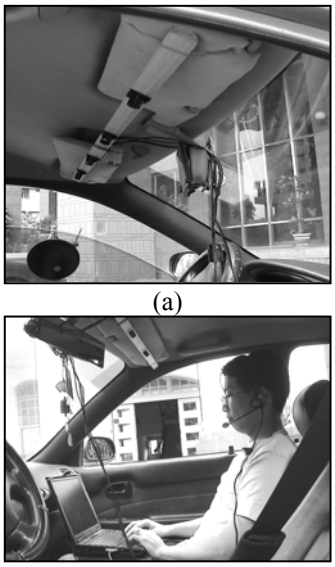
3.1. TAICAR Database

A group of researchers in the area of speech processing in Taiwan together initiate an in-car speech collection project called **TAICAR** (**Tai**wan in-**CAR** speech database) [6]. The objective of the TAICAR project is to produce an in-car Mandarin speech database which can be used as training and testing material for speech processing in car environment.

The detailed description of the recording elements is given below:

- A notebook PC with an Intel Pentium processor is the kernel of the speech data collection system.
- A DAQP PCMCIA multi-channel signal recording card, which is capable to record up to 16 channels of signal, is pluged into the notebook as the recording interface.
- Four omni-directional microphones form a linear microphone array (channel 0-3).
- One omni-directional microphone is place before the speaker (channel 4).
- One uni-directional microphone is worn on the head of the speaker (channel 5).

A pre-amplification circuit is utilized before the speech signal is feed to the PCMCIA card. The photographs showing the position of microphone array, and navigator are given in Fig. 3. Fig. 4 is a snapshot of the speech data.



(a) (b)

Fig. 3: (a) Microphone array placement. (b) Speaker and recording notebook.

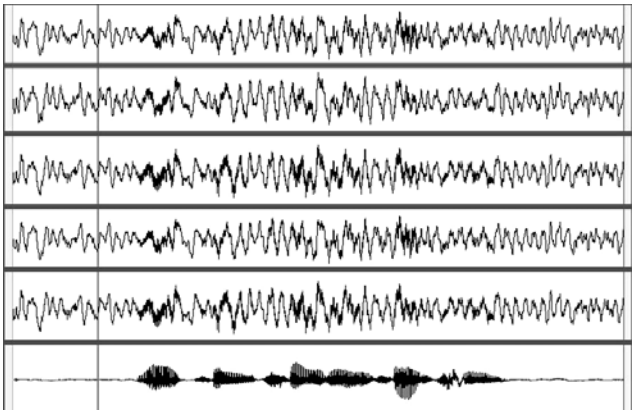


Fig. 4. Speech waveform of the utterance “EQ 7637” (in Mandarin), from channel 0 to channel 5 (top to bottom).

3.2. Performance Evaluation

We present analysis for the speech enhancement performance of three subspace decomposition methods: (1) using discrete cosine transform (DCT); (2) using KLT; (3) using the proposed approach. The results are shown in figures 5, 6 and 7. In these figures, the legends are (a) original noisy speech; (b) enhanced speech; (c) spectrogram of noisy speech and (d) spectrogram of enhanced speech. It is obviously that KLT and our approach are better than DCT. Furthermore, our approach outperforms the other methods in computational complexity. For an analysis frame length of n, the computational complexity of our work, DCT and KLT are $O(rn)$, $O(n\log n)$ and $O(n^3)$, respectively, where r is the number of eigenvectors.

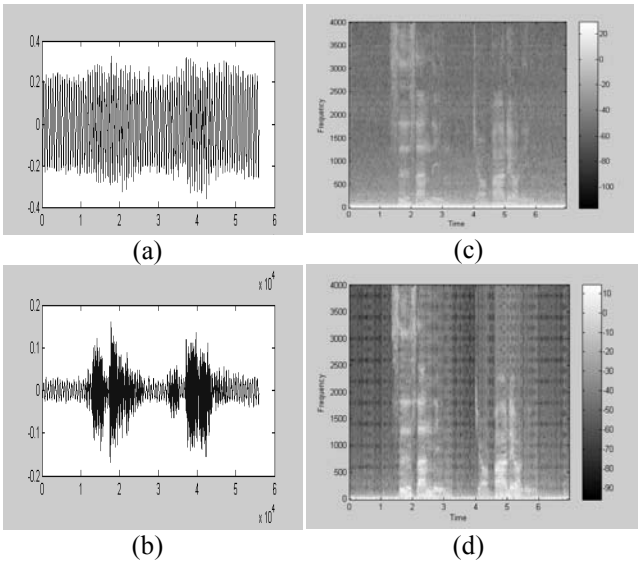
We also have another test to compare these different approaches. The evaluation was performed by a group of 20 listeners. Subjects were asked to rank the voice they heard. The voice consists of three type of noisy speech together with the enhanced results based on different approaches. The results are given in Table III. This evaluation is also known as mean opinion score (MOS) testing (5=Excellent; 4=Good; 3=Fair; 2=Poor; 1=Bad).

Table III: Five-grade MOS testing

	Speech from TAICAR Database		
	Car Ignited	Downtown Area	Highway
DCT	2.6	2.1	1.9
KLT	4.4	4.1	3.8
Ours	4.2	4.0	3.9

4. CONCLUSIONS

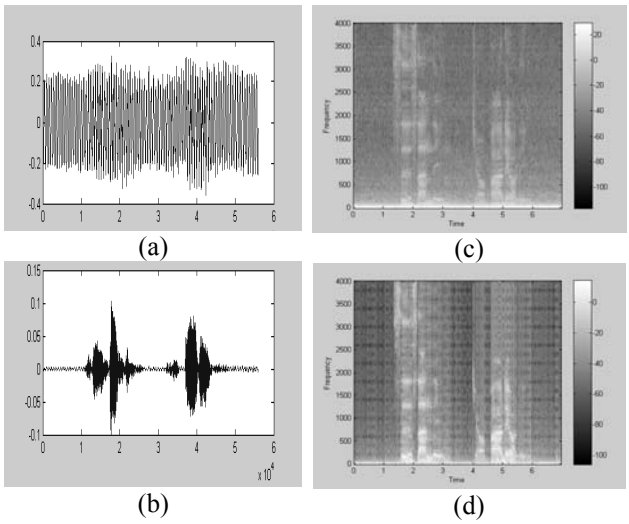
In this paper, a subspace tracking speech enhancement method was proposed. The proposed method incorporates psycho-acoustic model (perceptual filterbank) by adjusting the decomposition tree structure of the conventional wavelet packet transform. Experiments were carried out using the TAICAR in-car noisy speech database. According to the experiments and MOS evaluation, our method achieved enhancement performances very close to the KLT-based method. Another significant advantage of the proposed method is that the computational complexity was the best among the compared methods.



(a) (c)

(b) (d)

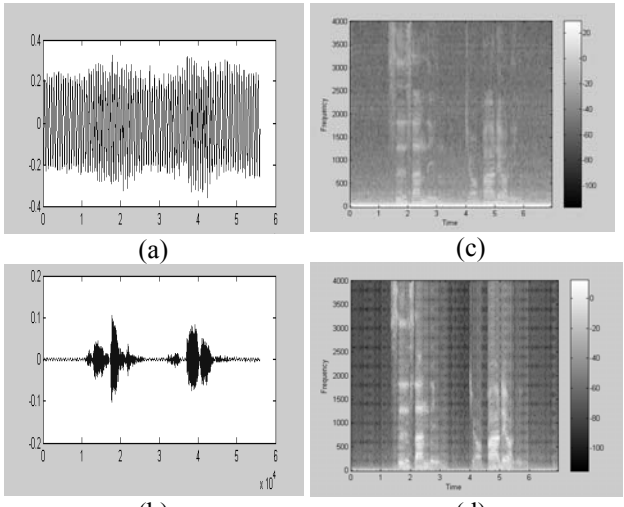
Fig. 5. Speech enhancement based on DCT.



(a) (c)

(b) (d)

Fig. 6. Speech enhancement based on KLT [1].



(a) (c)

(b) (d)

Fig. 7. Speech enhancement based on our approach.

5. REFERENCES

[1] Y. Ephraim and H. L. Van-Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 251–266, July 1995.

[2] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Signal Processing*, vol. 43, pp. 95–107, Jan. 1995.

[3] Shi-Huang Chen and Jhing-Fa Wang, "Speech Enhancement Using Perceptual Wavelet Packet Decomposition and Teager Energy Operator," accepted to appear in *The Journal of VLSI Signal Processing Systems*, Special Issue on Real World Speech Processing.

[4] O. Ghitza, "Auditory model and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 115-132, 1994.

[5] Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[6] Jhing-Fa Wang, Hsien-Chang Wang and Chung-Hsien Yang, "TAICAR - A Collection of In-Car Mandarin Speech Database in Taiwan," *O-COCOSDA2003 / PACLIC17*, Singapore.