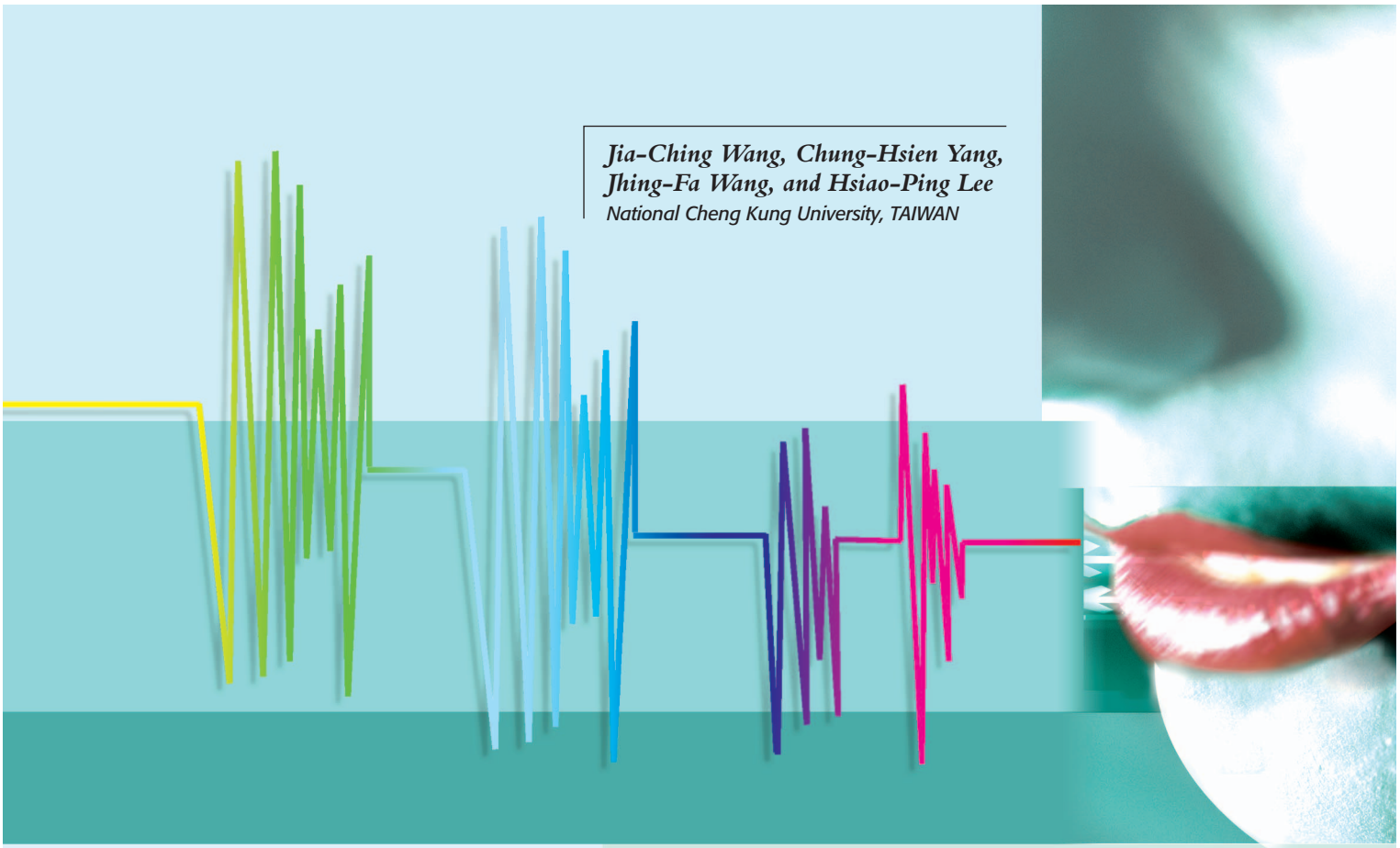

Jia-Ching Wang, Chung-Hsien Yang, Jhing-Fa Wang, and Hsiao-Ping Lee
National Cheng Kung University, TAIWAN

Robust Speaker Identification and Verification

Abstract: Acoustic characteristics have played an essential role in biometrics. In this article, we introduce a robust, text-independent speaker identification/verification system. This system is mainly based on a subspace-based enhancement technique and probabilistic support vector machines (SVMs). First, a perceptual filterbank is created from a psycho-acoustic model into which the subspace-based enhancement technique is incorporated. We use the prior SNR of each subband within the perceptual filterbank to decide the estimator's gain to effectively suppress environmental background noises. Then, probabilistic SVMs identify or verify the speaker from the enhanced speech. The superiority of the proposed system has been demonstrated by twenty speaker data taken from AURORA-2 database with added background noises.

I. Introduction

A biometric system makes a pattern recognition decision in accordance with the biometric features extracted from a human being. In recent years, various human characteristics such as the face, speech, fingerprint, and iris have been considered as discriminative features for automatic biometric recognition. In this article, we will address a speech-based biometric system, i.e. speaker recognition system. Basically, speaker recognition systems are divided into two main categories: speaker identification and speaker verification [1].

The block diagram of a general speaker identification/verification system is given in Figure 1. In a speaker identification system, an unknown speaker is identified as one of the speakers in the database. This assignment process is based on comparing the speech feature of this unknown person to that of each individual in the database. In a speaker verification system, a person's identity is validated based on his/her speech feature. The validation process usually relies on a likelihood threshold. In the past decade, there have been numerous speaker recognition algorithms presented in literature [2]–[10]. However, the performances of these speaker recognition systems have typically been drastically degraded in real-world noisy environments. To decrease the environmental noise

1556-603X/07/$25.00©2007IEEE

problem, this article introduces a robust speaker recognition architecture, which combines an SNR-sensitive subspace-based enhancement technique and probabilistic support vector machines (SVMs).

Among several strategies that have emerged for decreasing environment noise problems, front-end enhancement undoubtedly is a powerful one. Ephraim and Van Trees [11] proposed a subspace-based speech enhancement method. This method provides an optimal estimator that would minimize the speech distortion subject to the constraint such that the residual noise falls below a preset threshold. In this paper, we will use an SNR-sensitive subspace-based enhancement technique. First, the perceptual filterbank is obtained by adjusting the decomposition tree structure of the conventional wavelet packet transform [12]. This allows us to approximate the critical bands of the psychoacoustic model [13] as close as possible. The prior SNR of each subband within the perceptual filterbank is then used to determine the corresponding attenuation factor, which provides a trade-off between speech distortion and residual noise.

Considering the classifier design issues, modern speaker recognition systems apply statistical hidden Markov models (HMMs) and Gaussian mixture models (GMMs) [14]. Widespread uses of HMMs and GMMs for speaker modeling arise from the efficient parameter estimation procedures that involve maximizing the likelihood of the model data. However, since a maximum likelihood (ML) derived decision-surface is not optimal, discriminative approaches are a key ingredient for creating robust, more accurate models [15].

Support vector machine is a discriminative approach which has recently attracted significant attention because it discriminates between classes and can efficiently train nonlinear decision boundaries. Due to these advantages, we developed an SVM classifier based on a probabilistic score decided by the ratio of the distance between the test vector and the optimal hyperplane to the margin.

The rest of this paper is organized as follows: Section II describes the front-end speech enhancement in which an SNR-sensitive subspace-based enhancement technique is introduced. In Section III, a short overview of the support vector machine is given. Then, speaker identification and verification algorithms based on probabilistic SVMs are presented. Section IV presents the experimental results of the proposed system and Section V gives the conclusion and final remarks.

II. Front-End Speech Enhancement

The block diagram of our speech enhancement algorithm is given in Figure 2. The input noisy speech is first divided into critical band time series by the wavelet analysis filterbank. The subspace-based enhancement is performed in each critical band. The gain adaptation for estimating the clean speech is based on the prior SNR. The wavelet synthesis filterbank is applied to the gain-modified vector of critical band signal to reconstruct the enhanced full-band speech. The following subsections will describe more details of the speech enhancement method.

A. Subspace-Based Speech Enhancement

The speech enhancement issue can be considered as a clean speech signal $\bar{x}$ being corrupted by an additive noise $\bar{n}$. The resulting noisy speech signal $\bar{y}$ is expressed as

$$\bar{y} = \bar{x} + \bar{n}, \tag{1}$$

where $\bar{x} = [x_1, x_2, \ldots, x_M]^{\mathrm{H}}$, $\bar{n} = [n_1, n_2, \ldots, n_M]^{\mathrm{H}}$, and $\bar{y} = [y_1, y_2, \ldots, y_M]^{\mathrm{H}}$. The observation period has been denoted as M. Henceforth, the vectors $\bar{x}, \bar{n}, \bar{y}$ will be considered as part of complex space $\mathbf{C}^M$.

The clean speech is assumed to be confined into a subspace of dimensionality $K(K < M)$, and $\mathbf{C}^M$ can be decomposed into a signal subspace and a noise subspace. Ephraim and Van Trees [11] realized this partitioning by postulating a linear model for the speech frame under analysis. The range and the null spaces were characterized as the signal and noise subspaces, respectively. The linear model for the clean speech assumes that every M-sample frame can be represented using the model:

$$\bar{x} = \boldsymbol{V}\,\bar{s} = \sum_{i=1}^{K} s_i v_i,\ K \le M, \tag{2}$$

where $\bar{s} = [s_1, s_2, \ldots, s_K]^{\mathrm{H}}$ is a sequence of zero mean complex random variables. $\boldsymbol{V} \in R^{M \times K}$ is known as the model matrix. Assuming that the columns of $\boldsymbol{V}$ are linearly independent, and then the rank of $\boldsymbol{V}$ is K. The range of $\boldsymbol{V}$ defines the

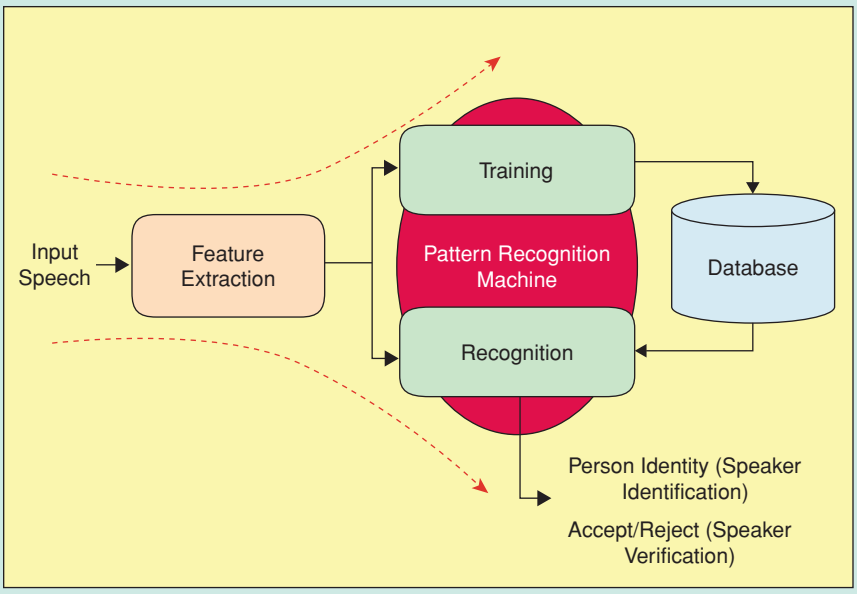

FIGURE 1 Block diagram of speaker identification/verification system.

Various human characteristics such as the face, speech, fingerprint, and iris have been considered as discriminative features for automatic biometric recognition.

signal subspace. The noise subspace is the null space of the model matrix. This subspace has rank $M - K$ and only contains vectors resulting from the noise process.

The subspace decomposition can be achieved using Karhunen-Loeve transform (KLT), i.e. eigenvector matrix. Let $\boldsymbol{R}_x$ and $\boldsymbol{R}_y$ denote the covariance matrix of the $\bar{x}$ and $\bar{y}$, respectively. The eigen-decomposition is performed on the covariance matrix $\boldsymbol{R}_x$ and the following form is obtained

$$\boldsymbol{R}_x = [\boldsymbol{U}_1 \boldsymbol{U}_2] \begin{bmatrix} \boldsymbol{\Lambda}_{x1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_1^H \\ \boldsymbol{U}_2^H \end{bmatrix}, \tag{3}$$

where $\boldsymbol{\Lambda}_{x1}$ is a $K \times K$ diagonal matrix with eigenvalues $\lambda_x(1), \lambda_x(2), \ldots, \lambda_x(K)$ as diagonal elements, i.e., $\text{diag}(\lambda_x(1), \lambda_x(2), \ldots, \lambda_x(K))$. The eigenvector matrix $\boldsymbol{U}$ has been partitioned into two sub-matrices, $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$. The matrix $\boldsymbol{U}_1$ contains eigenvectors corresponding to non-zero eigenvalues. These eigenvectors form a basis for the signal subspace. Meanwhile, $\boldsymbol{U}_2$ contains the eigenvectors which span the noise subspace.

Let $\boldsymbol{\Lambda}_{y1}$ denote $\text{diag}(\lambda_y(1)\lambda_y(2), \ldots, \lambda_y(K))$, and $\boldsymbol{\Lambda}_{y2}$ represent $\text{diag}(\lambda_y(K+1), \lambda_y(K+2), \ldots, \lambda_y(M))$. The notations $\boldsymbol{\Lambda}_{n1}$ and $\boldsymbol{\Lambda}_{n2}$ are in the same fashion. Similar to (3), the eigen-decomposition of $\boldsymbol{R}_y$ is given by

$$\begin{aligned} \boldsymbol{R}_y &= [\boldsymbol{U}_1 \boldsymbol{U}_2] \begin{bmatrix} \boldsymbol{\Lambda}_{y1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda}_{y2} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_1^H \\ \boldsymbol{U}_2^H \end{bmatrix} \\ &= [\boldsymbol{U}_1 \boldsymbol{U}_2] \begin{bmatrix} \boldsymbol{\Lambda}_{x1} + \boldsymbol{\Lambda}_{n1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda}_{n2} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_1^H \\ \boldsymbol{U}_2^H \end{bmatrix}. \end{aligned} \tag{4}$$

As indicated by (4), the clean speech lies only within the signal subspace while the noise spans the entire space. Therefore, only the contents of the signal subspace are used to estimate the clean speech signal.

The clean speech can be estimated using a linear estimator

$$\hat{x} = \boldsymbol{H}\bar{y}, \tag{5}$$

which $\boldsymbol{H}$ is a $K \times K$ matrix. The residual signal, $\bar{e}$, can then be represented as

$$\bar{e} = \hat{x} - \bar{x} = (\boldsymbol{H} - \boldsymbol{I})\bar{x} + \boldsymbol{H}\bar{n} = \bar{e}_x + \bar{e}_n, \tag{6}$$

where $\bar{e}_x$ refers to the signal distortion while $\bar{e}_n$ denotes the residual noise. The energy of the total error, ε thus can be calculated as

$$\varepsilon^2 = \varepsilon_x^2 + \varepsilon_n^2, \tag{7}$$

where $\varepsilon_x^2 = \text{tr}E\{\bar{e}_x \bar{e}_x^H\}$ and $\varepsilon_n^2 = \text{tr}E\{\bar{e}_n \bar{e}_n^H\}$ are the energies of the signal distortion and the residual noise, respectively.

The optimization target is to minimize the signal distortion energy while constraining the average residual noise power to be less than a positive constant, α. Thus

$$\boldsymbol{H}_{opt} = \arg\min_{\boldsymbol{H}} \varepsilon_x^2, \quad \text{subject to: } \varepsilon_n^2 \le \alpha\,. \tag{8}$$

The resulting optimal linear estimator from the time domain constraints has the form

$$\boldsymbol{H}_{opt} = \boldsymbol{R}_x(\boldsymbol{R}_x + \gamma \boldsymbol{R}_n)^{-1}, \tag{9}$$

where γ is the Lagrange multiplier.

Based on the eigen-decomposition of $\boldsymbol{R}_x$, (9) is rewritten as

$$\boldsymbol{H}_{opt} = \boldsymbol{U}\boldsymbol{\Lambda}_x(\boldsymbol{\Lambda}_x + \gamma\ \boldsymbol{U}^H \boldsymbol{R}_n \boldsymbol{U})^{-1}\boldsymbol{U}^H. \tag{10}$$

The $\boldsymbol{U}^H \boldsymbol{R}_n \boldsymbol{U}$ can be approximated by a diagonal matrix $\boldsymbol{\Lambda}_n$ [16], we thus have an approximated linear estimator

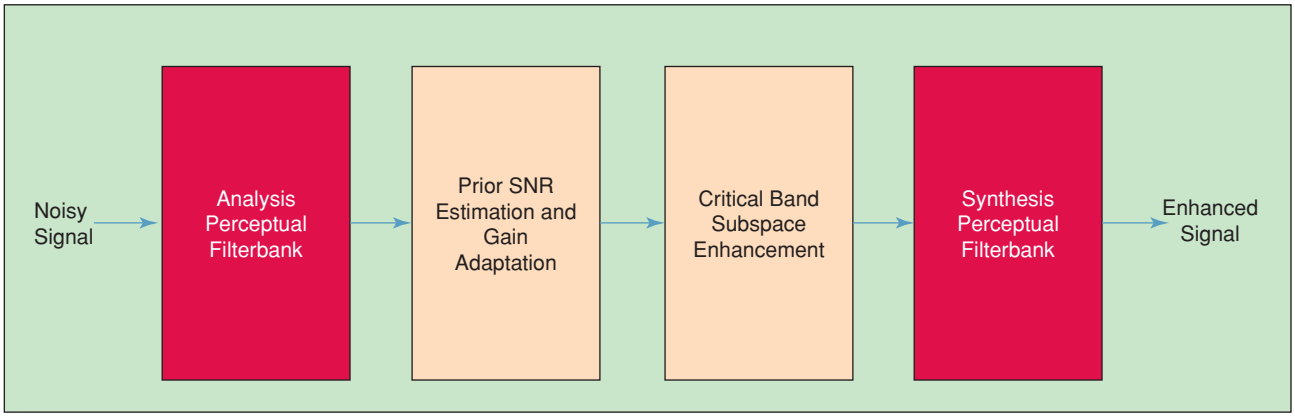

FIGURE 2 Block diagram of our speech enhancement algorithm.

Many researchers have proposed to apply the wavelet analysis to obtain a better time-frequency characteristics for non-stationary signals such as human speech.

$$\tilde{\boldsymbol{H}}_{opt} = \boldsymbol{U}\boldsymbol{\Lambda}_x(\boldsymbol{\Lambda}_x + \gamma\boldsymbol{\Lambda}_n)^{-1}\boldsymbol{U}^H. \tag{11}$$

Removing the noise subspace, we can rewrite the estimator as

$$\tilde{\boldsymbol{H}}_{opt} = \boldsymbol{U}\begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{U}^H, \tag{12}$$

where

$$\boldsymbol{G} = \boldsymbol{\Lambda}_{x1}(\boldsymbol{\Lambda}_{x1} + \gamma\boldsymbol{\Lambda}_{n1})^{-1}. \tag{13}$$

Hence, the signal estimate $\hat{x} = \tilde{\boldsymbol{H}}_{opt}\bar{y}$ is obtained by applying the KLT to the noisy signal, appropriately modifying the components of the KLT $\boldsymbol{U}^H\bar{y}$ by a gain function, and by the inverse KLT of the modified components.

B. Perceptual Wavelet Filterbank

Many researchers have proposed to apply the wavelet analysis to obtain a better time-frequency characteristics for non-stationary signals such as human speech [12], [17]. Compared to a Fourier sinusoid, which oscillates forever, a wavelet is localized in time and lasts for only a few cycles. The wavelet transforms can be implemented through a two-channel filterbanks comprising a lowpass filter $h(n)$ and a highpass filter $g(n)$. Then the lowpass and highpass filter outputs are downsampled by two, which removes the odd-numbered components after filtering, respectively. This processing is called analysis filterbank and is depicted in Figure 3(a). Figure 3(b) illustrates the synthesis filterbank structure. This processing is evaluated by up-sampling the j scaling coefficient sequence a_j, and then convoluting it with the scaling function coefficients $h(n)$. The same process is done to the j wavelet coefficient sequence d_j and the results are added to give the $j+1$ level scaling coefficients a_{j+1}.

The splitting, filtering, and decimation shown in Figure 3(a) can be repeated on the scaling coefficients to give the idea of pyramid-structured wavelet transform. Furthermore, this pyramid-structured wavelet transform can be adjusted to approximate the human auditory system [13]. Such a design is called a perceptual filterbank. The Bark scale is the critical band scales simulating the human auditory system [18]. The Bark scale z can be approximately expressed in terms of the linear frequency by

$$\begin{aligned} z(f) = {} & 13\arctan(7.6\times10^{-4}f) \\ & + 3.5\arctan(1.33\times10^{-4}f)^2, \end{aligned} \tag{14}$$

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

$$\mathrm{CBW}(f_c) = 25 + 75\left(1 + 1.4\times10^{-6}f_c^2\right)^{0.69}, \tag{15}$$

where f_c is the center frequency in Hertz. Theoretically, the range of human auditory frequency spreads from 20 to 20,000 Hz and covers approximately 25 Barks. However, the human speech mostly spans within 4 kHz and there are only 17 critical bands existed in this bandwidth. To approach the 17 critical bands, the tree structure of the perceptual wavelet packet transform can be constructed as shown in Figure 4. It contains 16 decomposition cells with 5 decomposition stages.

C. Gain Estimation in Each Critical Band

The perceptual wavelet filterbank is integrated with the subspace-based enhancement technique. We apply individual subspace analysis in each critical band. The optimal linear estimator for i-th critical band thus has the following form

$$\tilde{\boldsymbol{H}}^i_{opt} = \boldsymbol{U}^i\begin{bmatrix} \boldsymbol{G}^i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}(\boldsymbol{U}^H)^i, \tag{16}$$

where $\boldsymbol{G}^i = \boldsymbol{\Lambda}^i_{x1}(\Lambda^i_{x1} + \gamma^i\boldsymbol{\Lambda}^i_{n1})^{-1}$ is a diagonal gain matrix for i-th critical band.

The gains within the same critical band are assumed to be equal. All the elements of $\boldsymbol{\Lambda}^i_{x1}$ and $\boldsymbol{\Lambda}^i_{n1}$ are respectively summed to get the signal power P^i_x and noise power P^i_n of i-th critical band. Accordingly, the gain for i-th critical band can be expressed by

$$G^i = \frac{P^i_x}{P^i_x + \gamma^i P^i_n} = \frac{P^i_x\left(P^i_n\right)^{-1}}{P^i_x\left(P^i_n\right)^{-1} + \gamma^i}, \tag{17}$$

where γ^i is the attenuation factor for i-th critical band.

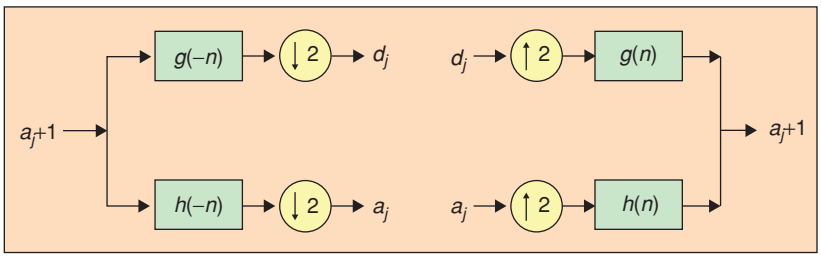

FIGURE 3 (a) Two-channel analysis filterbank; (b) two-channel synthesis filterbank.

Support vector machine is a discriminative approach which has recently attracted significant attention because it discriminates between classes and can efficiently train nonlinear decision boundaries.

The attenuation degree of each critical band is determined independently. In (17), the $P_x^i(P_n^i)^{-1}$ can be considered as the prior SNR in i-th critical band. The prior SNR provide a good clue for better trade off between speech distortion and residual noise [19]. It is calculated in accordance with the noise power spectrum estimated by a pre-obtained noise segment and the speech power spectrum derived by subtracting noise power spectrum from noisy speech power spectrum. The attenuation factor in each critical band is determined according to the prior SNR of the corresponding one. Assume the maximum attenuation value is κ. The attenuation factor of i-th critical band is decided by a monotonic decreasing function

$$\gamma^i = \frac{\kappa e^{-\chi^i}}{1 + e^{-S\chi^i}}, \tag{18}$$

where χ^i is the prior SNR of i-th critical band.

III. Speaker Identification and Verification

A. Support Vector Machines

The Support Vector Machine (SVM) theory is a new statistical technique and has drawn much attention in recent years. An SVM is a binary classifier that makes its decisions by constructing an optimal hyperplane that separates the two classes with the largest margin. It is based on the idea of the structural risk minimization (SRM) induction principle [20], which aims at minimizing a bound on the generalization error, rather than minimizing the mean square error. For the optimal hyperplane $\bar{\mathbf{w}} \cdot \bar{\mathbf{x}} + b = 0$, $\bar{\mathbf{w}} \in R^N$ and $b \in R$, the decision function of classifying a unknown point $\bar{\mathbf{x}}$ is defined as:

$$f(\bar{\mathbf{x}}) = \text{sign}(\overline{\mathbf{wx}} + b) = \text{sign}\left(\sum_{i=1}^{N_S} \alpha_i m_i \bar{\mathbf{x}}_i \cdot \bar{\mathbf{x}}\right), \tag{19}$$

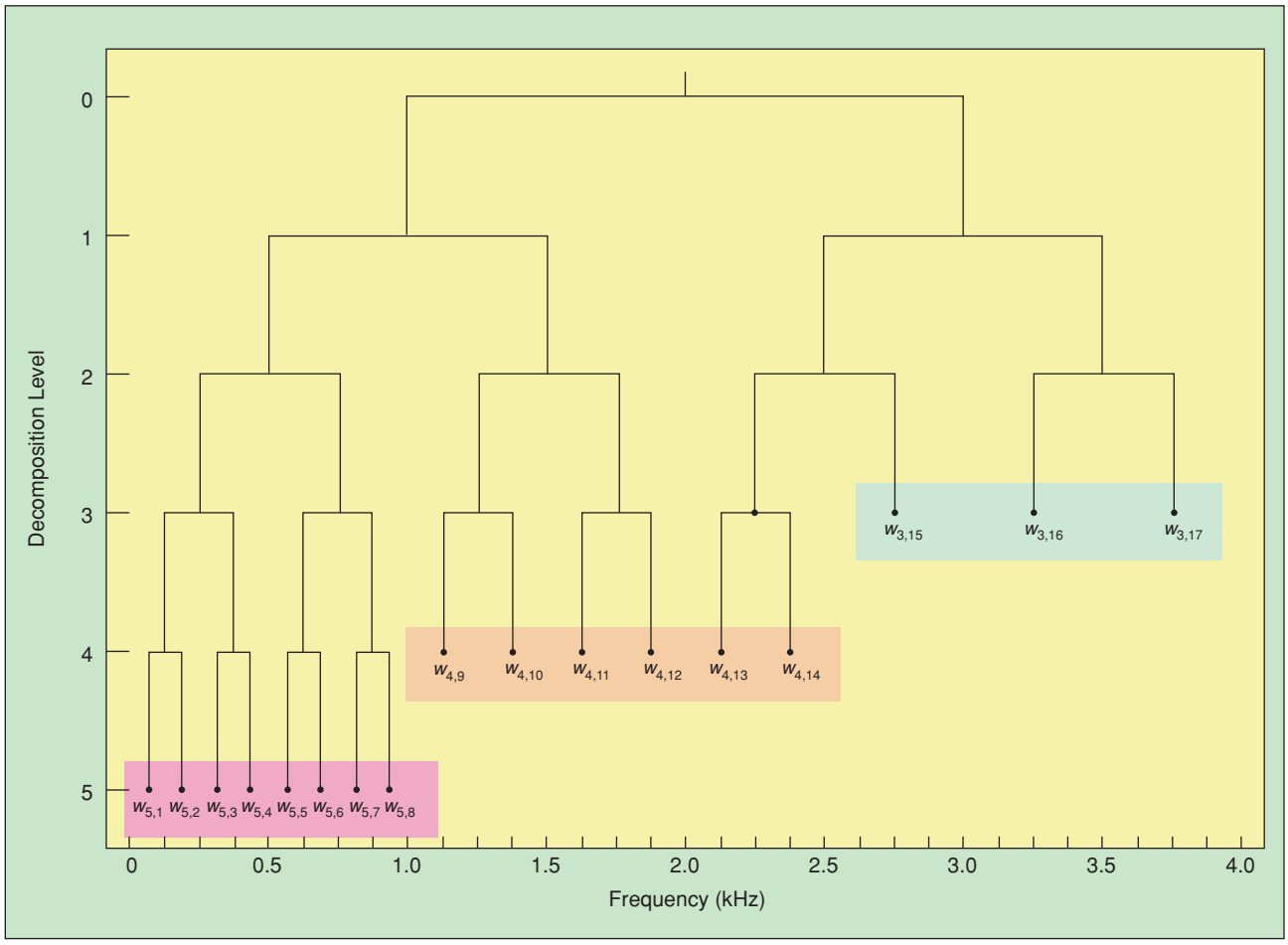

FIGURE 4 Tree structure of the perceptual filterbank.

where N_S is the support vector number, $\bar{\mathbf{x}}_i$ is the support vector, α_i is the Lagrange multiplier and $m_i \in \{-1, +1\}$ describes which class $\bar{\mathbf{x}}$ belongs to.

In most cases, searching suitable hyperplane in input space is too restrictive to be of practical use. The solution to this situation is mapping the input space into a higher dimension

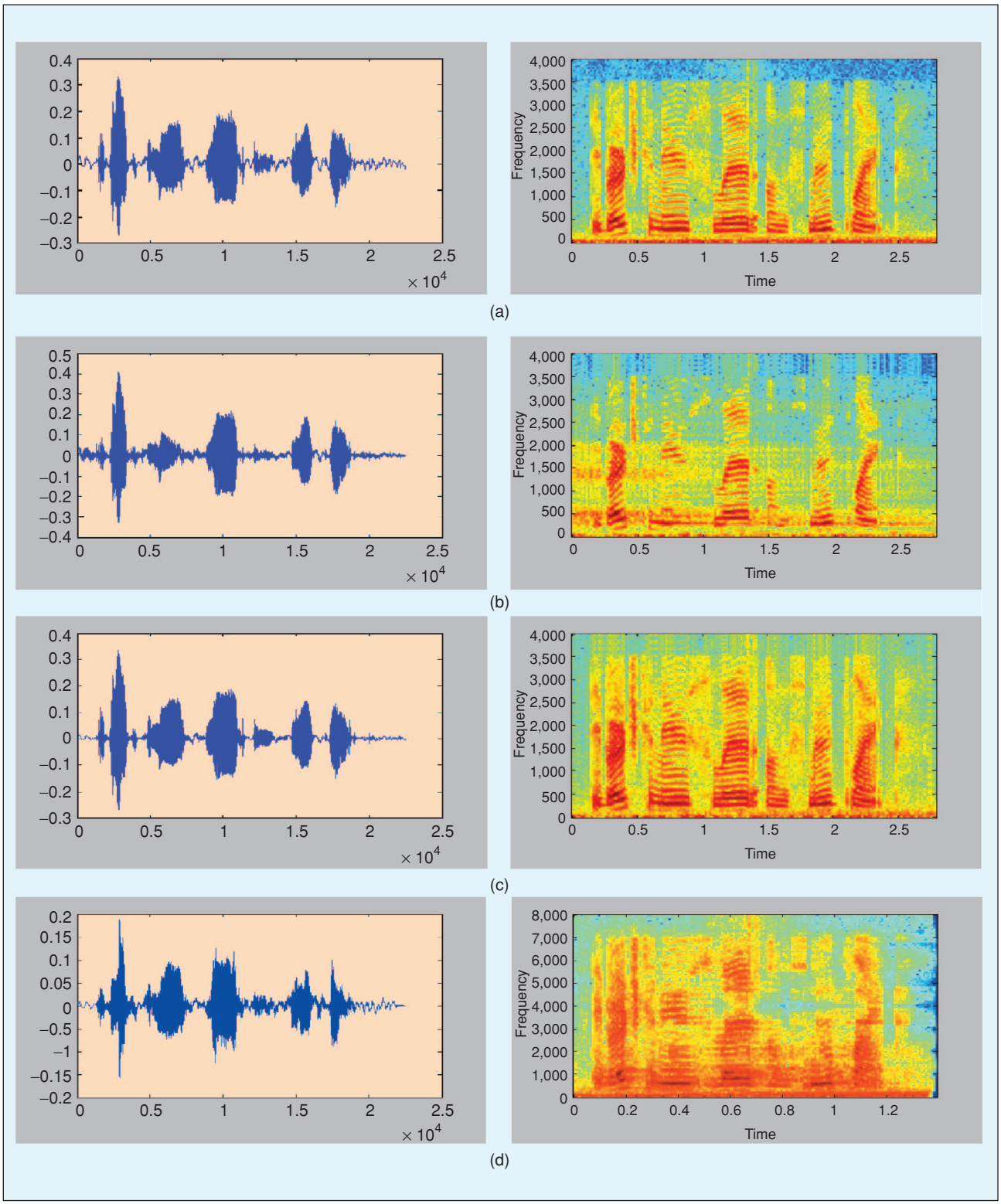

FIGURE 5 The waveform and spectrogram of: (a) noisy speech signal; (b) enhanced speech using spectral subtraction method; (c) enhanced speech using conventional subspace method; (d) enhanced speech using the proposed method.

The Support Vector Machine (SVM) theory is based on the idea of the structural risk minimization induction principle, which aims at minimizing a bound on the generalization error, rather than minimizing the mean square error.

feature space and searching the optimal hyperplane in this feature space. Let $\bar{\mathbf{z}} = \varphi(\bar{\mathbf{x}})$ denote the corresponding feature space vector with a mapping φ from R^N to a feature space Z. It is not necessary to know about φ. We just provide a function $K(*, *)$ called kernel which uses the points in input space to compute the dot product in feature space Z, that is

$$\bar{\mathbf{z}}_i \cdot \bar{\mathbf{z}}_j = \varphi(\bar{\mathbf{x}}_i) \cdot \varphi)\bar{\mathbf{x}}_j) = K(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j). \tag{20}$$

Finally, the decision function becomes

$$f(\bar{\mathbf{x}}) = \text{sign}\left(\sum_{i=1}^{N_S} \alpha_i m_i K(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}) + b\right). \tag{21}$$

Functions that satisfy Mercer's theorem [21] can be used as kernels. Typical kernel functions include linear kernel, polynomial and radial basis kernel, etc.

B. Speaker Identification and Verification Using Probabilistic SVMs

This subsection discusses our method to identify or verify a speaker. For a test utterance, this utterance is enhanced by our subspace-based speech enhancement first. Passing through the procedure of feature extraction, each frame will be transformed into a feature vector. After sending these feature vectors to the SVM classifier, each frame will be given a probabilistic score. Finally, we calculate the sum of all probabilistic scores for two-class classification.

In the speaker identification application, the recognition process starts from a two-class SVM classifier. The following describes how to compute the probabilistic scores. Assume a N_F-frame utterance is to be classified into speaker class C_m, $m \in \{-1, +1\}$ and $\bar{\mathbf{x}}_j$, $j = 1, \ldots, N_F$ is the corresponding feature vector. For speaker class $C_{m=1}$, the distance ratio of the distance between $\bar{\mathbf{x}}_j$ and optimal hyperplane to the margin distance is defined by

$$R\left(\bar{\mathbf{x}}^{(j)}\right) = \frac{\overline{\mathbf{w}\mathbf{x}}^{(f)} + b}{\|\bar{\mathbf{w}}\|} \Big/ \frac{1}{\|\bar{\mathbf{w}}\|} = \overline{\mathbf{w}\mathbf{x}}^{j} + b. \tag{22}$$

We then convert the distance ratio to a value between 0 and +1 through a sigmoid function [15]

$$score_{Frame}\left(C_{m=1}|\bar{\mathbf{x}}^{(j)}\right) = \frac{1}{1 + e^{-R(\bar{\mathbf{x}}^{(j)})}}. \tag{23}$$

This score denotes a kind of possibility that $\bar{\mathbf{x}}_j$ is belonged to $C_{m=1}$.

With $score_{Frame}(C_{m=1}\,|\bar{\mathbf{x}}^{(j)})$, the score describing that $\bar{\mathbf{x}}_j$ is belonged to $C_{m=-1}$. can be computed by

$$score_{Frame}\left(C_{m=-1}|\bar{\mathbf{x}}^{(j)}\right) = 1 - score_{Frame}\left(C_{m=1}|\bar{\mathbf{x}}^{(j)}\right). \tag{24}$$

Therefore, the total score for the N_F-frame testing speech utterance is given by

$$score_{Total}(C_m|\bar{\mathbf{x}}) = \sum_{j=1}^{N_F} score_{Frame}(C_m|\bar{\mathbf{x}}_j), \quad m \in \{-1, +1\}. \tag{25}$$

The two-class classification for choosing m as -1 or 1 is achieved by selecting a bigger one between $score_{Total}(C_{m=1}\,|\bar{\mathbf{x}})$ and $score_{Total}(C_{m=-1}\,|\bar{\mathbf{x}})$.

A multi-class classification system can be obtained from the two-class SVM classifier. Assume there are M speaker classes, each pair of the classes are used to train a SVM classifier, i.e. there are totally $M(M-1)/2$ SVM models. For a test utterance, the pairwise comparison [22] strategy is adopted to identify its speaker.

As for the speaker verification application, our method is also based on a 2-class SVM classifier. The claimed speaker and imposter speaker are the two speaker classes. The training data for imposter speaker class are collected by speeches uttered from numerous speakers. For an utterance with N_F frames, the probabilistic score of the claimed speaker class is obtained by (25). Dividing the core by N_F, we then get a normalized

TABLE 1 Performance comparison in average SegSNR (dB) for sentences corrupted by different noise lavels.

NOISE SOURCE	SPECTRAL SUBTRACTION	CONVENTIONAL SUBSPACE	PROPOSED METHOD
0 DB	−1.6048	0.7897	0.5027
5 DB	0.4765	3.8752	4.3012
10 DB	2.6337	7.2065	8.4046

TABLE 2 Performance evaluation of the proposed speaker identification/verification system.

TESTING SPEECH TYPE	SYSTEM	IDENTIFICATION (CORRECT RATE)	VERIFICATION (EQUAL ERROR RATE)
CLEAN SPEECH	SVMS	90.1%	9.3%
NOISY SPEECH	SVMS	20.3%	43.4%
NOISY SPEECH	ENHANCEMENT PLUS SVMS	48.6%	25.0%

probabilistic score to compare to an empirical threshold in order to decide acceptance or rejection.

In most cases, searching suitable hyperplane in input space is too restrictive to be of practical use. The solution is mapping the input space into a higher dimension feature space and searching the optimal hyperplane in this feature space.

IV. Experimental Results

In this section, an analysis of the performance of the proposed method is presented. The experiment was performed using natural speeches corrupted by additive noises. First, the front-end speech enhancement is evaluated. For comparative purposes, we also implemented and evaluated the spectral subtraction method of Berouti et al. [23] and the conventional subspace method. Figure 5 shows the waveforms and spectrograms of degraded speech and enhanced speech processed by three algorithms: (i) spectral subtraction; (ii) conventional subspace method; and (iii) the proposed method.

For objective evaluation, the segmental SNR (SegSNR) measure was used to evaluate these speech enhancement algorithms. Various in-car noises measured from different cars in TAICAR database [24] were adopted. The performance comparison using SegSNR is given in Table 1. The proposed method significantly outperforms the spectral subtraction and conventional subspace methods. The average improvements are 3.9011 dB and 0.4457 dB, respectively. Next, to evaluate the performance of the proposed robust speaker recognition system, twenty speakers, ten males and ten females, chosen from the AURORA-2 database [25] were used in our experiments. For each speaker, their clean utterances were used for training the SVMs in a clean environment with 13-dimension MFCCs [26] as one feature vector. Another twelve utterances from each speaker were first degraded by background noise and then individually used for testing the system performance. The radial basis kernel $K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \cdot \|\mathbf{x} - \mathbf{y}\|^2)$ with $\gamma = 1$ was adopted for all the experiments. The analysis frame used in this study had 256 samples, which was approximately 32 ms in length. The experimental results of the proposed robust speaker identification/verification system are listed in Table 2. The SVM classifier can achieve about 90.1% correct rate and 9.3% equal error rate [8] in identification and verification, respectively. However, the speaker recognition performance drastically reduces while background noise is introduced. With the front-end enhancement, the background noise problem is lessened and the recognition rate is highly improved from 20.3% to 48.6% in speaker identification. As for speaker verification, the equal error rate is also reduced from 43.4% to 25%.

V. Conclusions

This article has designed a robust speaker recognition system for identifying or verifying an unknown speaker. Our speaker recognition model is based on SVMs. The distance ratios are used to generate the probabilistic scores of SVMs. To alleviate environment noise problem, an SNR-sensitive subspace-based enhancement technique is developed. In our experimental results, the proposed SNR-sensitive subspace-based enhancement significantly outperforms the conventional subspace-based and spectral subtraction methods in terms of SNR. With this enhancement as a front-end process, the performance of our speaker recognition system under noisy environment is also notably improved.

References

[1] B.H. Juang and T.H. Chen, "The past, present, and future of speech processing," *IEEE Signal Processing Magazine*, vol. 15, no. 3, pp. 24–48, May 1998.

[2] T.G. Clarkson, C.C. Christodoulou, Y. Guan, D. Gorse, D.A. Romano-Critchley, and J.G. Taylor, "Speaker identification for security systems using reinforcement-trained pRAM neural network architectures," *IEEE Transactions on Systems, Man and Cybernetics, Part C*, vol. 31, no. 1, pp. 65–76, Feb. 2001.

[3] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[4] B.L. Pellom and J.H.L. Hansen, "An efficient scoring algorithm for Gaussian mixture model based speaker identification," *IEEE Signal Processing Letters*, vol. 5, no. 11, pp. 281–284, Nov. 1998.

[5] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

[6] K.T. Assaleh and R.J. Mammone, "New LP-derived features for speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 630–638, Oct. 1994.

[7] Q. Li, "A detection approach to search-space reduction for HMM state alignment in speaker verification," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 569–578, July 2001.

[8] H. Jiang and L. Deng, "A Bayesian approach to the verification problem: applications to speaker verification," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 874–884, Nov. 2001.

[9] B. Xiang and T. Berger, "Efficient text-independent speaker verification with structural Gaussian mixture models and neural network," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 447–456, Sept. 2003.

[10] M.W. Mak and S.Y. Kung, "Estimation of elliptical basis function parameters by the EM algorithm with application to speaker verification," *IEEE Transactions on Neural Networks*, vol. 11, no. 4, pp. 961–969, July 2000.

[11] Y. Ephraim and H.L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, July 1995.

[12] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1998.

[13] S.H. Chen and J.F. Wang, "Speech enhancement using perceptual wavelet packet decomposition and Teager energy operator," *Journal of VLSI Signal Processing*, vol. 36, no. 2–3, pp. 125–139, Feb. 2004.

[14] G.R. Doddington, M.A. Przybocki, A.F. Martin, and D.A. Reynolds, "The NIST speaker recognition evaluation—overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2–3, pp. 225–254, June 2000.

[15] A. Ganapathiraju, J.E. Hamaker, and J. Picone, "Applications of support vector machines to speech recognition," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2348–2355. Aug. 2004.

[16] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 87–95, Feb. 2001.

[17] I. Daubechies, *Ten Lectures on Wavelets*, CBMS, SIAM publ., 1992.

[18] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 819–829, June 1992.

[19] Y. Hu and P.C. Loizou, "A perceptually motivated approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no 5, pp. 457–465, Sept. 2003.

[20] V. Vapnik, *Statistical Learning Theory*, New York: Wiley, 1998.

[21] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Interscience Publishers, 1953.

[22] U. Kressel, "Pairwise classification and support vector machines", in *Advances in Kernel Methods—Support Vector Learning*, B. Scholkopf, C. Burges, and A. J. Smola, (Eds) MIT Press, Cambridge, Massachusetts, chapter 15, 1999.

[23] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1979, pp. 208–211.

[24] H.C. Wang, C.H. Yang, J.F. Wang, C.H. Wu, and J.T. Chien "TAICAR-the collection and annotation of an in-car speech database created in Taiwan," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 10, no. 2, pp. 237–250, June 2005.

[25] http://www.icp.inpg.fr/ELRA/home.html, the ELRA home page.

[26] J.C. Wang, J.F. Wang, and Y.S. Weng, "Chip design of MFCC extraction for speech recognition," *Integration, the VLSI Journal*, vol. 32, no. 1–3, pp. 111–131, Nov. 2002.