Robust Speaker Recognition Using SNR-Aware Subspace-Based Enhancement and Probabilistic SVMs

Jia-Ching Wang, Jhing-Fa Wang, Wai-He Kuok, Hsiao-Ping Lee, and Chung-Hsien Yang
Department of Electrical Engineering, National Cheng Kung University
1 University Road, Tainan, Taiwan, R.O.C.
Tel: 886-6-2757575 ext. 62341 Fax: 886-6-2080478

ABSTRACT

In this paper, we present a robust text-independent speaker recognition system. The proposed system mainly includes an SNR-aware subspace-based enhancement technique and probabilistic support vector machines (SVMs). First, we construct a perceptual filterbank from psycho-acoustic model and incorporate it with the subspace-based enhancement approach. The prior SNR of each subband within the perceptual filterbank is taken to decide the estimator's gain to effectively suppress environmental background noises. Next, this study uses probabilistic SVMs to identify the speaker from the enhanced speech. The superiority of the proposed system has been demonstrated by twenty speaker recognition from AURORA-2 database with in-car noises.

1. Introduction

A speaker recognition system attempts to identify a speaker in accordance with his speech utterances. However, the performance of a speaker recognition system is usually drastically degraded in a noisy environment. For lessening environment noise problem, this study presents a robust speaker recognition architecture, which comprises an SNR-aware subspace-based enhancement technique and probabilistic support vector machines (SVMs).

Among several strategies emerged for lessening environment noise problem, front-end enhancement undoubtedly is a powerful one. Ephraim and Van Trees [1] proposed a subspace-based speech enhancement which it seeks for an optimal estimator that would minimize the speech distortion subject to the constraint that the residual noise fell below a preset threshold. In this paper, an SNR-aware subspace-based enhancement technique is presented. First, 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. The prior SNR of each subband within the perceptual filterbank is used to determine the corresponding attenuation factor, which provides the trade-off between speech distortion and residual noise.

Considering the classifier design issue, modern speaker recognition systems applied statistical hidden Markov models (HMMs) and Gausian mixture models (GMMs) [2]. 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, as a maximum likelihood (ML) derived decision surface is not optimal, the discriminative approaches are a key ingredient for creating robust and more accurate models [3]. Support vector machine, a discriminative approach, attracts significant attention recently because they discriminate between classes and can be used to train nonlinear decision boundaries in an efficient manner. This motivates us to present an SVM-based speaker recognition system. The proposed SVM classifier is based on probabilistic score decided by the distance ratio of the distance between test vector and optimal hyperplane to the margin distance.

2. SNR-Aware Subspace-Based Speech Enhancement

2.1. Subspace-Based Speech Enhancement

The speech enhancement problem will be described as a clean speech signal $\bar{x}$ being transmitted through a distortionless channel that is corrupted by additive noise $\bar{n}$. The resulting noisy speech signal $\bar{y}$ can be expressed as

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

where $\bar{x} = [x_1, x_2, \ldots, x_M]^H$, $\bar{n} = [n_1, n_2, \ldots, n_M]^H$, and $\bar{y} = [y_1, y_2, \ldots, y_M]^H$. 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 C^M.

The subspace decomposition can be achieved using 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

$$\boldsymbol{R}_x = [\boldsymbol{U}_1 \boldsymbol{U}_2] \begin{bmatrix} \boldsymbol{\Lambda}_{x1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_1^H \\ \boldsymbol{U}_2^H \end{bmatrix}, \tag{2}$$

where $\boldsymbol{\Lambda}_{x1}$ is a $K \times K$ diagonal matrix with eigenvalues $\lambda_x(1)$, $\lambda_x(2)$, …, $\lambda_x(K)$ as diagonal elements. 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 $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ represent the identity matrices $\boldsymbol{I}_{K\text{x}K}$ and $\boldsymbol{I}_{(M\text{-}K)\text{x}(M\text{-}K)}$, respectively. Similar to (2), the eigen-decomposition of $\boldsymbol{R}_y$ is given by

$$\begin{aligned}\mathbf{R}_y &= \begin{bmatrix}\mathbf{U}_1\mathbf{U}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_{y1} & \mathbf{0}\\ \mathbf{0} & \sigma_n^2\mathbf{I}_2\end{bmatrix}\begin{bmatrix}\mathbf{U}_1^H\\ \mathbf{U}_2^H\end{bmatrix}, \quad (3)\\ &= \begin{bmatrix}\mathbf{U}_1\mathbf{U}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_{x1}+\sigma_n^2\mathbf{I}_1 & \mathbf{0}\\ \mathbf{0} & \sigma_n^2\mathbf{I}_2\end{bmatrix}\begin{bmatrix}\mathbf{U}_1^H\\ \mathbf{U}_2^H\end{bmatrix}\end{aligned}$$

where $\boldsymbol{\Lambda}_{y1}$ is a $K \times K$ diagonal matrix with eigenvalues $\lambda_y(1)$, $\lambda_y(2)$, … , $\lambda_y(K)$ as diagonal elements.

As indicated by (3), 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

$$\hat{x} = \boldsymbol{H}\bar{y}, \quad (4)$$

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

$$\bar{e} = \hat{x} - \bar{x} = (\boldsymbol{H} - \boldsymbol{I})\bar{x} + \boldsymbol{H}\bar{n} = \bar{e}_x + \bar{e}_n, \quad (5)$$

where $\bar{e}_x$ refers to the signal distortion while $\bar{e}_n$ denotes the residual noise. The energy of the signal distortion can be calculated from (5)

$$\varepsilon_x^2 = \text{tr}E\{\bar{e}_x\bar{e}_x^H\} = \text{tr}\{(\boldsymbol{H}-\boldsymbol{I})\boldsymbol{R}_x(\boldsymbol{H}-\boldsymbol{I})^H\}. \quad (6)$$

Similarly, the energy of the residual noise can be derived from (6)

$$\varepsilon_n^2 = \text{tr}E\{\bar{e}_n\bar{e}_n^H\} = \sigma_n^2\text{tr}\{\boldsymbol{H}\boldsymbol{H}^H\}. \quad (7)$$

The energy of the total error, ε thus can be calculated as

$$\varepsilon^2 = \varepsilon_x^2 + \varepsilon_n^2 \quad (8)$$

The time domain constrained estimator minimizes signal distortion while constraining the average residual noise power to be less than $\alpha\sigma_n^2$. Thus

$$\begin{aligned}\boldsymbol{H}_{opt} &= \arg\min_{\boldsymbol{H}} \varepsilon_x^2\\ \text{subject to: } &\frac{1}{M}\varepsilon_n^2 \le \alpha\sigma_n^2 \quad (9)\end{aligned}$$

where $0 \le \alpha \le 1$. The resulting filter from the TDC estimation has the form

$$\mathbf{H}_{opt} = \mathbf{R}_x(\mathbf{R}_x + \beta\sigma_n^2\mathbf{I})^{-1}. \quad (10)$$

Applying the eigen-decomposition (2) of $\boldsymbol{R}_x$ to (10), we can rewrite the optimal linear estimator as

$$\mathbf{H}_{opt} = \mathbf{U}\begin{bmatrix}\mathbf{G}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{U}^H \quad (11)$$

where

$$\mathbf{G}_1 = \boldsymbol{\Lambda}_{x1}(\boldsymbol{\Lambda}_{x1} + \beta\sigma_n^2\mathbf{I}_1)^{-1} \quad (12)$$

Hence, the signal estimate $\hat{x} = \boldsymbol{H}_{opt}\bar{y}$ 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 inverse KLT of the modified components.

2.2. 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. 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. 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

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

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

$$\text{CBW}(f_c) = 25 + 75(1 + 1.4\times10^{-6}f_c^2)^{0.69}, \quad (14)$$

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.

The tree structure of the perceptual wavelet packet transform can be constructed as shown in Fig. 1. 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. The resulting 17-band perceptual wavelet packet transform of the Bark scale and the CBW are plotted in Figs. 2 and 3, respectively.

2.3. Prior Subband SNR-Aware Gain Estimation

The perceptual filterbank is integrated with the subspace-based enhancement technique. For each subband within the perceptual filterbank, individual subspace analysis is applied. Therefore, the optimal linear estimator for i-th subband has the following form

$$\mathbf{H}_{opt}^i = \mathbf{U}^i\begin{bmatrix}\mathbf{G}_1^i & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}(\mathbf{U}^i)^H. \quad (15)$$

where

$$\mathbf{G}_1^i = \boldsymbol{\Lambda}_{x1}^i(\boldsymbol{\Lambda}_{x1}^i + \gamma^i\mathbf{I}_1)^{-1} \quad (16)$$

The j-th diagonal element of (16) can be expressed by

$$\mathbf{G}_1^i(j) = \frac{\lambda_x^i(j)}{\lambda_x^i(j) + \gamma^i} \cdot \quad (17)$$

In (17), γ^i is the attenuation factor for *i*-th subband. This factor corresponds to the term $\beta\sigma_n^2\mathbf{I}$ in (12) and controls the trade-off between speech distortion and residual noise in *i*-th subband. A larger value of the attenuation factor will yield more speech distortion and less residual noise and vice versa. How to decide the attenuation factor thus plays an essential role for the enhancement process.

Instead of applying the same attenuation value to the whole frequency span, deciding the attenuation degree of each subband within the perceptual filterbank respectively is a better solution. The attenuation factor in each subband is determined according to the prior SNR of the corresponding subband.

The prior SNR of *i*-th subabnd 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. Assume the the maximum attenuation value is κ and the prior SNR of *i*-th subabnd is SNR^i, the attenuation factor of *i*-th subband is decided by a monotonic decreasing function

$$\gamma^i = \frac{\kappa e^{-SNR_i}}{1 + e^{-SNR_i}} \cdot \quad (18)$$

3. Speaker Recognition

3.1. Support Vector Machines

The SVM theory is a new statistical technique and has drawn much attention on this topic 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 structural risk minimization (SRM) induction principle [4] that aims at minimizing a bound on the generalization error, rather than minimizing the mean square error. For the optimal hyperplane $\overline{\mathbf{w}} \cdot \overline{\mathbf{x}} + b = 0$, $\overline{\mathbf{w}} \in R^N$ and $b \in R$, the decision function of classifying a unknown point $\overline{\mathbf{x}}$ is defined as:

$$f(\overline{\mathbf{x}}) = \text{sign}(\overline{\mathbf{w}}\overline{\mathbf{x}} + b) = \text{sign}(\sum_{i=1}^{N_S} \alpha_i m_i \overline{\mathbf{x}}_i \cdot \overline{\mathbf{x}}), \quad (19)$$

where N_S is the support vector number, $\overline{\mathbf{x}}_i$ is the support vector, α_i is the Lagrange multiplier and $m_i \in \{-1,+1\}$ describes which class $\overline{\mathbf{x}}$ 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 feature space and searching the optimal hyperplane in this feature space. Let $\overline{\mathbf{z}} = \varphi(\overline{\mathbf{x}})$ 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

$$\overline{\mathbf{z}}_i \cdot \overline{\mathbf{z}}_j = \varphi(\overline{\mathbf{x}}_i) \cdot \varphi(\overline{\mathbf{x}}_j) = K(\overline{\mathbf{x}}_i, \overline{\mathbf{x}}_j) \cdot \quad (20)$$

Finally, the decision function becomes

$$f(\overline{\mathbf{x}}) = \text{sign}(\sum_{i=1}^{N_S} \alpha_i m_i K(\overline{\mathbf{x}}_i, \overline{\mathbf{x}}) + b) \cdot \quad (21)$$

Functions that satisfy Mercer's theorem [5] can be used as kernels. Typical kernel functions include linear kernel, polynomial and radial basis kernel, etc.

3.2. Speaker Recognition Using Probabilistic SVMs

This subsection discusses our method to identify a speaker. For an enhanced utterance from unknown speaker, the waveform is segmented into separate frames first. The proposed method starts from a 2-class SVM classifier. For a test utterance, this utterance is enhanced by our SNR-aware 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. If the sum of all probabilistic scores in an utterance is greater than zero, this utterance is classified to +1 speaker class. Otherwise, it is classified to -1 speaker class.

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 $\overline{\mathbf{x}}_j$, j = 1, …, N_F is the corresponding feature vector. For speaker class C_m, the distance ratio of the distance between $\overline{\mathbf{x}}_j$ and optimal hyperplane to the margin distance is defined by

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

This study then converts the distance ratio to a value between 0 and +1 through a sigmoid function

$$score_{SVM}(C_m \left| \overline{\mathbf{x}}^{(j)} \right.) = \frac{1}{1 + e^{-R(\overline{\mathbf{x}}^{(j)})}} \cdot \quad (23)$$

This score denotes a kind of possibility that $\overline{\mathbf{x}}_j$ is belonged to C_m.

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 [6] strategy is adopted to identify its speaker.

4. Experimental Results

First, an objective evaluation with SNR for the proposed SNR-aware subspace-based method was performed. The performance comparison with conventional subspace-based method and spectral subtraction method is listed in Table I. The proposed method significantly outperforms the spectral subtraction and conventional subspace methods. The average improvements are 7.3907 dB and 2.6763 dB, respectively.

To evaluate the performance of the proposed robust speaker recognition system, twenty speakers, ten males and ten females, chosen from the AURORA-2 database 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 as one feature vector. Another four utterances from each speaker were first degraded by in-car noise and then individually used for testing the system performance. 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 recognition system are listed in Table II. This experiment demonstrates the superiority of the proposed system.

5. Conclusions

This study has implemented a robust speaker recognition system. Our speaker recognition model is based on SVMs. This study uses distance ratios to generate the probabilistic scores of SVMs. To alleviate environment noise problem, an SNR-aware subspace-based enhancement technique is presented. In our experimental results, the proposed SNR-aware subspace-based enhancement significantly outperforms the conventional subspace-based and spectral subtraction methods in terms of SNR. With this enhancement as front-end process, the performance of our speaker recognition system under in-car noise environment is also notably improved.

References

[1] 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.

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

[3] 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.

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

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

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

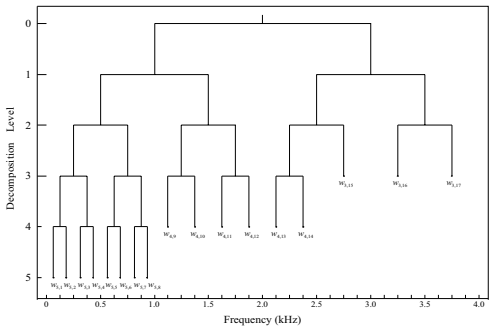

Fig.1. Tree structure of the perceptual filterbank.

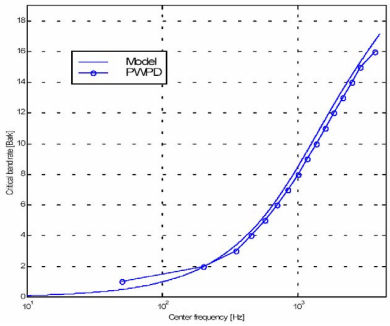

Fig. 2. Bark scale as a function of center frequency.

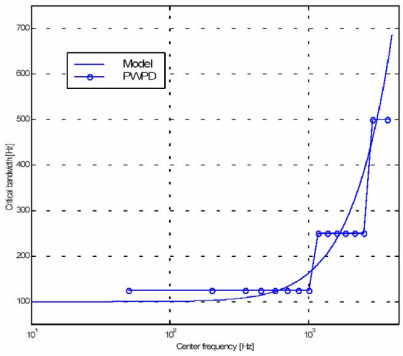

Fig. 3. Critical bandwidth as a function of center frequency.

Table I. Performance comparison in SNR (dB) for utterances corrupted by different in-car noises

Enhancement method	Honda noise	Toyota noise	Opel noise	Ave.
No enhancement	0.3261	-0.5933	-0.2285	-0.1652
Spectral subtraction	4.0408	3.9228	1.1338	3.0344
Conventional subspace	9.0472	5.8111	8.3881	7.7488
Proposed	13.7227	7.8719	9.6808	10.4251

Table II. Performance evaluation of the proposed speaker recognition system

Testing speech type	System	Male	Female	Ave.
Clean speech	SVMs	90%	100%	95%
Noisy speech	SVMs	22.5%	20%	21.25%
Noisy speech	Enhancement plus SVMs	50%	40%	45%