ELSEVIER

Speech Communication 30 (2000) 273–293

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

A robust training algorithm for adverse speech recognition

Wei-Tyng Hong [a,*], Sin-Horng Chen [b]

[a] *E000/CCL, Industrial Technology Research Institute, Chutung, Hsinchu, Taiwan, ROC*
[b] *Department of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan, ROC*

Received 10 November 1998

Abstract

In this paper, a new robust training algorithm is proposed for the generation of a set of bias-removed, noise-suppressed reference speech HMM models in adverse environment suffering from both channel bias and additive noise. Its main idea is to incorporate a signal bias-compensation operation and a PMC noise-compensation operation into its iterative training process. This makes the resulting speech HMM models more suitable to the given robust speech recognition method using the same signal bias-compensation and PMC noise-compensation operations in the recognition process. Experimental results showed that the speech HMM models it generated outperformed both the clean-speech HMM models and those generated by the conventional *k*-means algorithm for two adverse Mandarin speech recognition tasks. So it is a promising robust training algorithm.

Keywords: Robust training algorithm; PMC noise-compensation; Signal bias-compensation; Mandarin speech recognition

1. Introduction

Background noise and channel bias are the two major interference factors that seriously degrade the performances of speech recognizers operating in adverse environments such as telephone speech through public switching network. Recently, IBM built an HMM-based Mandarin telephone speech recognition system using a large telephone speech database called 'Mandarin call home database' (Liu et al., 1996). The vocabulary contained about 44 000 words. The word and syllable error rates were, respectively, 70.5% and 58.7%, which were much worse than those achieved in microphone-speech recognition (Lee and Juang, 1996). In the past, many studies have been devoted to the field of robust speech recognition for adverse environment (Juang, 1991; Furui, 1992; Gong, 1995; Junqua and Haton, 1996). Major efforts of those studies were put on developing robust recognition algorithms to compensate or to eliminate noise/channel effect based on a given set of reference speech models trained usually in clean-speech environment. In the non-linear noise subtraction method (Lockwood and Boudy, 1992; Mokbel and Chollet, 1995), a noise model was first estimated from the non-speech precursor of the testing utterance and then subtracted from the speech part in linear spectrum domain in order to obtain noise-suppressed features to be recognized using the clean-speech reference models. In (Acero and Stern, 1990, 1991), the CDCN

* Corresponding author.
E-mail addresses: jfhong@taiwan.com (W.-T. Hong), schen@cc.nctu.edu.tw (S.-H. Chen).

0167-6393/00/$ - see front matter
PII: S0167-6393(99)00057-6

(codeword-dependent cepstral normalization) algorithm was proposed to estimate equalization vectors for the best transformation, in the maximum likelihood sense, from the universal codebook into the testing acoustic space in order to eliminating both the noise and channel effects. In the RASTA method (Hermansky and Morgan, 1994), a filter was used to eliminate the speaker/channel bias for obtaining bias-removed recognition features. In the parallel-model-combination (PMC) method (Gales and Young, 1996), clean-speech HMM models were combined with the current noise model to form noise-compensated composite HMM models for recognizing noisy speech. In the state-based Wiener filtering method (Hansen and Clements, 1991; Ephraim, 1992; Vaseghi and Milner, 1997), a two-stage recognition method was used. It first used the Viterbi algorithm in the first stage to find the best state sequence for the input testing noisy speech, and then applied state-based Wiener filtering to estimate the clean-speech and recognized it using the clean-speech HMM models in the second stage. In (Zhao, 1996), a two-step procedure was employed to detect a spectral bias vector for the input testing utterance by using Gaussian distributed phone models. It then removed the estimated bias vector from the testing utterance for recognition. In the stochastic matching algorithm (Sankar and Lee 1996; Lee, 1998), the parameters of mapping functions between the testing speech and reference HMM models were estimated iteratively using the expectation maximization (EM) algorithm (Dempster et al., 1977). In (Minami and Furui, 1996), an integrated method for adapting HMM models to additive noise and channel distortion was proposed. This method first estimated the signal-to-noise ratio by maximizing the likelihood of the PMC-compensated HMM models to the input speech, and then estimated the cepstral bias by the Sankar's method (Sankar and Lee, 1996). The procedure is iteratively applied until a convergence is reached.

Apart from the above-mentioned main research stream, the robust training issue is also important for adverse speech recognition when the clean-speech reference models are not available. Its main concern is to train a set of robust reference speech models directly from a database collected in adverse environment for adverse speech recognition. The issue is important because the set of reference speech models obtained by the conventional segmental k-means algorithm (Juang and Rabiner, 1990) is usually not robust. This is mainly owing to the high variability on the characteristics of the training speech signals collected in the adverse environment. For example, a training data set collected from telephone calls through the public switching network will suffer diverse recording conditions caused by different background noises, different types of transducers, different telephone channels, etc. This will make speech patterns distribute more widely in the feature space so as to overlap to each other more seriously and cause the trained speech models degrade on their discrimination capabilities.

In the past, many robust training algorithms have been proposed. In the signal bias removal (SBR) algorithm (Rahim and Juang, 1996), a codebook-based iterative signal bias removing technique was performed on both the training and testing phases for minimizing the channel-induced variations. In (Anastaskos et al., 1997), the speaker-specific characteristics were first modeled by a linear-regressive transformation between the speaker-independent models and the speaker-dependent models. A speaker-adaptive training algorithm designed basing on the EM algorithm was then employed to iteratively estimate the parameters of the transformation and the compact speaker-normalized HMM models. In (Gong, 1997), a source normalization training algorithm, which modeled the environmental corruption as a form of linear transformation, was proposed to estimate the HMM models. The noise and channel effects were modeled implicitly in the linear transformation. In the testing stage, the MLLR adaptation (Gales and Woodland, 1996) was applied to estimate the state-dependent transformation matrices and the bias terms for recognition. Those training algorithms have been shown to be effective on removing the channel biases and/or the speaker variations. However, the noise effect is still seldom considered in the robust training issue.

In this study, we are interested in the robust training issue with both the signal bias and noise effects being considered. A robust training algorithm, referred to as the robust environment-effects suppression training (REST) algorithm, is proposed. The design goal of the REST algorithm is twofold. One is to countervail the large variability of the corrupted training samples for obtaining a set of compact reference

speech HMM models with both signal bias and noise being suppressed. The other is to make the generated compact reference speech HMM models better for a given robust speech recognition method. The REST algorithm is an iterative training procedure that sequentially optimizes the following three operations: parameter estimation for environment characterization, environment-effect compensation for speech segmentation, and environment-effect suppression for HMM model re-estimation. The parameter estimation for environment characterization is to detect the signal bias and to estimate the noise statistics for each training utterance. It assumes that each utterance has its own environmental characteristics. Based on an assumed environment contamination model, the environment-effect compensation uses the estimated environment characterization parameters to adapt the HMM models to match with the current training utterance for optimal segmentation. Using the segmentation results and the same environment contamination model, the environment-effect suppression is to remove the signal bias and the noise out of the corrupted speech for updating the HMM models. Owing to the involvement of the environment-effect compensation operation in the training process of the REST algorithm, we expect that it will generate better reference speech HMM models for the robust recognition method which employs the same environment-effect compensation operation in the recognition process. This is especially true for the case when the environment-effect compensation operation is not perfect due either to the non-existence of a perfect one or to the use of an inaccurate environment contamination model in its derivation.

The organization of the paper is stated as follows. Section 2 presents the proposed REST algorithm in detail. Section 3 describes the robust speech recognition method using the reference speech HMM models generated by the REST algorithm. Effectiveness of the REST algorithm is evaluated by simulations discussed in Section 4. Some conclusions are given in Section 5.

2. The REST algorithm

The proposed REST training algorithm consists of an iterative procedure which sequentially performs the following three steps:

1. optimally segment each training utterance by using the environment-compensated HMM models,
2. estimate the environment characteristics and enhance the speech by eliminating the noise using the state-based Wiener filtering method and by removing the signal bias using the SBR method, and
3. re-estimate the speech HMM models.

Operations performed in these three steps are derived based on a presumed environment contamination model. A schematic diagram of the model is displayed in Fig. 1. It assumes that, for each utterance, the observed speech z is generated from the clean speech x by corrupting first with a convolutional channel b and then with an additive noise n. Here b is assumed to be time-invariant and n is stationary throughout the utterance. In linear spectrum domain, the model can be expressed by

$$y_t(f) = b(f) \times x_t(f), \tag{1a}$$

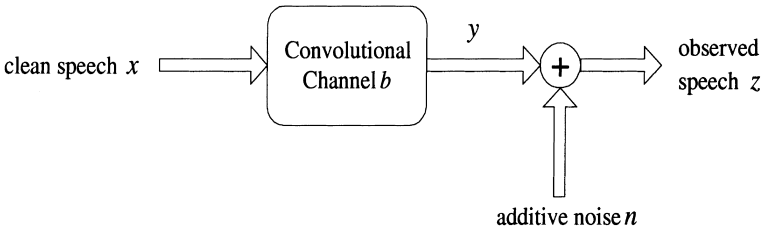

Fig. 1. A schematic diagram of the environment contamination model.

$$z_t(f) = y_t(f) + n_t(f), \tag{1b}$$

where the subscript t denotes the frame index and $y_t(f)$ is an intermediate signal showing the corruption of the clean-speech with the channel bias only. We can also express the relation of x and y in cepstrum domain by

$$y_t(m) = b(m) + x_t(m), \tag{1c}$$

where m denotes the order of cepstral coefficient. Obviously, it is troublesome to directly estimate the original clean speech x in either linear spectrum domain or cepstrum domain when both noise interference and channel distortion exist. We had better, as suggested by above formulations, to separately deal with the channel distortion in cepstrum domain and the noise interference in linear spectrum domain. In the following discussions, we specify a signal in linear spectrum domain and in cepstrum domain by attaching it with parameters f and m, respectively.

The REST algorithm is derived as follows. Assume that the training data set contains R utterances. Let $\Lambda_e \equiv \{\Lambda_n^{(r)}, b^{(r)}\}_{r=1,\ldots,R}$ denote the set of environmental interference models of the whole training data set, where $b^{(r)}$ and $\Lambda_n^{(r)} = \{\mu_n^{(r)}, \Sigma_n^{(r)}\}$ are, respectively, the signal bias and the noise model of the rth training utterance; $\mu_n^{(r)}$ and $\Sigma_n^{(r)}$ are the mean vector and covariance matrix of $\Lambda_n^{(r)}$. Let $Z^{(r)} = (z_1^{(r)}, \ldots, z_{T_r}^{(r)})$ and $X^{(r)} = (x_1^{(r)}, \ldots, x_{T_r}^{(r)})$ be, respectively, the observed and clean-speech feature vector sequence of the rth utterance, and Λ_x denote the set of environment-effect normalized speech HMM models that we want to generate. Based on the maximum likelihood criterion, the goal of an ideal robust training algorithm is to jointly estimate Λ_x and Λ_e with given $\{Z^{(r)}\}_{r=1,\ldots,R}$ by

$$(\Lambda_x^*, \Lambda_e^*) = \arg\max_{(\Lambda_x, \Lambda_e)} L\Big(\{Z^{(r)}\}_{r=1,\ldots,R} \,|\, \Lambda_x, \Lambda_e\Big), \tag{2}$$

where $L(\cdot)$ is the likelihood function of the observation sequence $Z^{(r)}$ given the parameter set of (Λ_x, Λ_e). But, due to the fact that it is generally difficult to derive a close form solution for the above joint maximization problem, we therefore use a three-step iterative training procedure in the REST algorithm to obtain a sub-optimal solution. The three steps are:

1. Form the environment-compensated speech HMM models $\Lambda_z^{(r)}$ by using the current (Λ_x, Λ_e) and use it to optimally segment the training utterance $Z^{(r)}$.
2. Based on the segmentation result, estimate $\Lambda_n^{(r)}$ and enhance the adverse speech $Z^{(r)}$ to obtain $Y^{(r)}$ by the state-based Wiener filtering method; and then, estimate $b^{(r)}$ and further enhance the speech $Y^{(r)}$ to obtain $X^{(r)}$ by the SBR method.
3. Update the current speech HMM models Λ_x using the enhanced speech $\{X^{(r)}\}_{r=1,\ldots,R}$. We discuss these three steps in more detail as follows.

The first step of the REST algorithm is to optimally segment each training utterance using the current speech HMM models $\Lambda_{x,k-1}$ and the environmental interference model $\Lambda_{e,k-1}$ given by the previous iteration, where the subscript k denotes the index of iteration. The task can be accomplished, based on the maximum likelihood criterion, by solving the following optimization problem to find the best state sequence $U_k^{(r)} = (u_{1,k}^{(r)}, \ldots, u_{T_r,k}^{(r)})$ and the best mixture component sequence $V_k^{(r)} = (v_{1,k}^{(r)}, \ldots, v_{T_r,k}^{(r)})$ of the optimal segmentation:

$$\begin{aligned}(U_k^{(r)}, V_k^{(r)}) &= \arg\max_{(U^{(r)}, V^{(r)})} \Pr\left(Z^{(r)}, U^{(r)}, V^{(r)} \middle| \Lambda_{x,k-1}, \Lambda_{e,k-1}\right) \\ &= \arg\max_{\left(\left(u_1^{(r)}, \ldots, u_{T_r}^{(r)}\right), \left(v_1^{(r)}, \ldots, v_{T_r}^{(r)}\right)\right)} \left\{ \prod_{t=1}^{T_r} a_{u_{t-1}^{(r)}, u_t^{(r)}} \Pr\left(z_t^{(r)} \middle| u_t^{(r)}, v_t^{(r)}, \Lambda_{z,k-1}^{(r)}\right) \right\},\end{aligned} \tag{3}$$

where $a_{i,j}$ denotes the transition probability from state i to state j. Eq. (3) is solved in this study by first forming the environment-compensated speech HMM models $\Lambda_{z,k-1}^{(r)}$ using $\Lambda_{x,k-1}$ and $\Lambda_{e,k-1}$, and then using the Viterbi search to simultaneously find $U_k^{(r)}$ and $V_k^{(r)}$. The formation of $\Lambda_{z,k-1}^{(r)}$ from $\Lambda_{x,k-1}$ and $\Lambda_{e,k-1}$ is based on the assumed environment contamination model defined in Eqs. (1b) and (1c), and realized by the following two sub-steps:

(1.1) Calculate $\Lambda_{y,k-1}^{(r)}$ in cepstrum domain by

$$\mu_{y,j,q,k-1}^{(r)}(m) = \mu_{x,j,q,k-1}(m) + b_{k-1}^{(r)}(m), \tag{4a}$$

$$\Sigma_{y,j,q,k-1}^{(r)}(m) = \Sigma_{x,j,q,k-1}(m), \tag{4b}$$

where $\mu_{y,j,q,k-1}^{(r)}(m)$ and $\Sigma_{y,j,q,k-1}^{(r)}(m)$ are, respectively, the mean vector and covariance matrix of the qth Gaussian mixture in the jth state of $\Lambda_{y,k-1}^{(r)}$, and $b_{k-1}^{(r)}(m)$ is the bias vector given in $\Lambda_{e,k-1}$.

(1.2) Use the PMC method to form $\Lambda_{z,k-1}^{(r)}$ by first transforming $\Lambda_{y,k-1}^{(r)}$ from cepstrum domain to linear spectrum domain, then combining it with $\Lambda_{n,k-1}^{(r)}$ in linear spectrum domain, and lastly transforming the result back to cepstrum domain.

The second step of the REST algorithm is to enhance the adverse speech by first suppressing the noise using the state-based Wiener filtering method (Hansen and Clements, 1991; Ephraim, 1992; Vaseghi and Milner 1997) and by then removing the signal bias by the SBR method (Rahim and Juang, 1996). It consists of the following two sub-steps:

(2.1) Noise suppression: Given the segmentation information $U_k^{(r)}$, estimate the noise model $\Lambda_{n,k}^{(r)}$ and eliminate it from the input adverse speech $z_t^{(r)}(f)$, in linear-spectrum domain, by the state-based Wiener filtering method to obtain the intermediate signal $y_{t,k}^{(r)}(f)$. The noise model $\Lambda_{n,k}^{(r)}$ and its average power spectrum density $P_{n,k}^{(r)}(f)$ of the rth utterance are re-estimated from the non-speech frames by

$$\mu_{n,k}^{(r)}(m) = \frac{\sum_{t=1}^{T_r} z_t^{(r)}(m) \times I\left(u_{t,k}^{(r)} \in \text{non-speech}\right)}{\sum_{t=1}^{T_r} I\left(u_{t,k}^{(r)} \in \text{non-speech}\right)}, \tag{5a}$$

$$\Sigma_{n,k}^{(r)}(m) = \frac{\sum_{t=1}^{T_r} \left(z_t^{(r)}(m)\right)^2 \times I\left(u_{t,k}^{(r)} \in \text{non-speech}\right)}{\sum_{t=1}^{T_r} I\left(u_{t,k}^{(r)} \in \text{non-speech}\right)} - \left(\mu_{n,k}^{(r)}(m)\right)^2, \tag{5b}$$

$$P_{n,k}^{(r)}(f) = \frac{\sum_{t=1}^{T_r} \hat{P}_{z,t}^{(r)}(f) \times I\left(u_{t,k}^{(r)} \in \text{non-speech}\right)}{\sum_{t=1}^{T_r} I\left(u_{t,k}^{(r)} \in \text{non-speech}\right)}, \tag{5c}$$

where $\hat{P}_{z,t}^{(r)}(f)$ is the periodogram of $z_t^{(r)}$, which is defined as

$$\hat{P}_{z,t}^{(r)}(f) = \frac{1}{L}\left|z_t^{(r)}(f)\right|^2, \tag{6}$$

and L is the analysis length of the FFT operation; $I(\cdot)$ is the zero–one indicator function. Basing on Eq. (1b) of the assumed environment contamination model, the Wiener filter for the jth state of speech model and the rth training utterance is constructed and expressed by

$$W_{j,k}^{(r)}(f) = \frac{P_{y,j,k-1}(f)}{P_{y,j,k-1}(f) + P_{n,k}^{(r)}(f)}, \tag{7a}$$

where $P_{y,j,k-1}(f)$ is the average power density spectrum corresponding to the jth state of the bias-compensated speech HMM models. After forming all state-based Wiener filters, we calculate the enhanced signal by

$$y_{t,k}^{(r)}(f) = W_{u_t,k}^{(r)}(f) \times z_t^{(r)}(f), \quad \text{for } t = 1, \ldots, T_r \text{ and } u_t \neq \text{non-speech}. \tag{7b}$$

(2.2) SBR: Given with the segmentation information $(U_k^{(r)}, V_k^{(r)})$, estimate the signal bias and remove it from the intermediate signal $y_{t,k}^{(r)}(f)$ to obtain the environment-normalized speech estimate. The SBR method is realized by first transforming $y_{t,k}^{(r)}(f)$ to $y_{t,k}^{(r)}(m)$, then making a simplified assumption of $\Sigma_{z,j,q}^{(r)} =$ identity matrix in Eq. (A.11) of Appendix A to obtain

$$b_k^{(r)}(m) = \frac{\sum_{t=1}^{T_r} \left(y_{t,k}^{(r)}(m) - \mu_{x,u_{t,k}^{(r)},v_{t,k}^{(r)},k-1}(m) \right) \times I\left(u_{t,k}^{(r)} \notin \text{non-speech} \right)}{\sum_{t=1}^{T_r} I\left(u_{t,k}^{(r)} \notin \text{non-speech} \right)} \tag{8a}$$

and lastly removing the signal bias by

$$x_{t,k}^{(r)}(m) = y_{t,k}^{(r)}(m) - b_k^{(r)}(m). \tag{8b}$$

The third step of the REST algorithm is to re-estimate the speech HMM models $\Lambda_{x,k}$ and the average power density spectrum $\{P_{y,j,k-1}(f)\}_{j=1,\ldots,N_j}$ using, respectively, the enhanced speech signals $\{X_k^{(r)}(m)\}_{r=1,\ldots,R}$ and $\{Y_k^{(r)}(m)\}_{r=1,\ldots,R}$ based on the current segmentation information $\{(U_k^{(r)}, V_k^{(r)})\}_{r=1,\ldots,R}$, where N_j denotes the total number of states in HMM models.

The combination of all operations in above three steps can be interpreted as a sequential optimal estimation procedure listed in the following:

For iteration k

For utterance $r = 1$ to R, do

$$(U_k^{(r)}, V_k^{(r)}) = \arg\max_{(U^{(r)}, V^{(r)})} \Pr\left(Z^{(r)}, U^{(r)}, V^{(r)} \middle| \Lambda_{x,k-1}, \Lambda_{e,k-1}\right), \tag{9a}$$

$$\Lambda_{n,k}^{(r)} = \arg\max_{\Lambda_n^{(r)}} \Pr\left(Z^{(r)} \middle| \Lambda_n^{(r)}, \left(U_k^{(r)}, V_k^{(r)}\right)\right), \tag{9b}$$

$$Y_k^{(r)} = \arg\max_{Y^{(r)}} \Pr\left(Y^{(r)} \middle| Z^{(r)}, U_k^{(r)}, \Lambda_{n,k}^{(r)}, \{P_{y,j,k-1}\}_{j=1,\ldots,N_j}\right), \tag{9c}$$

$$b_k^{(r)} = \arg\max_{b^{(r)}} \Pr\left(Y_k^{(r)} \middle| b^{(r)}, \left(U_k^{(r)}, V_k^{(r)}\right), \Lambda_{x,k-1}\right), \tag{9d}$$

$$X_k^{(r)} = \arg\max_{X^{(r)}} \Pr\left(X^{(r)} \middle| Y_k^{(r)}, b_k^{(r)}\right). \tag{9e}$$

End loop for r

$$\{P_{y,j,k}\}_{j=1,\ldots,N_j} = \arg\max_{\{P_{y,j}\}_{j=1,\ldots,N_j}} \Pr\left(\{Y^{(r)}\}_{r=1,\ldots,R} \middle| \{P_{y,j}\}_{j=1,\ldots,N_j}, \left(U_k^{(r)}\right)_{r=1,\ldots,R}\right), \tag{9f}$$

$$\Lambda_{x,k} = \arg\max_{\Lambda_x} \Pr\left(\left\{X_k^{(r)}\right\}_{r=1,\ldots,R} \middle| \Lambda_x, \left(U_k^{(r)}, V_k^{(r)}\right)_{r=1,\ldots,R}\right). \tag{9g}$$

Repeat for k until the average likelihood score converges.

A similar idea was used in (Lim and Oppenheim, 1978; Hansen and Clements, 1991) to employ a sequential MAP estimation procedure in an iterative algorithm to sequentially estimate the linear prediction coefficients, gain, and the noise-free speech waveform for frame-level speech enhancement.

The REST algorithm can also be derived by using the EM algorithm (Dempster et al., 1977). So its convergence can be guaranteed. Detailed derivations of the EM procedure for estimating (Λ_x, Λ_e) is given in Appendix A.

Like other iterative algorithms, the REST algorithm must be initialized by giving an initial set of speech HMM models, an initial set of state averaged power density spectra, an initial channel bias vector, and an initial noise model. The initial speech HMM models and the initial state averaged power density spectra can be constructed by a conventional ML training algorithm using either an enhanced version of the given adverse-speech training set or another training set with high SNR. In the study, we adopt the former approach to use an enhanced speech training set obtained by subtracting the given initial noise model from the adverse-speech training set. The initial noise models are obtained from non-speech frames of the adverse-speech training set detected by an RNN-based speech segmentation method (Hong and Chen, 1997). It uses an RNN classifier, directly trained from adverse speech, to classify the input speech pattern into three broad-classes: *initial*, *final* and non-speech. The speech segmentation method has been shown to perform well in noisy environment (Hong et al., 1999). The initial bias vector is obtained by the SBR method using the above enhanced speech training set.

3. The PMC–SBC method for Mandarin base-syllable recognition

Mandarin Chinese is a tonal language. Each Chinese character is pronounced as a syllable with a tone. There are, in total, about 1300 syllables. If the tones are disregarded, there are only 411 phonologically allowed base-syllables. The phonetic structures of these 411 base-syllables are very regular and relatively simple as compared with English. A base-syllable can be decomposed into an optional *initial* and a *final*. There are in total 22 *initials* (including a null) and 39 *finals*. Although, the base-syllable set is only in medium size, its recognition is actually very difficult because it comprises many highly confusable sets. Specifically, all 411 base-syllables can be categorized into 39 confusable sets according to their *finals*. Like the English E-set, all base-syllables in each confusable set differ only in their *initial* consonants and are therefore difficult to be distinguished (Chang et al., 1993; Lee and Juang, 1996). Besides, cross-set confusion between these 39 sets are also easy to occur. Medial confusion and nasal-ending confusion are the two most commonly occurred types of cross-set confusion. Highly discriminative speech models are therefore needed to tackle the difficult task. In this study, a set of sub-syllable HMM models containing 100 3-state right-*final*-dependent *initial* models and 39 5-state context-independent *final* models is used as basic recognition units (Wang and Chen, 1998). In each state, a mixture Gaussian distribution with diagonal covariance matrices is used. The number of mixture in each state is variable and depends on the number of training samples, but a fixed maximum value is set for it. Besides, a single-state, single-mixture, utterance-dependent model is used for noise.

An integrated PMC-based Mandarin base-syllable recognition method, which is a modified version of the PMC method for additive and convolutional noise (Gales and Young, 1995; Nakamura et al., 1996) by additionally considering broad-class based likelihood compensation (Hong and Chen, 1997), is employed in this work to test the reference speech HMM models generated by the proposed REST training algorithm. It can be regarded as the combination of the PMC method and a signal bias compensation (SBC) method and is referred to as the PMC–SBC method. A block diagram of the new recognizer is displayed in Fig. 2. Each input testing utterance is first processed in the RNN-based Speech Segmentation (Hong and Chen, 1997) to detect non-speech frames. The RNN-based speech segmentation uses a three-layer simple RNN to discriminate each input frame among three broad-classes of *initial*, *final* and non-speech.

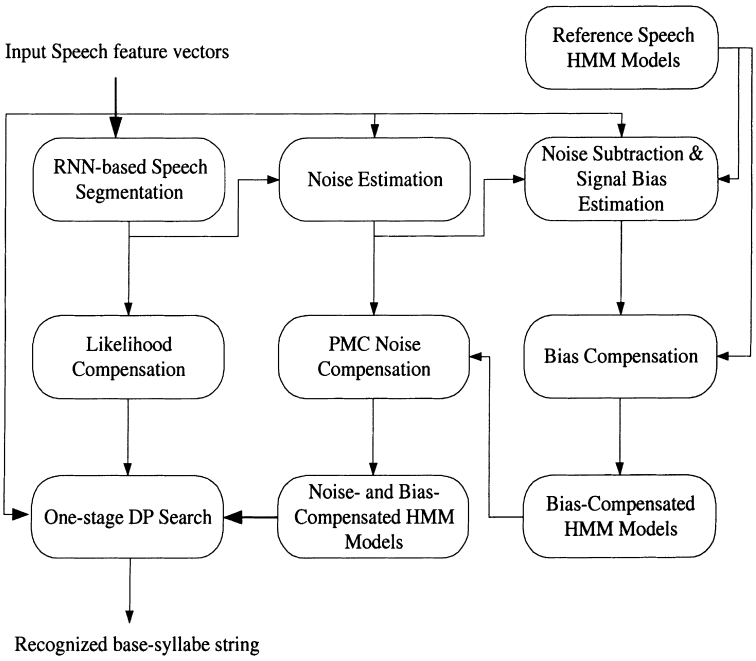

Fig. 2. A block diagram of the PMC–SBC method for testing the REST algorithm.

Non-speech frames are then detected by comparing the RNN non-speech output with a pre-determined threshold and used in the noise estimations to estimate the noise model. The input utterance is then processed in the Noise Subtraction and Signal Bias Estimation by first subtracting the noise model estimate to obtain an enhanced speech and then transforming to cepstrum domain to estimate the signal bias by the SBR method (Rahim and Juang, 1996). The SBR method estimates the signal bias by first encoding the feature vectors of the enhanced speech using a codebook and then calculating the average encoding residuals. The codebook is formed by collecting the mean vectors of mixture components of all reference speech HMM models. The bias estimate is then used in the Bias Compensation to convert all reference speech HMM models into bias-compensated speech HMM models. These models are then further converted, in the PMC Noise Compensation, into noise- and bias-compensated speech HMM models using the above noise model estimate. The PMC noise-compensation method used adopts the log-normal approximation (Gales and Young, 1993) for its noise-combination operator. These noise- and bias-compensated speech HMM models are then used in the One-stage DP Search to generate the recognized base-syllable sequence for the input adverse testing utterance. The One-stage DP Search uses a Viterbi search algorithm invoking with cumulative bounded-state-duration constraints (Wang and Chen, 1998) to accomplish its task with the help of the Likelihood Compensation. The likelihood compensation (LC) scheme used is the one proposed previously for improving the PMC-based recognition method for noisy Mandarin speech (Hong and Chen, 1997; Hong et al., 1999). The LC scheme uses the broad-class classification information, provided by the RNN outputs, to help reduce the recognition errors caused by the misalignments of syllable boundaries. Due to its importance, the LC scheme is briefly discussed as follows. Although the PMC method is effective on adapting the clean-speech HMM models to match with the testing noise environment, the discrimination capabilities of the noise-compensated HMM models are still subject to be degraded resulted from the noise perturbation on the distributions of the recognition features of speech patterns. This noise-perturbation effect will make all speech phones more difficult to be distinguished not

only to each other but also from the background noise. The PMC method can do nothing to compensate this effect. This noise-induced confusing effect was also confirmed in a recent study by Junqua et al. (1994) on a simple 10-digit noisy speech recognition task. They found that a large portion of recognition errors is owing to word boundary misalignments caused by the confusing between speech signals and the background noise. To partially cure the weakness of the PMC method, the LC scheme uses the broad-class classification information provided by the RNN to assist in the recognition. It directly takes the three RNN outputs as weighting factors to add additional scores to the log-likelihood scores of HMM states associated with the three broad classes, i.e.,

$$\rho_j^c(z_t) = \begin{cases} \rho_j(z_t) + \alpha \log(W_I(t)), & j \in \textit{initial}, \\ \rho_j(z_t) + \alpha \log(W_F(t)), & j \in \textit{final}, \\ \rho_j(z_t) + \alpha \log(W_N(t)), & j \in \text{non-speech}, \end{cases} \tag{10}$$

where $W_I(t)$, $W_F(t)$ and $W_N(t)$ are the *initial*, *final* and non-speech outputs of the RNN, $\rho_j(z_t)$ is the log-likelihood score of state j, and α is a scaling factor to control the degree of the likelihood compensation. It is noted that, if hard-decisions are performed in the broad-class classification to make $W_I(t)$, $W_F(t)$ and $W_N(t)$ become 0–1 functions, the LC scheme is equivalent to a restricted recognition search scheme in which only sub-syllables belonging to the detected broad-class are needed to be considered.

4. Evaluation

Performance of the proposed REST algorithm was evaluated on two multi-speaker Mandarin base-syllable recognition tasks. Due to the fact that the previous studies on robust training for eliminating the noise effect were still very few, we examined the effectiveness of the REST training algorithm on eliminating the noise effect in detail in the first task. Both the REST training algorithm and the PMC–SBC recognition method were simplified by discarding the parts related to the signal bias compensation. In the second task, the complete function of the REST algorithm on eliminating both the signal bias and noise effects was examined. In the following experiments, the base-syllable accuracy rate defined below was used to evaluate the recognition performance:

$$\text{base-syllable accuracy rate} = \left(1 - \frac{\text{Subs} + \text{Dels} + \text{Ins}}{\text{number of testing base-syllables}}\right) \times 100(\%), \tag{11}$$

where Subs, Dels and Ins denoted the numbers of substitution, deletion and insertion errors, respectively.

4.1. Performance evaluation I

In the first task, the performance of the REST algorithm on the adverse environment with only additive noise interference was examined. The noisy speech databases used in this study were generated by artificially adding noises to a clean-speech database composing of 1200 utterances of four speakers including two males and two females. Each utterance comprised several syllables and was pronounced in such a way that every syllable was clearly pronounced. The database contained in total 6197 syllables including 5124 training syllables and 1073 testing syllables. All speech signals were digitally recorded in a laboratory using a PC with a 16-bit Sound Blaster card and a head-set microphone. A sampling rate of 16 kHz was used. Two noisy-speech databases were artificially generated from the clean-speech database by adding noises of two different types including the Lynx helicopter noise from NOISEX-92 (Varga, 1993) and a computer-generated white Gaussian noise. For simplicity, these two noise types are referred to as Lynx and White noises, respectively. For each noise type, the training database contained three noisy-speech data sets of 12, 24 and 36 dB in SNR.

The open test used another three data sets for each noise type with 9, 18 and 30 dB in SNR. All speech signals were first pre-processed for each of 20 ms Hamming-windowed frame with 10 ms shift. Then, a set of 25 recognition features including 12 MFCC, 12 delta MFCC and a delta log-energy was computed for each frame. The maximum number of mixture components in each HMM state was set to be 5.

We first examined the efficiency of the speech HMM models generated by the REST algorithm using the *F*-ratio measure (Nicholson et al., 1997). The *F*-ratio is a measure of class separability in the acoustic feature space and can be roughly defined by

$$F\text{-ratio} = \frac{\text{variance of means}}{\text{mean of variances}}. \tag{12}$$

In this test, the classes were defined to include all states of the speech HMM models. The variance of means is the sample variance of all state means of these HMM models, and the mean of variances is the sample mean of all state variances. Obviously, a larger *F*-ratio measure indicates a larger separation among the states of the speech HMM models, which in turn roughly indicates that they have a higher discrimination capability. In the study, two schemes of the REST training algorithm with two different sets of initial models were tested. The first set of initial models, denoted as INIT1, was formed by the clean-speech HMM models, clean-speech state average power density spectra, and the exact noise models. Since INIT1 was an ideal model, the first scheme was not practical and hence was taken for reference only. The other set of initial models, denoted as INIT2, was a practical one and was generated by firstly segmenting all training utterances by the RNN-based speech segmentation method (Hong and Chen, 1997), secondly estimating the initial utterance-dependent noise models from non-speech frames of those training utterances, and lastly estimating the initial speech HMM models and the initial state average power density spectra from the enhanced version of the original training set obtained by subtracting the initial noise model. Figs. 3 and 4 show the feature-based *F*-ratio measures of the resulting HMM models for the two cases using Lynx and White noises, respectively. It can be seen from these two figures that the *F*-ratio measures for both schemes of the REST algorithm with INIT1 and INIT2 are comparable and are all better than the HMM_B models (to be defined later) trained by the conventional *k*-means algorithm. This is especially true for the lower-order recognition features. So the speech HMM models generated by the proposed REST algorithm are more compact and hence expected to possess better discrimination capability. Fig. 5 shows the learning curve of the REST algorithm. It can be found from Fig. 5 that the average log-likelihood score increases monotonically with respect to the iteration number. This empirically shows the convergence of the REST algorithm.

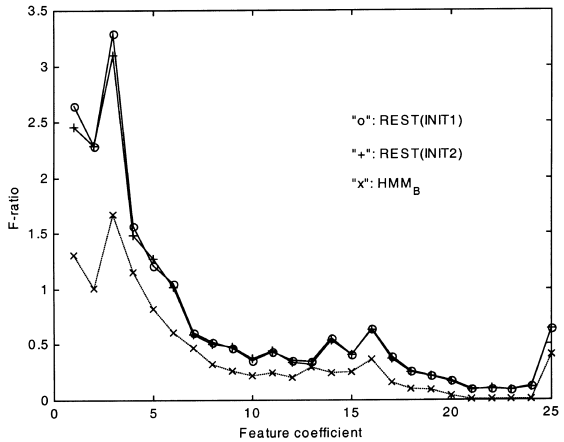

Fig. 3. The *F*-ratio measures of the speech HMM models trained from the noisy speech training database corrupted with Lynx noise.

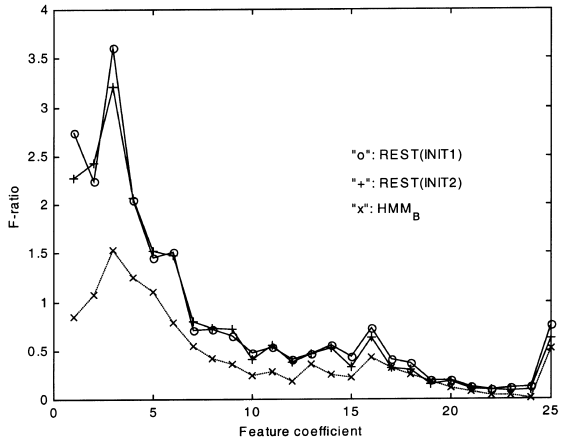

Fig. 4. The *F*-ratio measures of the speech HMM models trained from the noisy speech training database corrupted with White noise.

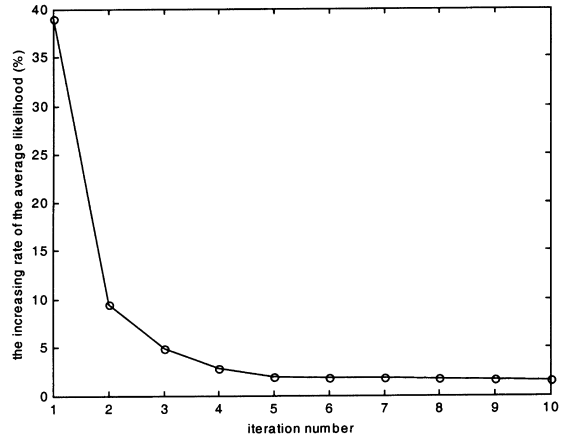

Fig. 5. The learning curve of the REST algorithm for the first task.

We then examined the recognition performance of the speech HMM models generated by the REST algorithm. The performance of the HMM method when both training and testing data were clean speech was also tested and taken as a benchmark. Its base-syllable recognition rate was 80.5%. In this test, four sets of reference speech HMM models were compared. They included:

M1. HMM_C: The HMM models trained from the clean-speech database by the ML-based segmental *k*-means algorithm.
M2. HMM_B: The HMM models trained from the noisy-speech database with three different SNRs by the ML-based segmental *k*-means algorithm.
M3. HMM_R: The HMM models trained from the noisy-speech database with three different SNRs by the proposed REST algorithm.
M4. HMM_M: The HMM models trained from a noisy-speech data set with SNR matched with the testing speech by the ML-based segmental *k*-means algorithm. That is, the HMM models trained from 9, 18 or 30 dB noisy-speech data set were used to recognize noisy speech with the same SNR.

For comparing the performances of these four sets of reference speech HMM models on noisy speech recognition, the following three recognition schemes were used:

S1-1. The 'NC' scheme: The conventional HMM recognition method without noise compensation.
S1-2. The 'PMC' scheme: The conventional PMC method (Gales and Young, 1993) with noise model being estimated based on RNN-based speech segmentation. Its noise-compensation operation used the log-normal approximation.
S1-3. The 'PMC/LC' scheme: An extended version of the 'PMC' scheme invoking with the likelihood compensation scheme. It is a degenerated version of the PMC–SBC method discussed in Section 3 with the parts related to signal-bias compensation being discarded.

Tables 1 and 2 show the experimental results of the open tests for the two cases using Lynx and White noises, respectively. It is noted that, in the implementation of the PMC recognition method using $\mathrm{HMM_B}$ as reference models, the mean of the estimated noise model, $\hat{\mu}_n^{(r)}(f)$, was intuitively modified by

$$\hat{\mu}_n^{(r)}(f) \leftarrow \begin{cases} \hat{\mu}_n^{(r)}(f) - \hat{\mu}_{n_0}(f), & \text{if } \hat{\mu}_n^{(r)}(f) > \hat{\mu}_{n_0}(f), \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

to count the noise effect embedded in the $\mathrm{HMM_B}$ models. Here $\hat{\mu}_{n_0}(f)$ is the noise mean of the training database estimated in the training process of generating the $\mathrm{HMM_B}$ models. From Tables 1 and 2, the following observations can be found:

O1. For $\mathrm{HMM_B}$, the NC scheme performed fair for both noise types with SNR = 18 dB and SNR = 30 dB. But it performed very bad for both noise types with SNR = 9 dB.
O2. For $\mathrm{HMM_M}$, the NC scheme performed very well for both noise types with all the three SNRs.
O3. For $\mathrm{HMM_B}$, the PMC scheme performed only slightly better than the NC scheme for both noise types with SNR = 18 dB and SNR = 30 dB, and much better for SNR = 9 dB.
O4. The NC scheme with $\mathrm{HMM_M}$ performed better than the PMC scheme with $\mathrm{HMM_C}$ for both noise types with all the three SNRs.

Table 1
The recognition results of the open tests for noisy speech corrupted with Lynx noise (unit: %)

	$\mathrm{HMM_B}$		$\mathrm{HMM_C}$		$\mathrm{HMM_R}$		$\mathrm{HMM_M}$
SNR (dB)	NC	PMC	PMC	PMC/LC	PMC	PMC/LC	NC
9	−12.1	34.9	39.1	42.3	43.6	48.7	45.0
18	51.4	52.0	58.6	62.5	62.8	67.7	66.3
30	62.3	65.1	71.2	75.1	73.6	78.3	75.6

Table 2
The recognition results of the open tests for noisy speech corrupted with White noise (unit: %)

	$\mathrm{HMM_B}$		$\mathrm{HMM_C}$		$\mathrm{HMM_R}$		$\mathrm{HMM_M}$
SNR(dB)	NC	PMC	PMC	PMC/LC	PMC	PMC/LC	NC
9	−35.9	29.8	26.9	33.0	35.0	38.1	33.6
18	42.8	45.2	48.3	52.0	54.2	58.0	57.0
30	58.4	59.9	65.4	71.8	68.2	73.8	68.6

O5. For both PMC and PMC/LC schemes, HMM_R performed better than HMM_C for both noise types with all the three SNRs.
O6. For both HMM_R and HMM_C, the PMC/LC scheme performed much better than the PMC scheme.
O7. The PMC/LC scheme with HMM_R performed better than the NC scheme with HMM_M.

Based on these observations, the following conclusions can be drawn:

C1-1. From O1–O2, the conventional HMM method without noise compensation can be used in noisy speech recognition only when the noise level of the training data set is the same as that of the testing speech. If the training database contains noisy speech with diverse noise levels, its performance will degrade seriously.
C1-2. From O1–O3, the HMM models generated by the conventional k-means training algorithm are good for the NC scheme in the noise-level matched condition, fair in the noise-level interpolation condition, and bad in the noise-level extrapolation condition.
C1-3. From O1 and O3, the performance improvements for the HMM method using HMM_B reference models by the PMC noise compensation are very limited.
C1-4. From O4, the log-normal approximation of the noise-compensation operation used in the PMC scheme is not perfect.
C1-5. From O5 and C1-4, the REST algorithm is a very efficient training algorithm to generate noise-suppressed HMM models directly from a noisy speech database with diverse noise levels. The resulting HMM models perform very well in the PMC scheme for testing noisy speech with untrained noise levels. They are even better than the clean-speech HMM models for the PMC method when the noise-compensation operation is not perfect. So it is a very promising robust training algorithm.
C1-6. From O6–O7, the likelihood compensation scheme is very helpful for the PMC-based noisy speech recognition. Actually, the PMC/LC scheme using HMM_R reference models performed best in all cases of the test.

An extra test on noisy English digit recognition using the NOISEX-92 database (Varga and Steeneken, 1993) was performed to examine the validity of the proposed REST algorithm. The database contains utterances of isolated digits and digit triples uttered by one male and one female speakers. Here only the part of isolated-digit utterances was used. The database contains in total 400 digits including 200 training tokens and 200 testing tokens. Each testing utterance comprises 100 digits and was uttered in such a way that every digit was clearly pronounced. All speech signals were first pre-processed for each of 25 ms Hamming-windowed frame with 10 ms shift. Then, 12 MFCC were computed for each frame and taken as the recognition features. For each digit, an 8-state HMM model with observations in each state being modeled by a mixture Gaussian distribution was trained. The number of mixture components in each state was set to be 2. Besides, a single-state, single-mixture model was used for noise.

In the test, we considered the performance of the REST algorithm on the adverse environment with additive noise interference only. Noisy-speech databases were artificially generated from the clean-speech database by adding computer-generated white Gaussian noise. The noisy training database contained four data sets of 0, 6, 12 and 24 dB in SNR. The open test used another five data sets of −3, 0, 3, 9 and 18 dB in SNR. The same accuracy rate defined in Eq. (11) was used to evaluate the recognition performance. We note that the benchmark of the recognition performance achieved by the conventional ML-trained HMM method for the clean-speech case is 100%. Three recognition schemes used in the first test were compared. They included:

Table 3
The recognition results of the NOISEX-92 database corrupted by White noise (unit: %)

SNR (dB)	HMM_B–NC	HMM_C–PMC	HMM_R–PMC
−3	39.5	72.0	82.5
0	63.5	86.0	94.5
3	82.0	94.0	99.0
9	93.0	98.5	99.5
18	94.0	99.5	99.5

1. HMM_B–NC: The conventional HMM method without noise compensation using HMM models trained from noisy speech.
2. HMM_C–PMC: The PMC method using clean-speech HMM models.
3. HMM_R–PMC: The PMC method using HMM models trained by the REST algorithm.

Table 3 shows the experimental results. It can be seen from the table that HMM_B–NC performed the worst, HMM_C–PMC the next, and HMM_R–PMC the best. This result is consistent with what we have obtained in the first test of the study on adverse Mandarin speech recognition.

4.2. Performance evaluation II

In the second task, the performance of the REST algorithm on adverse environment with both channel bias and noise interferences was examined. A simulated telephone-speech database generated by corrupting a clean-speech database with both convolutional channel bias and additive white noise was used in this study. The clean-speech database was generated by 10 speakers including 8 males and 2 females. It was a super-set of the clean-speech database used in the first task with the same recording condition. It contained, in total, 3050 utterances including 2572 training utterances (12 800 syllables) and 478 testing utterances (2666 syllables). To generate the adverse-speech database, each clean-speech utterance was first corrupted by a computer-generated white Gaussian noise and then passed through a filter which simulated a telephone channel. This was realized simply by first adding the white noise in time domain and then adding the channel bias in frequency domain. It is noted that the assumed environment contamination model shown in Fig. 1 is still suitable for modeling the simulated database. In the training database generation, noises with levels of 12, 24 and 36 dB in SNR were separately added to three subsets of the clean-speech training database. These three subsets contained utterances of three, three and four speakers, respectively. In the testing database generation, noises with levels of 9, 18 and 30 dB in SNR were added to the whole clean-speech testing database. To simulate the channel variations on the telephone speech through the public switching network, a set of 227 simulated filters was generated from a large telephone-speech database provided by Chunghwa Telecommunication Laboratories. Each filter was obtained by performing a frame-based cepstrum average to the long utterance of a telephone call through the public switching network. Fig. 6 shows their frequency responses. Among these 227 channel filters, 195 were used to generate the training database while all others were used in the testing database generation. It is noted that the stationarity of the environment characteristics for each utterance is guaranteed in this simulated adverse-speech database via the use of utterance-dependent channel filter and noise level.

The same format of speech HMM models as the first task was used here. The only difference was that the maximal number of mixtures used in each HMM state was increased to 20. In the REST algorithm, the initial condition was generated from the same adverse training database by a four-step procedure. First, segment all training utterances by the RNN-based speech segmentation method (Hong and Chen, 1997). Second, estimate the initial utterance-dependent noise model from the non-speech frames of each training utterance. Third, estimate the initial speech HMM models and the initial state average power density

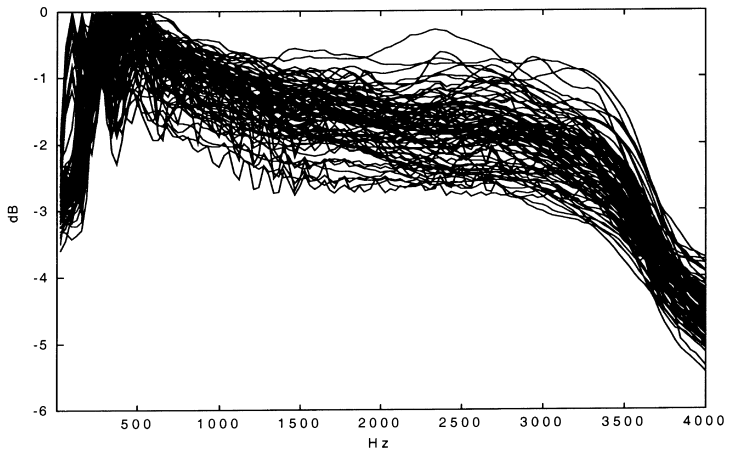

Fig. 6. The frequency responses of the simulated telephone channels.

spectra from the enhanced version of the original training set obtained by subtracting the initial noise models. Last, estimate the initial channel bias vectors from the same enhanced training set by the SBR method (Rahim and Juang, 1996).

In this test, the following recognition schemes were compared:

S2-1. The 'BASELINE' scheme: The conventional HMM method using the reference speech models trained directly from the adverse training database by the segmental *k*-means algorithm.
S2-2. The 'CLEAN' scheme: The PMC–SBC recognition method using the clean-speech reference HMM models, but without invoking the LC scheme.
S2-3. The 'REST-bias' scheme: The SBC recognition method using the reference HMM models trained by the REST algorithm without considering noise suppression.
S2-4. The 'REST-noise' scheme: The PMC recognition method using the reference HMM models trained by the REST algorithm without considering signal bias removal.
S2-5. The 'REST' scheme: The PMC–SBC recognition method using the REST-trained reference HMM models, but without invoking the LC scheme.
S2-6. The 'REST/LC' scheme: The PMC–SBC recognition method using the REST-trained reference HMM models.

Table 4 shows the base-syllable recognition results of these six schemes for adverse speech corrupted with channel bias and White noise. It can be found from Table 4 that, according to the recognition rate, these six schemes can be ordered as: REST/LC > REST > REST-noise or REST-bias > BASELINE > CLEAN. Based on the experimental results, the following conclusions can be made:

Table 4
The recognition results of the open tests for adverse speech corrupted with channel bias and White noise (unit: %)

SNR (dB)	BASELINE	CLEAN	REST-bias	REST-noise	REST	REST/LC
9	23.4	14.8	24.5	29.3	33.0	35.2
18	46.7	27.3	50.2	48.4	53.7	56.5
30	60.2	45.6	62.7	61.8	65.5	66.7

C2-1. The conventional HMM method using the reference models trained by the k-means algorithm performed fair in adverse speech recognition.
C2-2. The result that the CLEAN scheme performed much worse than the BASELINE scheme is mainly owing to the imperfection of the channel bias compensation performed in the SBC method. Actually, the CLEAN scheme was totally fail to compensate the mismatch between the testing speech and the clean-speech HMM model. This primarily resulted from the large deviation on the estimated signal bias from the real channel bias.
C2-3. Although the channel bias-compensation operation of the SBC method is imperfect, the REST training algorithm can still take its advantage by embedding it into the iterative training process to make the resulting HMM models more suitable to be used with the channel bias compensation of the testing process. This has been confirmed by the fact that both the REST-bias and REST scheme performed better than the BASELINE scheme.
C2-4. The HMM models generated by the REST algorithm which considers both noise suppression and signal bias removal are better than those obtained by the REST algorithm considering only noise suppression or signal bias.
C2-5. The likelihood compensation scheme is still effective on assisting in the adverse speech recognition.

A final test to check whether the REST training scheme is operable for clean-speech environment was lastly done. It is worthwhile to note that some robust training algorithms, designed for improving the performance of speech recognizers under adverse-speech environment, performed not well for clean-speech environment. In the test, two sets of HMM models were generated, respectively, by the conventional ML training method and by the REST training scheme using the same clean-speech database. The base-syllable recognition rate was 76.05% for the ML method and 76.24% for the REST scheme. This result confirmed that the REST algorithm did not degrade the system performance when the training data were clean speech.

5. Conclusions

A robust training algorithm for generating a set of speech HMM models directly from a training database collected in adverse environment for adverse speech recognition has been discussed in this paper. Its main advantage lies on the incorporation of the signal bias-compensation and PMC noise-compensation operations of a given robust adverse speech recognition method into its iterative training process so as to make the resulting speech HMM models more suitable to be used in the given robust adverse speech recognition method. Its effectiveness on generating robust speech HMM models has been confirmed by simulations. Experimental results showed that the HMM models it generated were even better than the clean-speech HMM models for use in the given robust adverse speech recognition method when the PMC noise-compensation and/or channel bias-compensation operations are imperfect. So it is a promising robust training algorithm.

Acknowledgements

This work was supported by the National Science Council of Taiwan under Contract no. NSC87-2213-E-009-056. The telephone-speech database was provided by the Chunghwa Telecommunication Laboratories.

Appendix A. The EM procedure for estimating $\{\Lambda_x, \Lambda_e\}$

Eq. (2) can be solved using an iterative EM procedure (Dempster et al., 1977) which tries to find the local optimal estimate of $\hat{\Theta} \equiv \{\hat{\Lambda}_x, \hat{\Lambda}_e\}$ with the following two intermediate parameter sequences involved: the hidden state sequences $\{U^{(r)} = (u_1^{(r)}, \ldots, u_{T_r}^{(r)})\}_{r=1,\ldots,R}$ and the mixture component sequences $\{V^{(r)} = (v_1^{(r)}, \ldots, v_{T_r}^{(r)})\}_{r=1,\ldots,R}$. The first (expectation) step of the EM procedure is to compute the auxiliary Q-function defined as

$$Q(\Theta, \hat{\Theta}_{k-1}) = E\left\{ \log L\left(\{Z^{(r)}, U^{(r)}, V^{(r)}\}_{r=1,\ldots,R} \middle| \Theta\right) \middle| \{Z^{(r)}\}_{r=1,\ldots,R}, \hat{\Theta}_{k-1} \right\}. \tag{A.1}$$

Here the subscript $k-1$ denotes the iteration index. In the second (maximization) step, new values of $\hat{\Theta}_k$ are computed based on the maximization of $Q(\Theta, \hat{\Theta}_{k-1})$:

$$\hat{\Theta}_k = \arg\max_{\Theta} Q(\Theta, \hat{\Theta}_{k-1}). \tag{A.2}$$

The detailed derivation of the EM procedure is described as follows.

Let

$$\Lambda_z^{(r)} = \begin{cases} G\left(\Lambda_x; \Lambda_n^{(r)}, b^{(r)}\right) & \text{for adverse-speech model,} \\ \Lambda_n^{(r)} & \text{for non-speech model} \end{cases} \tag{A.3}$$

be the environment-compensated HMM models, constructed from Λ_x and ($\Lambda_n^{(r)}$, $b^{(r)}$), for the rth observation utterance $Z^{(r)}$. Here $G(\cdot)$ denotes a mapping function that transforms Λ_x to match with the current environment of $Z^{(r)}$. By assuming that, in $\Lambda_z^{(r)}$, observations are mixture-Gaussian-distributed, we can calculate the mean vector $\mu_{z,j,q}^{(r)}$ and covariance matrix $\Sigma_{z,j,q}^{(r)}$ of the qth mixture component in the jth state of $\Lambda_z^{(r)}$, based on the assumed environment contamination model defined in Eqs. (1b) and (1c), by

$$\mu_{z,j,q}^{(r)} = \begin{cases} (\mu_{x,j,q} + b^{(r)}) \otimes \Lambda_n^{(r)}, & j \in \text{adverse-speech model,} \\ \mu_n^{(r)}, & j \in \text{non-speech model,} \end{cases} \tag{A.4a}$$

$$\Sigma_{z,j,q}^{(r)} = \begin{cases} \Sigma_{x,j,q} \otimes \Lambda_n^{(r)}, & j \in \text{adverse-speech model,} \\ \Sigma_n^{(r)}, & j \in \text{non-speech model,} \end{cases} \tag{A.4b}$$

where $\otimes$ denotes the PMC noise-compensation operator (Gales and Young, 1993), and $\mu_{x,j,q}$ and $\Sigma_{x,j,q}$ are, respectively, the mean vector and covariance matrix of the qth mixture component in the jth state of Λ_x. By further assuming that the state-based Wiener filtering is the inverse operation of the PMC (Gales and Young, 1993; Vaseghi and Milner, 1997), we can express the compensated cepstral mean $\mu_{z,j,q}^{(r)}$ in Eq. (A.4a) by (Gales and Young, 1993; Vaseghi and Milner, 1997)

$$\mu_{z,j,q}^{(r)} = \begin{cases} \mu_{x,j,q} + b^{(r)} + h_j, & j \in \text{adverse-speech model,} \\ \mu_n^{(r)}, & j \in \text{non-speech model,} \end{cases} \tag{A.5}$$

where h_j is the cepstral coefficients of the state-based Wiener filter of the jth state which is constructed from an estimate of the signal power density spectrum at the jth state and an estimate of the noise power density spectrum of the rth utterance.

Based on the above expression of $\Lambda_z^{(r)}$, the auxiliary Q-function can be rewritten as (Sankar and Lee, 1996)

$$
\begin{aligned}
Q(\Theta, \hat{\Theta}_{k-1}) &= Q\left(\left(\Lambda_x, \{\Lambda_n^{(r)}, b^{(r)}\}_{r=1,\ldots,R}\right), \hat{\Theta}_{k-1}\right) \\
&= Q\left(\left(\{\Lambda_z^{(r)}\}_{r=1,\ldots,R}\right), \hat{\Theta}_{k-1}\right) \\
&= Q_{k-1} + \sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,k-1}^{(r)}(j,q)\log \Pr\left(z_t^{(r)}, u_t = j, v_t = q \middle| \Theta\right) \\
&= Q_{k-1} + \sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,k-1}^{(r)}(j,q)\log N(z_t^{(r)}; \mu_{z,j,q}^{(r)}, \Sigma_{z,j,q}^{(r)}),
\end{aligned} \tag{A.6}
$$

where

$$
\gamma_{t,k-1}^{(r)}(j,q) \equiv \Pr\left(z_t^{(r)}, u_t^{(r)} = j, v_j^{(r)} = q \middle| \hat{\Theta}_{k-1}\right) \tag{A.7}
$$

is the probability of the observation $z_t^{(r)}$ produced from the qth mixture component of the jth state; N_j and N_q denote, respectively, the total numbers of states and mixture components; $N(\cdot)$ represents normal distribution; and Q_{k-1} is a function depending only on the transition probability and mixture probability of $\hat{\Lambda}_{z,k-1}^{(r)}$ (which are assumed to be the same as those of $\hat{\Lambda}_{x,k-1}$). But, due to the fact that it is generally difficult to derive a close form solution for the above joint maximization problem, a multi-stage sequential maximization procedure is employed to approximate the local optimum of $\hat{\Theta}_k$. In each stage, only one type of parameters is optimally estimated.

We first estimate the parameters of noise model $\hat{\Lambda}_{n,k}^{(r)}$ to maximize the Q-function in Eq. (A.6). They can be obtained by

$$
\hat{\mu}_{n,k}^{(r)} = \frac{\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,n,k-1}^{(r)}(j,q) z_t^{(r)}}{\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,n,k-1}^{(r)}(j,q)}, \tag{A.8a}
$$

$$
\hat{\Sigma}_{n,k}^{(r)} = \frac{\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,n,k-1}^{(r)}(j,q)\left(z_t^{(r)} - \hat{\mu}_{n,k}^{(r)}\right)\left(z_t^{(r)} - \hat{\mu}_{n,k}^{(r)}\right)^{\mathrm{T}}}{\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,n,k-1}^{(r)}(j,q)}, \tag{A.8b}
$$

where $\gamma_{t,n,k-1}^{(r)}(j,q) \equiv \gamma_{t,k-1}^{(r)}(j,q) I(j \in \text{ non-speech})$ and $I(\cdot)$ is the zero–one indicator function.

We then estimate the signal bias $\hat{b}_k^{(r)}$. After replacing $\Lambda_n^{(r)}$ and Λ_x with $\hat{\Lambda}_{n,k}^{(r)}$ and $\hat{\Lambda}_{x,k-1}$, the Q-function becomes

$$
\begin{aligned}
&Q\left(\left(\hat{\Lambda}_{x,k-1}, \left\{\hat{\Lambda}_{n,k}^{(r)}, b^{(r)}\right\}_{r=1,\ldots,R}\right), \hat{\Theta}_{k-1}\right) \\
&\quad = Q'_{k-1} + \sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,s,k-1}^{(r)}(j,q)\log N\left(z_t^{(r)}; \hat{\mu}_{x,j,q,k-1} + b^{(r)} - h_{j,k}, \Sigma_{z,j,q,k}^{(r)}\right) \\
&\quad = Q'_{k-1} + \sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,s,k-1}^{(r)}(j,q)\log N\left(\left(z_t^{(r)} + h_{j,k} - b^{(r)}\right); \hat{\mu}_{x,j,q,k-1}, \Sigma_{z,j,q,k}^{(r)}\right) \\
&\quad = Q'_{k-1} + \sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,s,k-1}^{(r)}(j,q)\log N\left(\left(y_{t,j,k}^{(r)} - b^{(r)}\right); \hat{\mu}_{x,j,q,k-1}, \Sigma_{z,j,q,k}^{(r)}\right),
\end{aligned} \tag{A.9}
$$

where $\gamma_{t,s,k-1}^{(r)}(j,q) \equiv \gamma_{t,k-1}^{(r)}(j,q) I(j \in \text{speech})$; $h_{j,k}$ and $\Sigma_{z,j,q,k}^{(r)}$ are updated versions of h_j and $\Sigma_{z,j,q}^{(r)}$ with $\Lambda_n^{(r)}$ and Λ_x being replaced with $\hat{\Lambda}_{n,k}^{(r)}$ and $\hat{\Lambda}_{x,k-1}$, and $y_{t,j,k}^{(r)}$ is the Wiener-filtered version of $z_t^{(r)}$ at the jth state. By solving

$$\frac{\partial Q\left(\left(\hat{\Lambda}_{x,k-1}, \left\{\hat{\Lambda}_{n,k}^{(r)}, b^{(r)}\right\}_{r=1,\ldots,R}\right), \hat{\Theta}_{k-1}\right)}{\partial b^{(r)}} = 0, \tag{A.10}$$

the pth element of $\hat{b}_k^{(r)}$ can be obtained by (Sankar and Lee, 1996)

$$\hat{b}_k^{(r)}(p) = \frac{\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,s,k-1}^{(r)}(j,q)\left(\Sigma_{z,j,q,k}^{(r)}(p,p)\right)^{-1}\left(y_{t,j,k}^{(r)}(p) - \hat{\mu}_{x,j,q,k-1}(p)\right)}{\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,s,k-1}^{(r)}(j,q)\left(\Sigma_{z,j,q,k}^{(r)}(p,p)\right)^{-1}}, \tag{A.11}$$

where $y_{t,j,k}^{(r)}(p)$ and $\hat{\mu}_{x,j,q,k-1}(p)$ denote, respectively, the pth elements of $y_{t,j,k}^{(r)}$, and $\hat{\mu}_{x,j,q,k-1}$, and $\Sigma_{z,j,q,k}^{(r)}(p,p)$ is the (p,p)th element of $\Sigma_{z,j,q,k}^{(r)}$.

We then estimate $\hat{\Lambda}_{x,k}$. After replacing $\Lambda_n^{(r)}$ and $b^{(r)}$ with $\hat{\Lambda}_{n,k}^{(r)}$ and $\hat{b}_k^{(r)}$, the Q-function becomes

$$Q\left(\left(\Lambda_x, \left\{\hat{\Lambda}_{n,k}^{(r)}, \hat{b}_k^{(r)}\right\}_{r=1,\ldots,R}\right), \hat{\Theta}_{k-1}\right) = Q'_{k-1} + \sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,s,k-1}^{(r)}(j,q)\log N\left(x_{t,j,k}^{(r)}; \mu_{x,j,q}, \Sigma_{x,j,q}\right), \tag{A.12}$$

where

$$x_{t,j,k}^{(r)} = y_{t,j,k}^{(r)} - \hat{b}_k^{(r)} = z_t^{(r)} + h_{j,k} - \hat{b}_k^{(r)} \tag{A.13}$$

is the signal bias-removed and Wiener-filtered signal of the jth state. The Q-function is now in the same form as that in the conventional EM algorithm for estimating HMM's parameters. So, the mean and covariance of $\hat{\Lambda}_{x,k}$ can be estimated in the same way by

$$\hat{\mu}_{x,j,q,k} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,s,k-1}^{(r)}(j,q)x_{t,j,k}^{(r)}}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,s,k-1}^{(r)}(j,q)}, \tag{A.14a}$$

$$\hat{\Sigma}_{x,j,q,k} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,s,k-1}^{(r)}(j,q)\left(x_{t,j,k}^{(r)} - \hat{\mu}_{x,j,q,k}\right)\left(x_{t,j,k}^{(r)} - \hat{\mu}_{x,j,q,k}\right)^{\mathrm{T}}}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}\gamma_{t,s,k-1}^{(r)}(j,q)}. \tag{A.14b}$$

The HMM state transition probabilities and the mixture component coefficients can also be estimated by the standard EM method.

It can be verified that the Q-function will increase at each stage of the sequential maximization procedure, i.e.,

$$\begin{aligned} Q\left(\hat{\Theta}_{k-1}; \hat{\Theta}_{k-1}\right) &= Q\left(\left(\hat{\Lambda}_{x,k-1}, \left\{\hat{\Lambda}_{n,k-1}^{(r)}, \hat{b}_{k-1}^{(r)}\right\}_{r=1,\ldots,R}\right); \hat{\Theta}_{k-1}\right) \\ &\leqslant Q\left(\left(\hat{\Lambda}_{x,k-1}, \left\{\hat{\Lambda}_{n,k}^{(r)}, \hat{b}_{k-1}^{(r)}\right\}_{r=1,\ldots,R}\right); \hat{\Theta}_{k-1}\right) \\ &\leqslant Q\left(\left(\hat{\Lambda}_{x,k-1}, \left\{\hat{\Lambda}_{n,k}^{(r)}, \hat{b}_{k}^{(r)}\right\}_{r=1,\ldots,R}\right); \hat{\Theta}_{k-1}\right) \\ &\leqslant Q\left(\left(\hat{\Lambda}_{x,k}, \left\{\hat{\Lambda}_{n,k}^{(r)}, \hat{b}_{k}^{(r)}\right\}_{r=1,\ldots,R}\right); \hat{\Theta}_{k-1}\right) \\ &= Q\left(\hat{\Theta}_k; \hat{\Theta}_{k-1}\right). \end{aligned} \tag{A.15}$$

This in turn leads to an increase on the likelihood of the training data in each iteration (Dempster et al., 1977), i.e.,

$$L\left(\{Z^{(r)}\}_{r=1,\ldots,R}\middle|\hat{\Theta}_k\right) \geqslant L\left(\{Z^{(r)}\}_{r=1,\ldots,R}\middle|\hat{\Theta}_{k-1}\right). \tag{A16}$$

Hence, the EM procedure is guaranteed to converge.

In practical implementation, the above EM procedure needs to be modified by invoking with the segmental k-means algorithm (Juang and Rabiner, 1990) in order to increase its computational efficiency. It adds an additional pre-segmentation stage into the above iterative re-estimation procedure. In each iteration, all training utterances are first optimally segmented by the Viterbi algorithm (Forney, 1973) to determine the best state sequences $\{\hat{U}_k^{(r)}\}_{r=1,\ldots,R}$ and the best mixture component sequences $\{\hat{V}_k^{(r)}\}_{r=1,\ldots,R}$. Then, parameters of all models are re-estimated based on the given $\{\hat{U}_k^{(r)}, \hat{V}_k^{(r)}\}_{r=1,\ldots,R}$. All formulations of the above EM procedure listed in Eqs. (A.3)–(A.14) still hold except that $\gamma_{t,s,k-1}^{(r)}(j,q)$ and $\gamma_{t,n,k-1}^{(r)}(j,q)$ are now associated only with $\{\hat{U}_k^{(r)}, \hat{V}_k^{(r)}\}$ and hence all $\sum_{j=1}^{N_j}\sum_{q=1}^{N_q}$ in Eqs. (A.6), (A.8), (A.9), (A.11), (A.12) and (A.14) have to be taken away.

A final modification of the above re-estimation procedure is needed to replace the optimal signal bias estimation with the conventional SBR method. By making a simplified assumption of $\hat{\Sigma}_{z,j,q,k-1}^{(r)} = I$, the modified version of Eq. (A.11) can be reduced to Eq. (8a). This completes the derivations of the REST training algorithm.

References

Acero, A., Stern, R.M., 1990. Environmental robustness in automatic speech recognition. In: Proceedings of ICASSP-90, pp. 849–852.

Acero, A., Stern, R.M., 1991. Robust speech recognition by normalization of the acoustic space. In: Proceedings of ICASSP-91, pp. 893–896.

Anastasakos, T., McDonough, J., Makhoul, J., 1997. Speaker adaptive training: a maximum likelihood approach to speaker normalization. In: Proceedings of ICASSP-97, pp. 1043–1046.

Chang, P.-C., Chen, S.-H., Juang, B.-H., 1993. Discriminative analysis of distortion sequences in speech recognition. IEEE Trans. Speech and Audio Process. 1, 326–333.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. 39, 1–38.

Ephraim, Y., 1992. Statistical-model-based speech enhancement systems. Proc. IEEE 80, 1526–1555.

Forney, G., 1973. The Viterbi algorithm. Proc. IEEE 61, 268–278.

Furui, S., 1992. Toward robust speech recognition under adverse conditions. In: Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions, pp. 31–24.

Gales, M.J.F., Woodland, P.C., 1996. Mean and variance adaptation within the MLLR framework. Comput. Speech and Language 10, 249–264.

Gales, M.J.F., Young, S.J., 1993. Cepstral parameter compensation for HMM recognition in noise. Speech Communication 12, 231–239.

Gales, M.J.F., Young, S.J., 1995. Robust speech recognition in additive and convolutional noise using parallel model combination. Comput. Speech and Language 9, 289–307.

Gales, M.J.F., Young, S.J., 1996. Robust continuous speech recognition using parallel model combination. IEEE Trans. Speech and Audio Process. 5, 352–359.

Gong, Y., 1995. Speech recognition in noisy environments: A survey. Speech Communication 16, 261–291.

Gong, Y., 1997. Source normalization training for HMM applied to noisy telephone speech recognition. In: Proceedings of EuroSpeech-97, Vol. 3, pp. 1555–1558.

Hansen, J.H.L., Clements, M.A., 1991. Constrained iterative speech enhancement with application to speech recognition. IEEE Trans. Signal Process. 39, 795–805.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech and Audio Process. 2, 578–589.

Hong, W.-T., Chen, S.-H., 1997. A robust RNN-based pre-classification for Noisy Mandarin speech recognition. In: Proceedings of EuroSpeech-97, Vol. 3, pp. 1083–1086.

Hong, W.-T., Liao, Y.-F., Wang, Y.-R., Chen, S.-H., 1999. RNN-based speech segmentation and its applications to robust noisy Mandarin speech recognition. J. Acoust. Soc. Amer., revised.

Juang, B.-H., 1991. Speech recognition in adverse environment. Comput. Speech and Language 5, 275–294.

Juang, B.-H., Rabiner, L.R., 1990. The segmental *K*-means algorithm for estimating parameters of hidden Markov models. IEEE Trans. Acoust. Speech Signal Process. 38, 1639–1641.

Junqua, J.-C., Halton, J.-P., 1996. Robustness in Automatic Speech Recognition: Fundaments and Applications. Kluwer Academic Press, Boston, MA.

Junqua, J.S., Mak, B., Reaves, B., 1994. A robust algorithm for word boundary detection in the presence of noise. IEEE Trans. Speech and Audio Process. 2, 406–412.

Lee, C.-H., 1998. On stochastic feature and model compensation approaches to robust speech recognition. Speech Communication 25, 29–47.

Lee, C.-H., Juang, B.-H., 1996. A survey on automatic speech recognition with an illustrative example on continuous speech recognition of Mandarin. J. Comput. Linguist. Chinese Language Process. 1, 1–36.

Lim, J.S., Oppenheim, A.V., 1978. All-pole modeling of degraded speech. IEEE Trans. Acoust. Speech Sig. Process. 26, 197–210.

Liu, F.-H., Picheny, M., Srinivasa, P., Monkowaski, M., Chen, J., 1996. Speech recognition on Mandarin call home: a large vocabulary, conversational and telephone speech corpus. In: Proceedings of ICASSP-96, Vol. 1, pp. 157–160.

Lockwood, P., Boudy, J., 1992. Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars. Speech Communication 11, 215–228.

Mokbel, C.E., Chollet, G.F.A., 1995. Automatic word recognition in cars. IEEE Trans. Speech and Audio Process. 3, 346–356.

Minami, Y., Furui, S., 1996. Adaptation method based on HMM composition and EM algorithm. In: Proceedings of ICASSP-96, pp. 327–330.

Nakamura, S., Takigucgi, T., Shikano, K., 1996. Noise and room acoustics distorted speech recognition by HMM composition. In: Proceedings of ICASSP-96, Vol. 1, pp. 69–72.

Nicholson, S., Milner, B., Cox, S., 1997. Evaluating features set performance using the *F*-ratio and *J*-measures. In: Proceedings of EuroSpeech-97, Vol. 1, pp. 413–416.

Rahim, M., Juang, B.-H., 1996. Signal bias removal by maximum likelihood estimation for robust telephone speech recognition. IEEE Trans. Speech and Audio Process. 4, 19–30.

Sankar, A., Lee, C.-H., 1996. A maximum-likelihood approach to stochastic matching for robust speech recognition. IEEE Trans. Speech and Audio Process. 4, 190–202.

Varga, A., Steeneken, H.J.M., 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication 12, 247–251.

Vaseghi, S.V., Milner, B.P., 1997. Noise compensation methods for hidden Markov model speech recognition in adverse environments. IEEE Trans. Speech and Audio Process. 5, 11–21.

Wang, Y.-R., Chen, S.-H., 1998. Mandarin telephone speech recognition for automatic telephone number directory service. In: Proceedings of ICASSP-98, Vol. 2, pp. 841–844.

Zhao, Y., 1996. Self-learning speaker and channel adaptation based on spectral variation source decomposition. Speech Communication 18, 65–77.