

Noise Robust Speech Recognition Using Feature Compensation Based on Polynomial Regression of Utterance SNR

Xiaodong Cui, *Student Member, IEEE*, and Abeer Alwan, *Senior Member, IEEE*

Abstract—A feature compensation (FC) algorithm based on polynomial regression of utterance signal-to-noise ratio (SNR) for noise robust automatic speech recognition (ASR) is proposed. In this algorithm, the bias between clean and noisy speech features is approximated by a set of polynomials which are estimated from adaptation data from the new environment by the expectation-maximization (EM) algorithm under the maximum likelihood (ML) criterion. In ASR, the utterance SNR for the speech signal is first estimated and noisy speech features are then compensated for by regression polynomials. The compensated speech features are decoded via acoustic HMMs trained with clean data. Comparative experiments on the Aurora 2 (English) and the German part of the Aurora 3 databases are performed between FC and maximum likelihood linear regression (MLLR). With the Aurora 2 experiments, there are two MLLR implementations: pooling adaptation data across all SNRs, and using three distinct SNR clusters. For each type of noise, FC achieves, on average, a word error rate reduction of 16.7% and 16.5% for Set A, and 20.5% and 14.6% for Set B compared to the first and second MLLR implementations, respectively. For each SNR condition, FC achieves, on average, a word error rate reduction of 33.1% and 34.5% for Set A, and 23.6% and 21.4% for Set B. Results using the Aurora 3 database show that, the best FC performance outperforms MLLR by 15.9%, 3.0% and 14.6% for well-matched, medium-mismatched and high-mismatched conditions, respectively.

Index Terms—Feature compensation, noise robust speech recognition, polynomial regression, signal-to-noise ratio (SNR) estimation.

I. INTRODUCTION

SPEECH recognition systems trained in quiet environments suffer from performance degradation in the presence of ambient acoustic noise. The degradation is mainly attributed to the mismatch between clean acoustic models and noisy speech data. Considerable efforts have been made to reduce this mismatch and improve recognition accuracy in noisy conditions [1]. Generally speaking, noise robust algorithms are applied in the front-end feature domain and/or in the back-end model domain.

In the front-end feature domain, spectral subtraction [2] is a commonly used method for noise suppression where

Manuscript received March 12, 2004; revised September 11, 2004. This work was supported in part by the NSF under Grant 0326214. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Kuldip K. Paliwal.

The authors are with the Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA (e-mail: xdcui@icsl.ucla.edu; alwan@icsl.ucla.edu).

Digital Object Identifier 10.1109/TSA.2005.853002

additive noise spectrum is estimated and subtracted from the noisy speech spectrum to recover the clean speech spectrum. Cepstral mean normalization (CMN) [3] which removes the mean of cepstral features is another popular approach to deal with convolutive channel noise. In [4], an signal-to-noise ratio (SNR)-dependent cepstral normalization (SDCN) algorithm is proposed to compensate the noisy speech features in the cepstral domain by removing the compensation vectors from them in a discrete HMM recognizer. The compensation vectors clustered by frame SNRs are estimated using “stereo” data which consist of clean and noisy speech signals recorded simultaneously. Statistical speech feature enhancement based on conditional minimum mean square error estimation is investigated in [5] where the joint prior of static and frame-differential dynamic cepstral features are utilized. Other research on noise-robust feature extraction includes RASTA [6] and perceptual linear predictive (PLP) [7] where noise-robust features are extracted based on human perceptual characteristics. More recently, noise robust front-end feature schemes have been widely investigated for distributed speech recognition (DSR), e.g., [8] and [9]. In [8], noise robust speech features are obtained using a combination of SNR-dependent waveform processing, two passes of Wiener filtering and blind equalization techniques and achieve impressive performance. In [9], low-complexity algorithms such as peak isolation [10], harmonic demodulation [11] and variable frame rate [12] are combined to generate speech features and result in very good recognition performance in adverse environments. The algorithms are inspired by the observation that in case of noise, peaks in speech spectra are more noise robust than valleys, and that formant transitions carry important perceptually-discriminative information.

In the back-end model domain, parallel model combination (PMC) [13] is widely used to adapt clean acoustic models to match the acoustic environment. According to the relationship between clean and noisy speech signals and based on the estimation of additive and convolutive noise, PMC transforms the means and covariances of Gaussian mixtures in clean acoustic HMMs toward the true distributions of the noisy speech features. The transformation generated by PMC can be fairly accurate but the computation is expensive since all acoustic model parameters have to be modified. If the environmental statistics are not stationary, model parameters need to be changed constantly. The compensation relationship between the distributions of clean and noisy speech features being statistically nontrivial, vector Taylor series (VTS) and vector polynomial approximations (VPS) are proposed in [14] and

[15], respectively, to approximate the nonlinear compensation relationship which allow environment adaptation to be carried out in a tractable way. The Jacobian approach in [16] approximates PMCs complicated model compensation computation by linearization using Jacobian matrices between the initial and test environments. Since the linearization is best performed within a small neighborhood of the initial conditions, the Jacobian method works best for the cases where training and test conditions are not much different. Maximum likelihood linear regression (MLLR) [17] is also an effective way to adapt the clean acoustic models to a new environment, although it was originally developed for speaker adaptation. Without having prior knowledge of the noise, MLLR obtains the environmentally matched models by rotating and shifting the means of the Gaussian mixtures of clean HMMs using linear regression. In comparison with PMC, MLLR is computationally attractive. A modified version of MLLR called piecewise-linear transformation (PLT) is studied in [18] where various types of noise are clustered based on their spectral property and one set of noisy acoustic models is trained for each cluster under a variety of SNR conditions. In recognition, the best matched HMM set is selected and further adapted by MLLR.

In [19], significant improvements are achieved in noisy speech recognition by using variable parameter hidden Markov models (VPHMM) whose Gaussian mean vectors under different environments are described by polynomial functions of an environment-dependent continuous variable. In recognition, one set of HMMs is instantiated according to the environment. By modeling the trend of the Gaussian means, VPHMM has smaller Gaussian variances which indicates higher model discriminative abilities. Typically, the estimation of the state emission parameter polynomials requires a relatively large amount of data under the target environments.

In this paper, feature compensation based on polynomial regression of the utterance signal-to-noise ratio (SNR) is investigated. The bias between the clean and noisy speech features in the cepstral domain is approximated by a set of polynomials with respect to utterance SNR. During the recognition stage, utterance SNR is first estimated and compensation bias is then computed and removed from the noisy speech cepstral feature. The compensated feature is fed into the clean acoustic HMMs decoding network. Feature compensation polynomials are estimated under the maximum likelihood (ML) criterion using the expectation-maximization (EM) algorithm. Depending on the amount of adaptation data available, the polynomials could be flexibly tied at different levels of granularity. By learning the trend of the bias as a function of SNR, the algorithm is able to predict the biases at unseen SNRs. The biases, not the means, are approximated by the polynomials in this paper because the biases allow more flexible tying schemes with limited adaptation data, e.g., global tying or a few tying classes, which is not easy to perform directly on the means. Furthermore, with the knowledge of clean acoustic models, biases can achieve more robust estimation compared with the means.

The remainder of this paper is organized as follows. In Section II, the motivation and formulation of feature compensation utilizing polynomial regression of utterance SNRs is given. The utterance SNR estimation method based on minimum statistics

tracking is described in Section III. The training and recognition scheme of the feature compensation algorithm and comparative experimental results with MLLR are shown in Section IV. Section V concludes the paper with a summary.

II. POLYNOMIAL REGRESSION OF SIGNAL-TO-NOISE RATIO

In this section, the motivation of using SNR-based regression polynomials for feature compensation is introduced. The ML estimation of the regression polynomials in an EM framework is also described.

A. Bias Approximation by SNR Polynomials

For additive noise, assuming that clean speech signals and noise are statistically independent in each filter bin, the power of a noisy speech signal in the k th filter bin of each frame is the sum of the power of clean speech and noise of the filter bin

$$Y_k^{\text{lin}} = X_k^{\text{lin}} + N_k^{\text{lin}} \quad (1)$$

where Y_k^{lin} , X_k^{lin} and N_k^{lin} denote noisy speech, clean speech and noise in the linear power domain of the k th filter bin respectively. For the k th filter bin in the log-power domain, (1) could be rewritten as

$$\begin{aligned} Y_k^{\text{log}} &= X_k^{\text{log}} + \log \left(1 + \frac{N_k^{\text{lin}}}{X_k^{\text{lin}}} \right) \\ &= X_k^{\text{log}} + \log \left(1 + \frac{1}{\text{SNR}_k} \right) \\ &= X_k^{\text{log}} + g_k \end{aligned} \quad (2)$$

where Y_k^{log} and X_k^{log} represent noisy and clean speech in the log-power domain, SNR_k is the signal-to-noise ratio and

$$g_k \triangleq \log \left(1 + \frac{1}{\text{SNR}_k} \right). \quad (3)$$

Applying Discrete Cosine Transform (DCT) on both sides of (2), we get the n th cepstral coefficient as

$$\begin{aligned} Y_n^{\text{cep}} &= \sum_k d_{nk} Y_k^{\text{log}} \\ &= \sum_k d_{nk} (X_k^{\text{log}} + g_k) \\ &= \sum_k d_{nk} X_k^{\text{log}} + \sum_k d_{nk} g_k. \end{aligned} \quad (4)$$

Since g_k is a function of the utterance SNR, (4) could be written as

$$Y_n^{\text{cep}} = X_n^{\text{cep}} + f_n(\text{SNR}) \quad (5)$$

where Y_n^{cep} and X_n^{cep} are the n th cepstral component of noisy and clean speech, d_{nk} 's are the DCT coefficients and $f_n(\text{SNR})$ denotes a function of SNR of the n th cepstral coefficient.

From (5), it is clear that the bias between the clean and noisy features is a nonlinear function of SNR. In this paper, this nonlinear function is approximated by a polynomial of order P regressing on SNR, that is

$$Y_n^{\text{cep}} \approx X_n^{\text{cep}} + \sum_{j=0}^P \tilde{c}_{jn}(\text{SNR})^j \quad (6)$$

where $\tilde{c}_{jn}$'s are the coefficients for the j th order items of the n th cepstrum. The above relation provides a way to recover the clean speech feature (X_n^{cep}) by compensating the noisy feature (Y_n^{cep}) with the polynomial approximated bias if one knows the utterance SNR

$$X_n^{\text{cep}} \approx Y_n^{\text{cep}} - \sum_{j=0}^P \tilde{c}_{jn}(\text{SNR})^j. \quad (7)$$

B. Feature Compensation

Assuming that the clean acoustic models are Gaussian mixture HMMs, the probability density function of observing feature $\mathbf{o}_t$ from state i is computed as

$$p(\mathbf{o}_t | s_t = i) = \sum_k \alpha_{ik} b_{ik}(\mathbf{o}_t) \quad (8)$$

where $b_{ik}(\mathbf{o}_t) \sim \mathcal{N}(\mathbf{o}_t; \mu_{ik}, \Sigma_{ik})$ is the k th multivariate Gaussian mixture in state i with weight α_{ik} and μ_{ik} and Σ_{ik} are the mean vector and covariance matrix associated with it, respectively.

The feature compensation algorithm removes the polynomial approximated bias from the noisy speech feature using the estimated SNR during the mixture Gaussian probability calculation, which is shown in

$$p(\mathbf{o}_t | s_t = i) = \sum_k \alpha_{ik} \mathcal{N} \left(\mathbf{o}_t - \sum_{j=0}^P \mathbf{c}_{ikj} \eta^j; \mu_{ik}, \Sigma_{ik} \right). \quad (9)$$

In (9), $\mathbf{o}_t$ is the noisy speech feature, μ_{ik} and Σ_{ik} are the mean and covariance of Gaussian mixtures in clean acoustic HMMs. η is the utterance SNR. $\mathbf{c}_{ikj}$'s are the coefficients of the regression polynomials of state i , mixture k and polynomial order j . $\mathbf{c}_{ikj}$ is a vector with the same dimension as the feature vector which means each component in the feature vector has its own regression polynomial with coefficients $\tilde{c}_{ikjn}$.

Depending on the adaptation data available from the new environment, the regression polynomials could be tied flexibly at different levels of granularity-mixtures, states, phonetic classes or globally shared for all HMMs.

C. Maximum Likelihood Estimation of Regression Polynomials

The ML estimation of the regression polynomials from the environmental adaptation data is performed under an EM [20] framework.

We assume that the incorporation of the feature compensation into the Gaussian mixture calculation does not affect the initial state probabilities, state transition probabilities and Gaussian mixture weights. Therefore, define the EM auxiliary function we are interested in as

$$Q_b(\lambda; \bar{\lambda}) = \sum_{r=1}^R \sum_{i \in \Omega_s} \sum_{k \in \Omega_m} \sum_{t=1}^{T^r} \gamma_t^r(i, k) \cdot \log b_{ik}(\mathbf{o}_t^r) \quad (10)$$

where R is the utterance number of adaptation data and T^r is the frame number of the r th utterance. $\Omega_s = \{1, 2, \dots, N\}$ and

$\Omega_m = \{1, 2, \dots, M\}$ are the state and mixture sets, respectively. $\gamma_t^r(i, k) = p(s_t^r = i, \xi_t^r = k | \mathcal{O}^r, \bar{\lambda})$ is the posterior probability of staying at state i mixture k at time t given the r th observation sequence $\mathcal{O}^r = \{\mathbf{o}_1^r, \dots, \mathbf{o}_{T^r}^r\}$.

Without loss of generality, we assume that each Gaussian mixture has one set of distinct regression polynomials. For other tying strategies, the derivations follow accordingly utilizing the collection of the corresponding statistics within each tying set. The extension to other strategies is straightforward and will be discussed later.

Optimizing $Q_b(\lambda; \bar{\lambda})$ with respect to $\mathbf{c}_{ikl}$, one obtains

$$\begin{aligned} \frac{\partial Q_b(\lambda; \bar{\lambda})}{\partial \mathbf{c}_{ikl}} &= \frac{\partial}{\partial \mathbf{c}_{ikl}} \sum_{r=1}^R \sum_{i=1}^N \sum_{k=1}^M \sum_{t=1}^{T^r} \gamma_t^r(i, k) \\ &\quad \cdot \log \mathcal{N} \left(\mathbf{o}_t^r - \sum_{j=0}^P \mathbf{c}_{ikj} (\eta^r)^j; \mu_{ik}, \Sigma_{ik} \right) \\ &= \frac{\partial}{\partial \mathbf{c}_{ikl}} \sum_{r=1}^R \sum_{t=1}^{T^r} \gamma_t^r(i, k) \\ &\quad \times \left[-\frac{1}{2} \left(\mathbf{o}_t^r - \sum_{j=0}^P \mathbf{c}_{ikj} (\eta^r)^j - \mu_{ik} \right)^T \right. \\ &\quad \times \Sigma_{ik}^{-1} \left(\mathbf{o}_t^r - \sum_{j=0}^P \mathbf{c}_{ikj} (\eta^r)^j - \mu_{ik} \right) \left. \right] \\ &= \sum_{r=1}^R \sum_{t=1}^{T^r} \gamma_t^r(i, k) \cdot \Sigma_{ik}^{-1} \\ &\quad \cdot \left(\mathbf{o}_t^r - \sum_{j=0}^P \mathbf{c}_{ikj} (\eta^r)^j - \mu_{ik} \right) \cdot (\eta^r)^l = 0 \\ &\quad l = 0, 1, \dots, P. \quad (11) \end{aligned}$$

By regrouping terms, (11) can be rewritten as

$$\begin{aligned} &\sum_{j=0}^P \left[\sum_{r=1}^R \sum_{t=1}^{T^r} \gamma_t^r(i, k) \cdot \Sigma_{ik}^{-1} \cdot (\eta^r)^{j+l} \right] \mathbf{c}_{ikj} \\ &= \sum_{r=1}^R \sum_{t=1}^{T^r} \gamma_t^r(i, k) \cdot \Sigma_{ik}^{-1} \cdot (\mathbf{o}_t^r - \mu_{ik}) \cdot (\eta^r)^l \\ &\quad l = 0, 1, \dots, P. \quad (12) \end{aligned}$$

In a similar way as [19], define

$$\psi(\zeta, \rho, \alpha, \beta) = \sum_{r=1}^R \sum_{t=1}^{T^r} \gamma_t^r(i, k) \cdot \Sigma_{ik}^{-1} \cdot \zeta^\alpha \rho^\beta. \quad (13)$$

Equation (12) is simplified into

$$\begin{aligned} &\sum_{j=0}^P \psi(\eta^r, \eta^r, l, j) \cdot \mathbf{c}_{ikj} = \psi(\eta^r, \mathbf{o}_t^r - \mu_{ik}, l, 1) \\ &\quad l = 0, 1, \dots, P. \quad (14) \end{aligned}$$

The $P + 1$ equations in (14) can be expressed in a matrix form

$$\mathbf{U}_{ik} \cdot \mathbf{c}_{ik} = \mathbf{v}_{ik}. \quad (15)$$

In (15), $\mathbf{U}_{ik}$ is a $(P+1) \times (P+1)$ dimensional block matrix

$$\mathbf{U}_{ik} = \begin{bmatrix} u_{ik}(0,0) & \cdots & u_{ik}(0,P) \\ \vdots & u_{ik}(l,j) & \vdots \\ u_{ik}(P,0) & \cdots & u_{ik}(P,P) \end{bmatrix} \quad (16)$$

with elements $u_{ik}(l,j)$ being a $D \times D$ matrix where D denotes the feature dimensionality

$$u_{ik}(l,j) = \psi_{ik}(\eta^r, \eta^r, l, j) \quad (17)$$

and

$$\mathbf{c}_{ik} = [\mathbf{c}_{ik0}^T, \dots, \mathbf{c}_{ikl}^T, \dots, \mathbf{c}_{ikP}^T]^T \quad (18)$$

is composed of $P+1$ coefficient vectors $\mathbf{c}_{ikl}$ ($l = 0, \dots, P$), each of which is D dimensional.

On the right side of the (15), $\mathbf{v}_{ik}$ is a $P+1$ dimensional block vector

$$\mathbf{v}_{ik} = [v_{ik}(0), \dots, v_{ik}(l), \dots, v_{ik}(P)]^T \quad (19)$$

where $v_{ik}(l)$ is a D dimensional vector

$$v_{ik}(l) = \psi_{ik}(\eta^r, \mathbf{o}_t^r - \mu_{ik}, l, 1). \quad (20)$$

From (15), the polynomial coefficients $\mathbf{c}_{ikl}$ ($l = 0, \dots, P$) can be computed by inverting the matrix $\mathbf{U}_{ik}$. This operation is computationally expensive if the covariance matrices in ψ are full. However, when the covariance matrices Σ_{ik} are diagonal (which is usually the case), the computational load could be significantly reduced as discussed in [21].

The above describes the formulation for estimating the polynomials that are distinct for each Gaussian mixture. For other tying schemes, the extension of the above derivation is straightforward. Suppose there are K classes $\{\omega_1, \omega_2, \dots, \omega_K\}$ within which the regression polynomials of different Gaussian mixtures are shared.

The optimization of $Q_b(\lambda; \bar{\lambda})$ with respect to $\mathbf{c}_{\omega_q l}$ ($q = 1, \dots, K$) changes to

$$\begin{aligned} \frac{\partial Q_b(\lambda; \bar{\lambda})}{\partial \mathbf{c}_{\omega_q l}} &= \frac{\partial}{\partial \mathbf{c}_{\omega_q l}} \sum_{r=1}^R \sum_{i=1}^N \sum_{k=1}^M \sum_{t=1}^{T^r} \gamma_t^r(i, k) \\ &\quad \cdot \log \mathcal{N} \left(\mathbf{o}_t^r - \sum_{j=0}^P \mathbf{c}_{ikj} (\eta^r)^j; \mu_{ik}, \Sigma_{ik} \right) \\ &= \frac{\partial}{\partial \mathbf{c}_{\omega_q l}} \sum_{r=1}^R \sum_{(i,k) \in \omega_q} \sum_{t=1}^{T^r} \gamma_t^r(i, k) \\ &\quad \times \left[-\frac{1}{2} \left(\mathbf{o}_t^r - \sum_{j=0}^P \mathbf{c}_{\omega_q j} (\eta^r)^j - \mu_{ik} \right)^T \right. \\ &\quad \times \Sigma_{ik}^{-1} \left. \left(\mathbf{o}_t^r - \sum_{j=0}^P \mathbf{c}_{\omega_q j} (\eta^r)^j - \mu_{ik} \right) \right] \\ &= \sum_{r=1}^R \sum_{(i,k) \in \omega_q} \sum_{t=1}^{T^r} \gamma_t^r(i, k) \\ &\quad \cdot \Sigma_{ik}^{-1} \cdot \left(\mathbf{o}_t^r - \sum_{j=0}^P \mathbf{c}_{\omega_q j} (\eta^r)^j - \mu_{ik} \right) \cdot (\eta^r)^l = 0 \\ &\quad l = 0, 1, \dots, P \quad (21) \end{aligned}$$

where $(i, k) \in \omega_q$ denotes the k th Gaussian mixture in state i that belongs to the tying class ω_q .

Similarly, the shared polynomial coefficients satisfy

$$\begin{aligned} \sum_{j=0}^P \left[\sum_{r=1}^R \sum_{(i,k) \in \omega_q} \sum_{t=1}^{T^r} \gamma_t^r(i, k) \cdot \Sigma_{ik}^{-1} \cdot (\eta^r)^{j+l} \right] \mathbf{c}_{\omega_q j} \\ = \sum_{r=1}^R \sum_{(i,k) \in \omega_q} \sum_{t=1}^{T^r} \gamma_t^r(i, k) \cdot \Sigma_{ik}^{-1} \cdot (\mathbf{o}_t^r - \mu_{ik}) \cdot (\eta^r)^l \\ l = 0, 1, \dots, P \quad (22) \end{aligned}$$

and $\mathbf{c}_{\omega_q l}$'s can be solved accordingly.

In [22] and [23], the phase relationship between the clean and noisy speech is investigated. In particular, the phase information is incorporated into the minimum mean square error estimation of clean speech features in the log-Mel power domain in [22] which achieves impressive performance improvements. The consequent non-Gaussian probability density function in the estimator is approximated by a single-point, second-order Taylor series expansion. Unlike the phase-sensitive model in [22], phase is not taken into account as shown in (1). However, phase, which is also a function of utterance SNR, is implicitly represented in the nonlinear bias which is, in turn, approximated by regression polynomials. Compared with the phase-sensitive model which uses a single expansion point for all clean speech mixture components, the regression polynomials can be considered as expansions at a particular Gaussian mixture mean (mixture-specific polynomials) or averaged Gaussian mixture means (tied polynomials) which are also optimized iteratively under the ML criterion. In addition, the proposed feature compensation algorithm is less computationally expensive than the phase-sensitive model.

III. UTTERANCE SIGNAL-TO-NOISE RATIO ESTIMATION

SNRs used in the feature compensation algorithm are estimated based on the minimum statistics tracking algorithm proposed in [24]. It is well known that significant portions of speech signals contain silence. In noisy conditions, silence portions are contaminated with background noise. Under the assumption that the power of noisy speech is the summation of the power of clean speech and background noise, tracking the power spectral minima can provide a fairly accurate estimate of the background noise power, and hence, a good estimate of the SNR. In addition, by tracking the minimum statistics, the algorithm can deal with nonstationary background noise of slowly changing statistical characteristics. One disadvantage of this approach is the bias between the mean and minimum value of the background noise. In this paper, an empirically determined constant factor of 2.0 is applied to correct for this bias. Power spectral minimum statistics are searched within a 0.5 second interval preceding each speech frame. Fig. 1 illustrates the noise power (denoted by “+”) estimated, without bias correction, from the noisy speech spectrum (denoted by “o”) for an utterance labeled 15 dB SNR in the Aurora 2 database. Power is calculated from each frame.

After the noise power is estimated, the clean speech power is computed by subtracting the noise power from the noisy speech power. In case negative values appear, a small positive floor

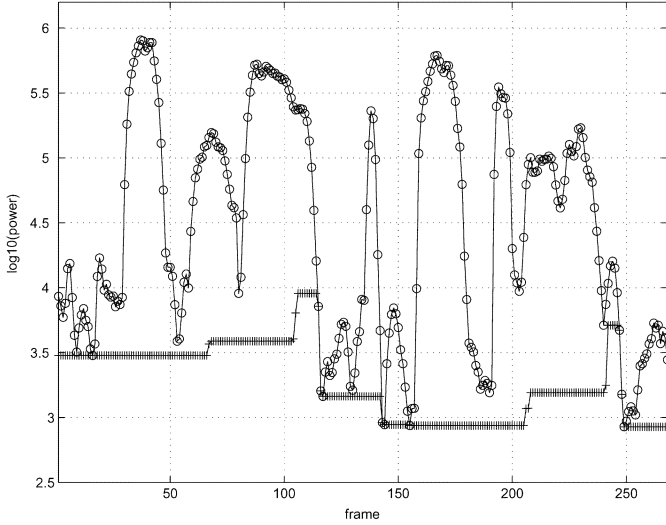


Fig. 1. Noise power(+) estimated by minimum statistics tracking from the noisy speech power spectrum (o) for the utterance “43o6571.” The utterance is labeled 15 dB SNR in the Aurora 2 database.

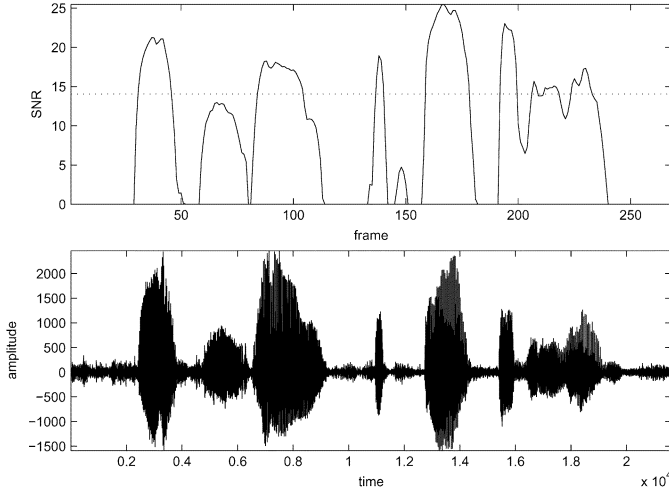


Fig. 2. Estimated frame-wise SNR (solid line in top panel), estimated utterance SNR (dotted line in top panel) and waveform (bottom panel) of the utterance “43o6571” labeled 15 dB SNR in the Aurora 2 database.

is used. Power estimates of clean speech and of noise lead to frame-wise SNR estimates

$$\text{SNR} = 10 \cdot \log_{10} \frac{\text{clean speech power}}{\text{noise power}}. \quad (23)$$

Fig. 2 demonstrates the estimated frame SNRs (solid line) and estimated utterance SNR (dotted line) of the same utterance as in Fig. 1. There is an SNR floor set at 0 dB for all frames since SNR estimates below 0 dB are assumed to be not reliable. The utterance SNR used in the polynomial regression feature compensation is the average of the nonzero frame SNRs of the utterance. The reason for employing utterance SNR instead of frame SNR is that the polynomial approximated bias is only meaningful with respect to the clean speech. Frame SNR reflects the power variation of different portions within the utterance. For example, for a clean utterance, there is no need for compensation although frame SNRs can vary over a wide range. In contrast, the averaged utterance SNR reflects how the overall signal is corrupted by noise.

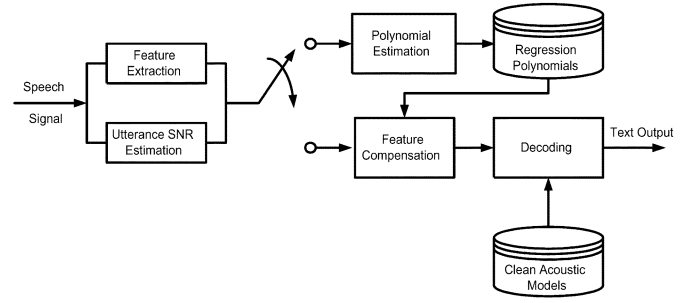


Fig. 3. Training and recognition scheme.

IV. EXPERIMENTAL RESULTS

A. Experimental Conditions

The proposed feature compensation algorithm by polynomial regression of the utterance SNR is trained and tested on the connected digits from the Aurora 2 [25] and the German part of the Aurora 3 databases.

For the Aurora 2 database experiments, there are 8440 clean utterances from 55 male and 55 female adult speakers in the clean training set from which the acoustic HMMs are trained. Speech data in testing sets A and B are used for evaluation. There are eight types of background noise in the Aurora 2 database, which are subway, babble, car and exhibition noise in Set A and restaurant, street, airport and station noise in Set B. Noisy speech data are generated by artificially adding the noise signals at a variety of SNR levels. Six SNR conditions are evaluated in the test which are clean, 20 dB, 15 dB, 10 dB, 5 dB, and 0 dB. For each SNR level and each type of noise in Sets A and B, there are 1001 utterances from 52 male and 52 female adult speakers. Thus, in total, each set consists of 24 024 utterances.

For the German part of the Aurora 3 database, the utterances are recorded in a real car environment that includes four different conditions: stopped with motor running (SMR), town traffic (TT), low speed rough road (LSRR), and high speed good road (HSGR). The data are recorded from different scenarios such as left front window open or closed, sunroof open or closed, etc. All the data are recorded by a close-talking microphone and a hands-free microphone. There are 2929 utterances in the database from which utterances are selected for three training and testing conditions: well-matched (WM), medium-mismatched (MM) and high-mismatched (HM). The WM experiment utilizes speech from both microphone types and all driving conditions for both training and testing. It includes 2032 utterances in training and 897 utterances in testing. The MM experiment utilizes 997 utterances as training data from hands-free microphone using all driving conditions except for the HSGR driving condition, and 241 utterances as testing data from hands-free microphone with HSGR driving condition. The HM experiment utilizes 1007 utterances as training data from the close-talking microphone and all driving conditions, and 394 utterances as testing data from the hands-free microphone for all driving conditions except the SMR driving condition. In all cases, there are 36 male and 43 female talkers in the training set, and 15 male and 18 female talkers in the testing set.

Fig. 3 shows the training and recognition scheme employed in the experiments. For each utterance, Mel-frequency cepstral coefficients (MFCC) features are extracted and the utterance SNR is estimated from the speech signals. The frame length is 25 ms

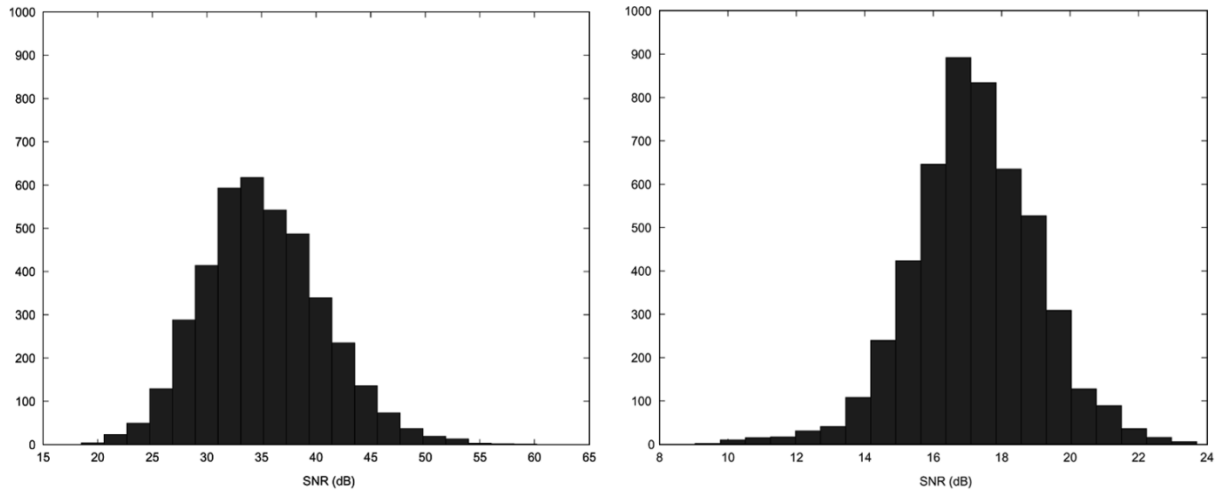


Fig. 4. Histograms of estimated utterance SNRs labeled as clean (left) and 20 dB SNR (right) data in Set A of Aurora 2 database.

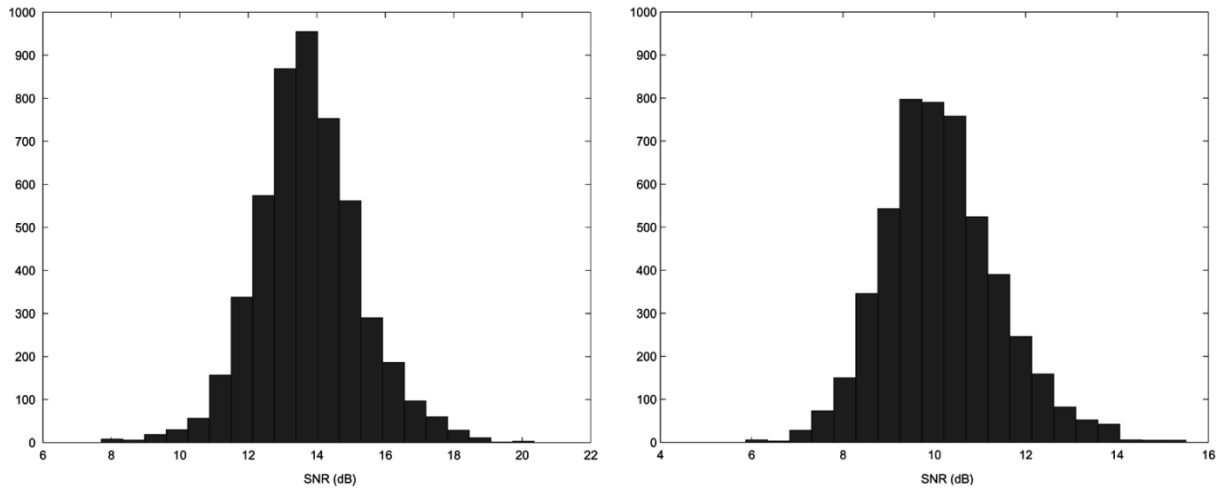


Fig. 5. Histograms of estimated utterance SNRs labeled as 15 dB (left) and 10 dB SNR (right) data in Set A of Aurora 2 database.

and the frame shift is 10 ms. The speech feature for each frame contains 12 static MFCCs (excluding C0) plus log energy (E) and their first and second order derivatives. Therefore, there are 39 components in each feature vector. In the training stage, regression polynomials are estimated from the adaptation data. In the recognition stage, for each frame in the utterance, the polynomial approximated bias is computed based on the utterance SNR and removed from the noisy MFCC features. Cepstral features are then decoded using the Viterbi decoding network with original acoustic HMMs.

The HMMs adopt a left-to-right topology and are word-based models with 16 emission states for each digit, three states for the silence model and 1 state for the short pause model. There are three mixtures in each state of digit models and six mixtures for silence and short pause models. All the Gaussian mixtures have diagonal covariance matrices. The above setup follows the Aurora 2 and 3 specifications.

B. Distribution of Utterance SNRs

It is interesting to compare the estimated utterance SNRs and the original SNR labels provided by the database. Figs. 4 and 5 show histograms of utterance SNRs in Set A estimated by the

minimum statistics tracking algorithm at four SNR conditions in the Aurora 2 database. Small variances (2–3 dB) are observed in the estimated utterance SNRs for each condition in those figures. Since frames with SNR below 0 dB are not included in the utterance SNR calculation and bias exists in the minimum statistics tracking algorithm, the utterance SNRs estimated are not exactly the same as the labeled SNRs. The calibration with the database labeling shows good utterance SNR estimation by the minimum statistics tracking algorithm. Figs. 6–8 indicate the utterance SNR distributions of the training and testing sets under three experimental conditions of the German part of the Aurora 3 database. For the FC algorithm, the absolute utterance SNR accuracy is not critical while the consistency of utterance SNR estimation between training and recognition is.

C. Regression Polynomial Orders

As stated before, the nonlinear bias can be approximated by polynomials of different orders. The higher the order of the polynomials, the smaller the approximation errors. However, higher order polynomials can also result in overfitting and more parameters to estimate. With limited learning data, unreliable

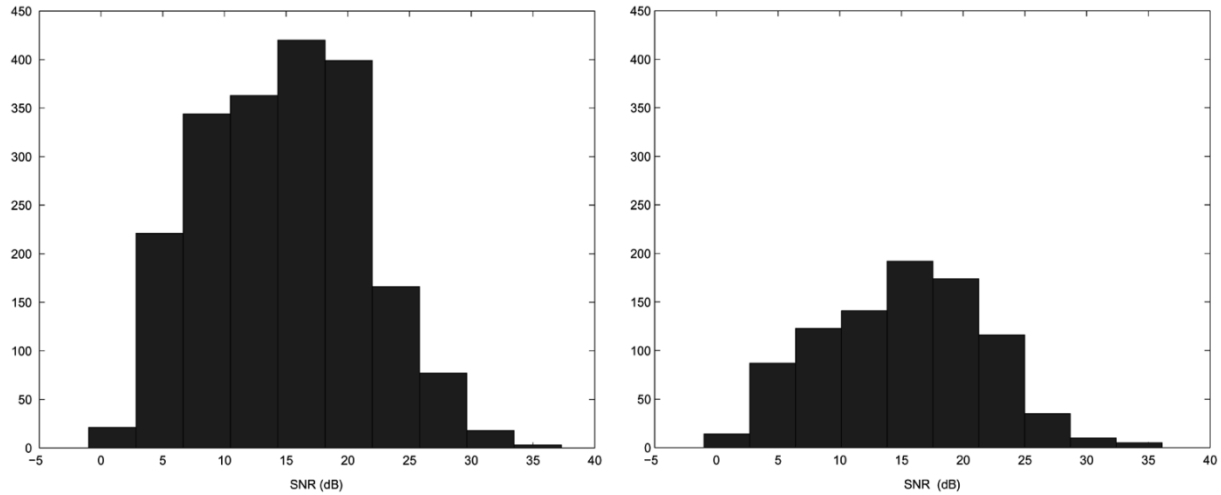


Fig. 6. Histograms of estimated utterance SNRs in training (left) and testing (right) of the well-matched condition of the Aurora 3 German database.

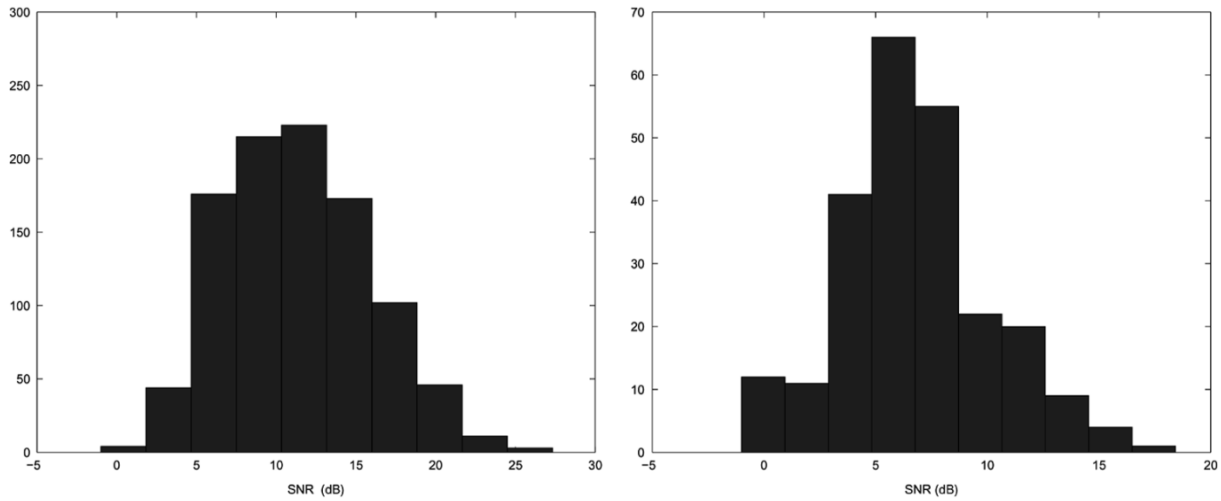


Fig. 7. Histograms of estimated utterance SNRs in training (left) and testing (right) of the medium-mismatched condition of the Aurora 3 German database.

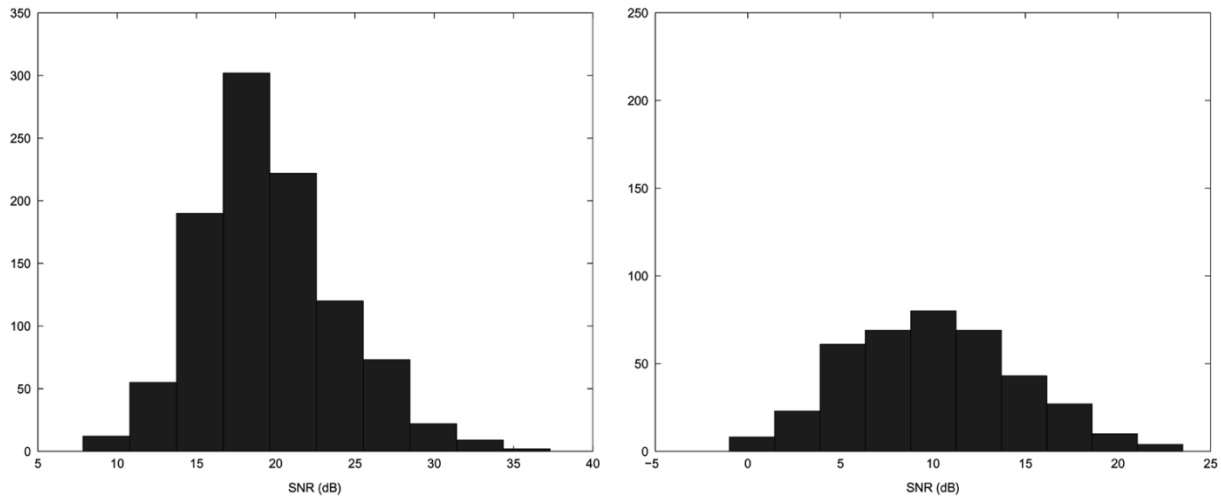


Fig. 8. Histograms of estimated utterance SNRs in training (left) and testing (right) of the high-mismatched condition of the Aurora 3 German database.

parameter estimation may occur. Table I shows the FC performance with respect to different regression polynomial orders. The state-tied regression polynomials are estimated from 300 utterances from Sets A and B (each) in the Aurora 2 database.

For the special case when the polynomial order is 0, it is equal to a state-based SNR independent bias removal method. From the table, performance improves when the polynomial order increases from 0 to 2. Third order polynomials give slightly worse

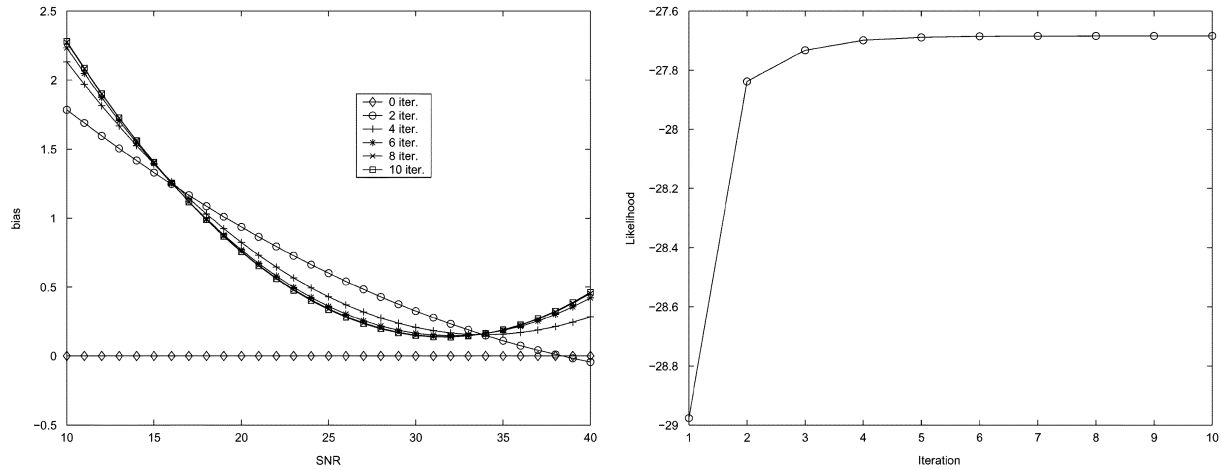


Fig. 9. Left panel shows estimated global polynomials as a function of SNR and iteration number. The right panel shows average likelihood of 50 utterances as a function of the number of EM iterations. Both panels use the energy feature component (E) for the airport noise data.

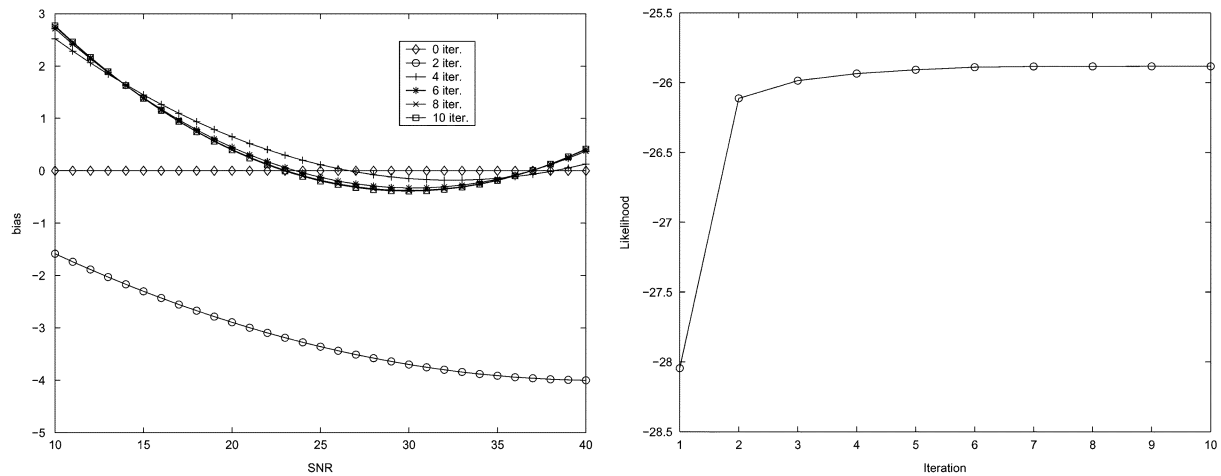


Fig. 10. Left panel shows estimated global polynomials as a function of SNR and iteration number. The right panel shows average likelihood of 50 utterances as a function of the number of EM iterations. Both panels use the energy feature component (E) for the station noise data.

TABLE I
AVERAGE WORD RECOGNITION ACCURACY (%) FOR SETS A AND B AN AURORA 2 WITH RESPECT TO POLYNOMIAL ORDER. THE POLYNOMIALS ARE STATE TIED AND ESTIMATED FROM 300 UTTERANCES

Data Sets	Polynomial Order			
	0	1	2	3
Set A	83.0	83.5	83.8	83.7
Set B	83.1	83.5	83.9	83.4

performance due to a larger number of parameters to be estimated. Considering the approximation goodness and number of parameters, first and second order polynomials are chosen for the experiments.

D. Estimated Regression Polynomials

The effectiveness of regression polynomial fitting to noisy speech data can be observed from the change of the average likelihood of the adaptation utterances. The effect of the number of EM iterations is illustrated in Figs. 9 and 10, the figures show

average likelihood of 50 utterances using one to ten EM iterations in the right panels for airport and station noise, respectively. The corresponding polynomials of the energy component of the features are illustrated on the left. The polynomials are shared globally and zero polynomials are chosen as the initial conditions under which no feature compensation is performed. From the figures, a significant increase of the average likelihood can be observed after the first iteration which is attributed to the feature compensation beginning to take effect and the regression polynomials change from zero polynomials to nonzero polynomials. Afterwards, the average likelihood increases monotonically until it converges to a stationary point. The monotonicity and convergence is guaranteed by the EM algorithm. Fast convergence of the FC algorithm can be observed in the figures. Typically, the algorithm converges after three or four iterations. The increase of the likelihood indicates a better fit of the features to the original models after the compensation.

The effect of the number of utterances is illustrated in Figs. 11 and 12 where the global polynomials are estimated for six EM iterations for a variety of feature components with different numbers of adaptation utterances with car and subway background noise. The shape of the polynomials varies dramatically when

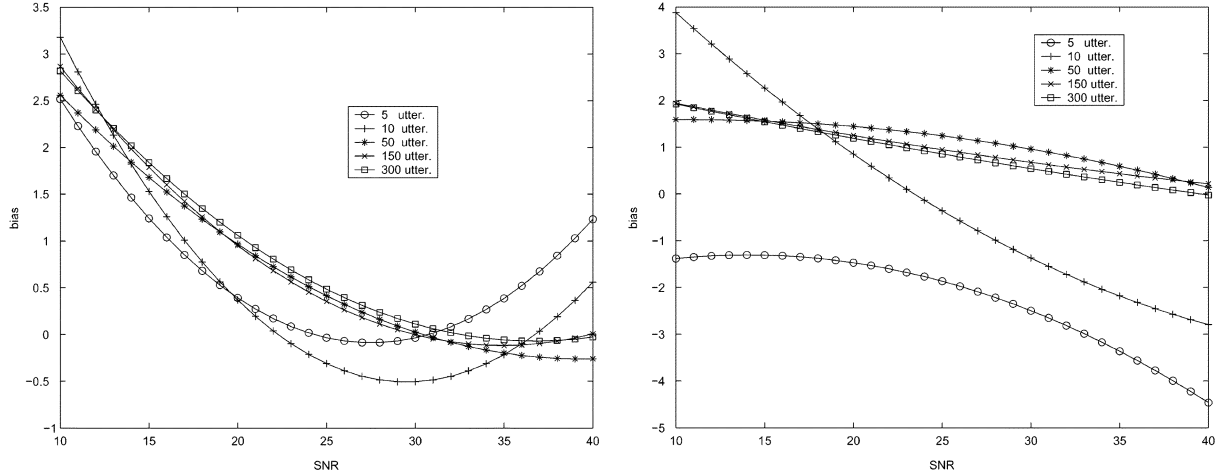


Fig. 11. Left and right panels show estimated global polynomials as a function of SNR and number of utterances for energy (E) and the first cepstral coefficient (C1), respectively, under car noise. The number of EM iterations is fixed at six.

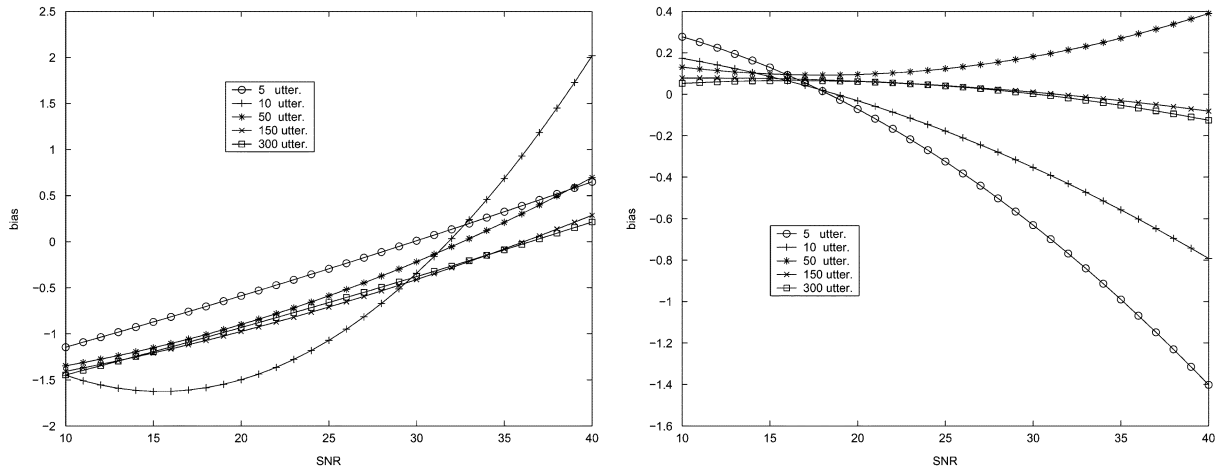


Fig. 12. Left and right panels show estimated global polynomials as a function of SNR and number of utterances for the sixth (C6) and tenth (C10) cepstral coefficients, respectively, under subway noise. The number of EM iterations is fixed at six.

the number of adaptation utterances is small. As the adaptation data size grows, the polynomials become stable since the statistics collected for polynomial estimation become more robust.

As observed from the estimated regression polynomials, biases exist under clean conditions (e.g., SNR > 20 dB). This will degrade the performance for clean speech to a certain degree. Table II shows the baseline (zero utterances) and FC performance averaged over all the clean conditions in the Aurora 2 database. Ten, 100, and 200 adaptation utterances are utilized which tie the polynomials at the global, state and mixture levels, respectively. Compared with the baseline, the FC algorithm degrades recognition performance. Therefore, in the decoding stage of the following experiments on the Aurora 2 database, no compensation is performed for SNRs higher than 20 dB.

E. Recognition Results

1) *Aurora 2 Database:* Comparative experiments are performed using the proposed feature compensation (FC) algorithm and the maximum likelihood linear regression (MLLR) algorithm. Different amounts of adaptation data are used ranging

TABLE II
WORD RECOGNITION ACCURACY (%) AVERAGED OVER ALL CLEAN CONDITIONS IN THE AURORA 2 DATABASE. FEATURE COMPENSATION IS PERFORMED WITH TEN, 100, AND 200 UTTERANCES. CLEARLY, USING FC WITH CLEAN DATA DEGRADES PERFORMANCE

Number of utterances	0	10	100	200
Accuracy(%)	99.0	98.8	97.4	97.6

from five to 300 utterances for each type of noise. Adaptation utterances are randomly chosen from Sets A and B for each type of noise and are excluded from the testing. The average length for each utterance is 4.6 digits. Regression polynomials are noise dependent and depending on the adaptation data size, they are tied at different levels of granularity: for adaptation sizes of five, 10, and 20 utterances, polynomials are tied globally; for adaptation sizes of 50, 100, and 150 utterances, polynomials are tied within states; for adaptation sizes of 200, 250, and 300 utterances, polynomials are Gaussian mixture specific. The transformations of MLLR are also tied in a similar manner. In the experiments, two MLLR schemes are tested—the transformation matrices are estimated by pooling all the adaptation data

TABLE III

WORD RECOGNITION ACCURACY (%) FOR FC AND MLLR FOR 4 TYPES OF NOISE IN SET A OF THE AURORA 2 DATABASE. MLLR1 REFERS TO THE CASE WHERE THE MLLR TRANSFORMATION MATRICES ARE ESTIMATED ACROSS ALL SNR LEVELS, WHILE MLLR2 REFERS TO MLLR TRANSFORMATION MATRICES BEING SNR-CLUSTER SPECIFIC. BASELINE MFCC RESULTS ARE PRESENTED AS ADAPTATION WITH ZERO UTTERANCES

Noise Type	Algorithm	Number of Utterances									
		0	5	10	20	50	100	150	200	250	300
subway	MLLR1	74.5	73.2	76.0	75.5	75.8	81.2	82.8	83.6	83.5	84.0
	MLLR2	74.5	70.4	77.5	77.5	78.3	81.8	81.9	82.7	83.5	82.4
	FC	74.5	79.0	78.9	79.7	81.5	84.0	84.8	85.4	87.2	87.4
babble	MLLR1	58.1	67.0	68.9	73.0	75.5	74.1	76.6	76.0	76.5	78.3
	MLLR2	58.1	70.7	64.5	69.4	74.1	73.6	74.8	75.4	76.9	75.8
	FC	58.1	71.4	71.8	74.9	81.6	83.7	84.7	85.2	85.9	86.5
car	MLLR1	70.0	70.9	70.0	73.5	75.9	77.8	78.9	80.4	79.8	80.5
	MLLR2	70.0	69.5	70.6	75.3	81.7	79.9	80.6	79.7	79.3	81.3
	FC	70.0	71.7	74.5	74.4	77.8	79.6	80.2	81.0	82.9	83.2
exhibition	MLLR1	71.0	73.3	73.9	72.2	72.9	76.9	78.5	79.3	79.5	81.0
	MLLR2	71.0	69.5	75.2	74.7	79.7	76.5	77.1	76.0	74.8	75.4
	FC	71.0	73.7	76.6	75.6	76.8	80.2	81.5	82.1	84.5	85.4

TABLE IV

WORD RECOGNITION ACCURACY (%) FOR FC AND MLLR ON 4 TYPES OF NOISE IN SET B OF THE AURORA 2 DATABASE. SEE Table III CAPTION FOR THE DEFINITION OF MLLR1 AND MLLR2. BASELINE MFCC RESULTS ARE PRESENTED AS ADAPTATION WITH ZERO UTTERANCES

Noise Type	Algorithm	Number of Utterances									
		0	5	10	20	50	100	150	200	250	300
restaurant	MLLR1	60.3	70.6	70.3	70.5	76.9	78.8	79.1	79.9	80.6	81.1
	MLLR2	60.3	66.3	78.2	75.2	78.7	80.1	80.6	79.9	77.3	79.6
	FC	60.3	72.0	74.1	75.6	82.9	83.5	86.6	87.1	88.2	88.4
street	MLLR1	67.8	68.7	77.1	74.4	78.8	78.3	80.1	80.2	80.7	82.3
	MLLR2	67.8	70.4	69.2	75.7	81.3	82.7	83.4	83.8	78.6	84.2
	FC	67.8	74.8	75.3	74.6	80.4	82.5	83.2	83.9	84.2	85.1
airport	MLLR1	60.9	73.8	75.3	74.2	76.1	78.7	80.5	81.1	81.9	83.1
	MLLR2	60.9	68.5	75.3	75.7	79.5	83.4	84.0	83.8	80.1	84.0
	FC	60.9	76.1	77.1	78.3	83.4	85.0	86.0	86.9	87.4	88.1
station	MLLR1	62.9	68.3	67.5	71.6	74.6	76.9	77.3	77.3	77.5	79.1
	MLLR2	62.9	71.7	75.2	74.7	69.4	80.4	81.1	80.9	75.3	80.7
	FC	62.9	71.7	76.0	76.0	79.0	81.2	82.0	82.8	83.5	84.4

across SNR levels, and by distinct SNR levels, respectively. For the second case, considering the SNR distinction and amount of data for robust estimation, three SNR clusters are employed: clean and 20 dB, 15 dB and 10 dB, and 5 dB and 0 dB. Different MLLR transformation matrices are estimated for three different SNR clusters and used to evaluate the corresponding SNR level utterances. For both MLLR implementations, the transformation matrices are three-block diagonal. The regression polynomial order is set to two.

Performance of FC and MLLR for each type of noise is presented in Table III for Set A, and in Table IV for Set B. In the tables, MLLR1 denotes the case where MLLR transformation matrices are estimated by all of the adaptation data and MLLR2 denotes the SNR-cluster specific case. The results are averaged across the 6 SNR levels (clean, 20 dB, 15 dB, 10 dB, 5 dB, and 0 dB) tested.

In most cases, MLLR2 outperforms MLLR1 when the number of adaptation utterances is small, e.g., five to 150

TABLE V

WORD RECOGNITION ACCURACY (%) FOR FC AND MLLR FOR EACH SNR LEVEL IN SET A OF THE AURORA 2 DATABASE. SEE Table III CAPTION FOR THE DEFINITION OF MLLR1 AND MLLR2. BASELINE MFCC RESULTS ARE PRESENTED AS ADAPTATION WITH ZERO UTTERANCES

SNR	Algorithm	Number of Utterances									
		0	5	10	20	50	100	150	200	250	300
clean	MLLR1	99.0	95.9	96.5	96.5	97.0	97.1	97.2	97.6	97.7	97.2
	MLLR2	99.0	97.8	98.7	98.8	98.7	98.9	98.9	99.0	98.6	98.9
	FC	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0
20 dB	MLLR1	95.3	92.8	94.1	93.5	94.3	95.4	95.7	95.8	95.9	96.2
	MLLR2	95.3	92.9	94.2	95.9	94.6	95.0	95.3	95.2	94.5	94.9
	FC	95.3	96.3	96.4	96.8	96.9	96.6	97.0	97.2	97.0	97.3
15 dB	MLLR1	87.5	86.6	89.0	89.4	89.5	91.8	92.8	92.8	93.1	93.6
	MLLR2	87.5	87.9	90.4	91.0	93.1	93.6	94.3	94.5	92.6	94.5
	FC	87.5	92.4	93.2	93.3	94.4	95.1	95.5	95.8	95.6	95.9
10 dB	MLLR1	67.7	72.5	76.4	78.7	78.9	82.7	84.9	85.1	85.6	87.0
	MLLR2	67.7	76.7	80.4	82.4	85.8	86.2	87.4	88.1	83.9	87.8
	FC	67.7	79.7	80.9	79.5	88.5	90.2	90.8	91.0	91.7	91.8
5 dB	MLLR1	39.5	51.9	52.6	57.7	60.4	63.7	68.8	69.8	70.3	72.8
	MLLR2	39.5	41.8	44.2	53.9	62.9	61.6	62.8	61.6	67.2	63.4
	FC	39.5	51.6	55.6	58.1	66.4	72.6	74.4	75.5	80.0	80.4
0 dB	MLLR1	17.0	26.8	24.7	25.4	30.0	34.2	36.0	37.9	36.4	39.0
	MLLR2	17.0	23.1	24.1	27.8	35.5	32.3	32.8	32.5	34.8	33.4
	FC	17.0	24.7	27.9	30.1	31.3	37.2	40.1	42.1	47.4	49.5

utterances, and not as good as MLLR1 when the adaptation data set has more than 200 utterances. This is due to the number of parameters to be estimated for the transformation matrices. When the adaptation data are limited, transformation matrices are tied in a larger scale, e.g., global or within states. On the other hand, mixture-specific matrices are employed when the number of utterances is higher than 200. In this case, the number of parameters to be estimated becomes larger and the transformation matrices are estimated less robustly.

Both MLLR and FC obtain improved performance with the growth of the adaptation data size. Compared with MLLR1, FC results in 17.3%, 26.9%, 8.4%, and 14.2% WER reduction for subway, babble, car and exhibition noise. On average, FC outperforms MLLR1 by 16.7% for Set A. Compared with MLLR2, FC results in 17.2%, 30.1%, 2.9%, and 16.5% WER reduction for the four types of noise in Set A, and overall improvement is 16.5%. Considering the restaurant, street, airport and station noise in Set B, FC obtains 25.8%, 20.1%, 23.4%, and 20.6% WER reduction over MLLR1, and 21.6%, 16.1%, 18.3%, and 12.3% over MLLR2. On average, FC outperforms MLLR1 by 20.5% and MLLR2 by 14.6% for Set B over all conditions. Among the eight types of noise in Sets A and B, FC achieves the highest improvement over MLLR in babble noise and least improvement with car noise.

Tables V and VI summarize the performance of FC and MLLR under each SNR condition for Sets A and B, respectively. For each set, the results are averaged over the 4 types of noise in the set. On average, FC achieves a WER reduction of 33.1% for MLLR1, 23.6% for MLLR2 for Set A, and 34.5% for MLLR1, 21.4% for MLLR2 for Set B. Among the six SNR levels, FC results in best performance for SNRs arranging from 10 to 20 dB, compared to both MLLR1 and MLLR2.

TABLE VI
WORD RECOGNITION ACCURACY (%) FOR FC AND MLLR FOR EACH SNR LEVEL IN SET B OF THE AURORA 2 DATABASE. SEE TABLE III CAPTION FOR THE DEFINITION OF MLLR1 AND MLLR2. BASELINE MFCC RESULTS ARE PRESENTED AS ADAPTATION WITH ZERO UTTERANCES

SNR	Algorithm	Number of Utterances									
		0	5	10	20	50	100	150	200	250	300
clean	MLLR1	99.0	95.4	96.9	94.9	97.3	97.5	97.5	97.7	97.9	97.5
	MLLR2	99.0	97.6	98.6	98.8	98.7	99.0	99.0	99.1	98.2	99.0
	FC	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0
20 dB	MLLR1	92.8	93.7	94.5	92.9	95.3	96.2	96.4	96.4	96.6	96.8
	MLLR2	92.8	93.0	94.9	93.8	95.3	96.0	95.8	95.9	95.9	95.9
	FC	92.8	96.7	97.1	97.2	96.8	97.1	97.3	97.6	97.4	97.7
15 dB	MLLR1	81.3	87.9	90.9	88.1	90.2	92.7	93.2	93.3	93.5	94.3
	MLLR2	81.3	89.3	92.6	93.4	93.4	94.9	95.1	94.9	91.6	95.2
	FC	81.3	93.0	93.0	93.6	95.1	96.0	96.0	96.2	95.9	96.4
10 dB	MLLR1	59.0	74.9	81.9	76.1	78.4	84.2	85.6	85.8	86.4	88.0
	MLLR2	59.0	78.7	85.1	86.4	86.1	89.9	90.5	90.5	82.1	90.6
	FC	59.0	80.0	80.7	82.0	90.8	91.8	92.3	92.5	92.7	92.9
5 dB	MLLR1	31.9	52.7	59.9	54.1	57.9	65.4	67.7	68.4	69.3	72.2
	MLLR2	31.9	39.1	50.5	53.9	57.8	69.1	70.8	70.0	64.1	69.4
	FC	31.9	49.2	54.5	56.7	71.8	76.7	78.0	79.2	80.9	82.0
0 dB	MLLR1	13.7	25.8	26.7	25.2	30.1	33.0	35.1	36.1	37.2	39.6
	MLLR2	13.7	17.5	27.1	26.7	32.1	41.2	42.4	42.1	35.2	42.5
	FC	13.7	20.5	24.6	28.2	34.9	37.7	44.1	46.5	49.0	51.1

TABLE VII
WORD RECOGNITION ACCURACY (%) FOR FC AND MLLR OF STATIC-TYING SCHEMES UNDER HIGH-MISMATCHED (HM), MEDIUM-MISMATCHED (MM) AND WELL-MATCHED (WM) CONDITIONS OF THE GERMAN PART OF THE AURORA 3 DATABASE. BASELINE MFCC RESULTS ARE PRESENTED AS ADAPTATION WITH ZERO UTTERANCES

Conditions	Algorithm	Number of Utterances			
		0	10	40	70
HM	MLLR	74.3	81.6	83.6	85.8
	FC	74.3	78.8	85.7	86.6
MM	MLLR	79.1	80.2	79.7	81.1
	FC	79.1	78.3	81.5	84.4
WM	MLLR	90.6	90.6	88.8	90.9
	FC	90.6	91.2	91.0	91.4

In Tables V and VI, no compensation is performed on speech features for utterances with SNRs higher than 20 dB.

2) *Aurora 3 Database:* For the German part of the Aurora 3 database, 10, 40, and 70 utterances are randomly chosen for adaptation and excluded from each testing condition. Table VII shows the performance of MLLR performance and the proposed FC algorithm. As before, the FC algorithm outperforms MLLR. To investigate whether dynamically grouping the Gaussians in the model set is beneficial, a regression tree is constructed for each condition (WM, MM, and HM) with each mixture being a leaf utilizing the originally trained acoustic HMMs. Instead of specifying static-tying classes explicitly in advance as static-tying strategies, a regression tree is created based on the centroid splitting algorithm using the Euclidean distance between the Gaussian mixture means as described in [26]. Depending on the amount of adaptation available, a regression tree can tie

TABLE VIII
WORD RECOGNITION ACCURACY (%) FOR FC(r) AND MLLR(r) OF DYNAMIC-TYING SCHEMES UNDER HIGH-MISMATCHED (HM), MEDIUM-MISMATCHED (MM) AND WELL-MATCHED (WM) CONDITIONS OF THE GERMAN PART OF THE AURORA 3 DATABASE. BASELINE MFCC RESULTS ARE PRESENTED AS ADAPTATION WITH ZERO UTTERANCES

Conditions	Algorithm	Number of Utterances			
		0	10	40	70
HM	MLLR(r)	74.3	80.7	84.8	86.0
	FC(r)	74.3	82.8	87.2	88.4
MM	MLLR(r)	79.1	80.2	80.8	81.3
	FC(r)	79.1	78.1	81.7	84.1
WM	MLLR(r)	90.6	89.1	90.4	91.0
	FC(r)	90.6	91.4	91.9	92.0

parameters (transformations in MLLR or polynomials in FC) at different levels dynamically. To tie the parameters, MLLR needs node occupations as well as a minimum number of mixtures (larger than the feature vector size) within one class for matrix inversion while FC needs node occupations only. In the experiments, the node occupation threshold is set to 700 and three-block transformation matrices are used. Table VIII shows the performance of MLLR (MLLR(r)) and FC (FC(r)) using the dynamic tying strategy. From the table, FC(r) gives better results than MLLR(r) and static tying FC under most cases. On average, FC(r) obtains performance improvements over the baseline (no adaptation) by 12.4%, 12.1%, and 46.1% for WM, MM, and HM conditions, respectively. The most significant improvement is for the HM condition where the original acoustic models are trained by relatively clean speech and a large mismatch exists between the training and testing environments. Although the FC algorithm is originally used to approximate the bias between clean and noisy speech, improvements are still observed for the MM and WM conditions where the original HMMs are trained with noisy speech. The improvements of FC(r) over MLLR(r) are 15.9%, 3.0%, and 14.6% for the three testing conditions, respectively.

V. CONCLUSION

A feature compensation algorithm based on polynomial regression of utterance signal-to-noise ratio (SNR) for noise-robust speech recognition is proposed and implemented. In this algorithm, a set of polynomials regressed on utterance SNRs are utilized to approximate the bias between the clean and noisy speech features. The bias is used to compensate noisy features with respect to clean acoustic models. The maximum likelihood estimation of the regression polynomials is provided within an EM algorithm framework.

ASR experiments are performed on the Aurora 2 and the German part of the Aurora 3 databases. Using the Aurora 2 database, eight types of noise in Sets A and B are tested with different adaptation data sizes. Experiments are designed to compare the performance of the feature compensation algorithm with two MLLR implementations: transformation matrices estimated by pooling all the adaptation data, and by distinct SNR clusters. Significant improvements over the two MLLR schemes

are observed for Sets A and B. The evaluation of the algorithm on the German part of the Aurora 3 database shows improvements under the well-matched, medium-mismatched and high-mismatched testing conditions. The most significant improvement is observed for the high-mismatched case.

REFERENCES

- [1] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Commun.*, vol. 16, pp. 261–291, 1995.
- [2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.
- [3] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, 1974.
- [4] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Norwell, MA: Kluwer, 1992.
- [5] L. Deng, J. Droppo, and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 3, pp. 218–233, May 2004.
- [6] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 578–589, 1994.
- [7] H. Hermansky, "Perceptual linear prediction (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.
- [8] D. Macho, L. Mauuary, B. Noe, Y. Cheng, D. Ealey, D. Jouviet, H. Kelleher, D. Perace, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on aurora databases," in *Proc. Int. Conf. Spoken Language Processing*, 2002, pp. 17–20.
- [9] X. Cui, M. Iseli, Q. Zhu, and A. Alwan, "Evaluation of noise robust features on the Aurora databases," in *Proc. Int. Conf. Spoken Language Processing*, 2002, pp. 481–484.
- [10] B. Stroppe and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, pp. 451–464, 1997.
- [11] Q. Zhu and A. Alwan, "Amplitude demodulation of speech and its application to noise robust speech recognition," in *Proc. Int. Conf. Spoken Language Processing*, 2000, pp. 341–344.
- [12] —, "On the use of variable frame rate analysis in speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2000, pp. 1783–1786.
- [13] M. Gales and S. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Process.*, vol. 4, pp. 352–359, 1996.
- [14] P. Moreno, "Speech Recognition in Noisy Environments," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, 1996.
- [15] B. Raj, E. Gouvea, P. Moreno, and R. Stern, "Cepstral compensation by polynomial approximation for environment-independent speech recognition," in *Proc. Int. Conf. Spoken Language Processing*, 1996, pp. 2340–2343.
- [16] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian approach to fast acoustic model adaptation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1997, pp. 835–838.
- [17] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.
- [18] Z. Zhang and S. Furui, "Piecewise-linear transformation-based HMM adaptation for noisy speech," *Speech Commun.*, vol. 42, pp. 43–58, 2004.
- [19] X. Cui and Y. Gong, "Variable parameter Gaussian mixture hidden Markov modeling for speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, 2003, pp. 12–15.
- [20] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.
- [21] Y. Gong, "Noise-dependent Gaussian mixture classifiers for robust rejection decision," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 2, pp. 57–64, Mar. 2002.
- [22] L. Deng, J. Droppo, and A. Acero, "Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 3, pp. 133–143, May 2004.
- [23] Q. Zhu and A. Alwan, "The effect of additive noise on speech amplitude spectra: A quantitative analysis," *IEEE Signal Process. Lett.*, vol. 9, no. 9, pp. 275–277, Sep. 2002.
- [24] R. Martin, "An efficient algorithm to estimate instantaneous SNR of speech signals," in *Proc. Eur. Conf. Speech Communication Technology*, 1993, pp. 1093–1096.
- [25] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ASR2000 Int. Workshop on Automatic Speech Recognition*, 2000, pp. 181–188.
- [26] *The HTK Book (Version 3.1)*. Cambridge, U.K.: Cambridge Univ. Press, 2001.

Xiaodong Cui (S'01) received the B.S. degree (with highest honors) from Shanghai Jiao Tong University, Shanghai, China, in 1996 and the M.S. degree from Tsinghua University, Beijing, China, in 1999, both in electrical engineering. He is currently pursuing the Ph.D. degree in the Speech Processing and Auditory Perception Laboratory, Department of Electrical Engineering, University of California, Los Angeles.

In 2002 and 2003, he held internship positions in DSP Solutions R&D Center, Texas Instruments, Dallas, TX. His research interests include speech recognition, digital speech processing, statistical signal processing, and pattern recognition.

Abeer Alwan (SM'00) received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1992.

Since then, she has been with the Electrical Engineering Department at the University of California, Los Angeles, as an Assistant Professor (1992–1996), Associate Professor (1996–2000), Professor (2000–present), and Vice Chair (2003–present). She established and directs the Speech Processing and Auditory Perception Laboratory at UCLA (<http://www.icsl.ucla.edu/~spapl>). Her research interests include modeling human speech production and perception mechanisms and applying these models to speech-processing applications such as noise-robust automatic speech recognition, compression, and synthesis. She was the Editor-in-Chief of the *Journal of Speech Communication* from 2000 to 2003.

Dr. Alwan received the NSF Research Initiation Award (1993), the NIH FIRST Career Development Award (1994), the UCLA-TRW Excellence in Teaching Award (1994), the NSF Career Development Award (1995), and the Okawa Foundation Award in Telecommunications (1997). She is an elected member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the New York Academy of Sciences. She served as an elected member on the Acoustical Society of America Technical Committee on Speech Communication (1993–1999), on the IEEE Signal Processing Technical Committees on Audio and Electroacoustics (1996–2000), and on Speech Processing (1996–2001). She is a Fellow of the Acoustical Society of America.