Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR

Author : **Xiaodong Cui and Abeer Alwan**

Professor : 陳嘉平

Reporter : 楊治鏞

Outline

- Introduction
- Polynomial regression of SNR
- Utterance SNR estimation
- Experimental result

Introduction

- 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.

Polynomial regression of SNR

- Bias approximation by SNR
- Feature compensation
- ML estimation of regression polynomials

Bias approximation by SNR (1/4)

$$Y_k^{lin} = X_k^{lin} + N_k^{lin}$$

- 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.

Bias approximation by SNR (2/4)

$$Y_k^{\log} = X_k^{\log} + \log\left(1 + \frac{N_k^{lin}}{X_k^{lin}}\right) = X_k^{\log} + \log\left(1 + \frac{1}{SNR_k}\right)$$

$$= X_k^{\log} + g_k$$

$$g_k \triangleq \log\left(1 + \frac{1}{SNR_k}\right)$$

- Where $Y_k^{\log}$ and $X_k^{\log}$ represent noisy and clean speech in the log-power domain.

Bias approximation by SNR (3/4)

- DCT
- We get the nth cepstral coefficient as

$$Y_n^{cep} = \sum_k d_{nk} Y_k^{\log} = \sum_k d_{nk} X_k^{\log} + \sum_k d_{nk} g_k$$

$$Y_n^{cep} = X_n^{cep} + f_n\left(SNR\right)$$

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

Bias approximation by SNR (4/4)

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

- Where $\tilde{c}_{jn}$'s are the coefficients for the j th order item of the nth cepstrum.

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

Feature compensation (1/2)

- Assuming that the clean acoustic models are Gaussian mixture HMMs, the probability density function of observing feature o_t from state i is computed as

$$p\left(o_t \mid s_t = i\right) = \sum_k \alpha_{ik} b_{ik}(o_t)$$

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

Feature compensation (2/2)

$$p(o_t \mid s_t = i) = \sum_k \alpha_{ik} N\left(o_t - \sum_{j=0}^{p} c_{ikj}\eta^j ; \mu_{ik}, \Sigma_{ik} \right)$$

- 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. c_{ikj}'s are the coefficients of the regression polynomials of state i, mixture k and polynomial order j.

ML estimation of regression polynomials

$$Q_b\left(\lambda;\bar{\lambda}\right)=\sum_{r=1}^{R}\sum_{i\in\Omega_s}\sum_{k\in\Omega_m}\sum_{t=1}^{T^r}\gamma_t^r\left(i,k\right)\cdot\log b_{ik}\left(o_t^r\right)$$

- Where R is the utterance number of adaptation data and T^r is the frame number of the rth utterance. $\Omega_s=\{1,2,\dots,N\}$ and $\Omega_m=\{1,2,\dots,M\}$ are the state and mixture sets, respectively. $\gamma_t^r\left(i,k\right)=p\left(s_t^r=i,\xi_t^r=k\mid O^r,\bar{\lambda}\right)$ is the posterior probability of staying at state i mixture k at time t given the rth observation sequence $O^r=\{\mathrm{o}_1^r,\dots,\mathrm{o}_{T^r}^r\}$.

ML estimation of regression polynomials

- Optimizing $Q_b(\lambda;\bar{\lambda})$ with respect to c_{ikl}, one obtains

$$\frac{\partial Q_b(\lambda;\bar{\lambda})}{\partial c_{ikl}} = \frac{\partial}{\partial c_{ikl}} \sum_{r=1}^{R}\sum_{i=1}^{N}\sum_{k=1}^{M}\sum_{t=1}^{T^r} \gamma_t^r(i,k)\cdot \log \mathrm{N}\left(\mathrm{o}_\mathrm{t}^\mathrm{r} - \sum_{j=0}^{P} c_{ikj}\left(\eta^r\right)^j ; \mu_{ik}, \Sigma_{ik} \right)$$

$$= \frac{\partial}{\partial c_{ikl}} \sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_t^r(i,k) \times \left[-\frac{1}{2}\left(o_t^r - \sum_{j=0}^{P} c_{ikj}\left(\eta^r\right)^j - \mu_{ik} \right)^T \times \Sigma_{ik}^{-1} \left(o_t^r - \sum_{j=0}^{P} c_{ikj}\left(\eta^r\right)^j - \mu_{ik} \right) \right]$$

$$= \sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_t^r(i,k)\, \Sigma_{ik}^{-1} \cdot \left(o_t^r - \sum_{j=0}^{P} c_{ikj}\left(\eta^r\right)^j - \mu_{ik} \right) \cdot \left(\eta^r\right)^l = 0$$

$$l = 0,1,\ldots,P$$

ML estimation of regression polynomials

$$\sum_{j=0}^{P}\left[\sum_{r=1}^{R}\sum_{t=1}^{T^r}\gamma_t^r(i,k)\cdot\Sigma_{ik}^{-1}\cdot(\eta^r)^{j+l}\right]c_{ikj}=\sum_{r=1}^{R}\sum_{t=1}^{T^r}\gamma_t^r(i,k)\cdot\Sigma_{ik}^{-1}\cdot(o_t^r-\mu_{ik})\cdot(\eta^r)^l$$

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

$$\sum_{j=0}^{P}\psi(\eta^r,\eta^r,l,j)\cdot c_{ikj}=\psi(\eta^r,o_t^r-\mu_{ik},l,1)$$

ML estimation of regression polynomials

$$\mathbf{U}_{ik} \cdot c_{ik} = \mathbf{v}_{ik}$$

$$\mathbf{U}_{ik} = \left[\begin{pmatrix} \mu_{ik}(0,0) & \dots & \mu_{ik}(0,P) \\ \vdots & \ddots & \vdots \\ \mu_{ik}(P,0) & \cdots & \mu_{ik}(P,P) \end{pmatrix} \right]$$

$$\mu_{ik}(l,j) = \psi_{ik}(\eta^r, \eta^r, l, j)$$

$$c_{ik} = [c_{ik0}^T, \dots, c_{ikl}^T, \dots, c_{ikP}^T]^T$$

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

ML estimation of regression polynomials

- When the covariance matrices Σ_{ik} are diagonal (which is usually the case), the computational load could be significantly reduced.
- Suppose there are K classes $\{\omega_1, \omega_2, ..., \omega_K\}$ within which the regression polynomials of different Gaussian mixtures are shared.
- $c_{\omega_q l} \left(q = 1, \ldots, K \right)$

Utterance SNR estimation (1/2)

- Minimum statistics tracking algorithm
- Power spectral minimum statistics are searched within a 0.5 second interval preceding each speech frame.
- Factor 2.0

$$SNR(i) = 10 * \log_{10}\left(\frac{\overline{P}_x(i) - \min\left(factor * P_n(i), \overline{P}_x(i)\right)}{factor * P_n(i)} \right)$$

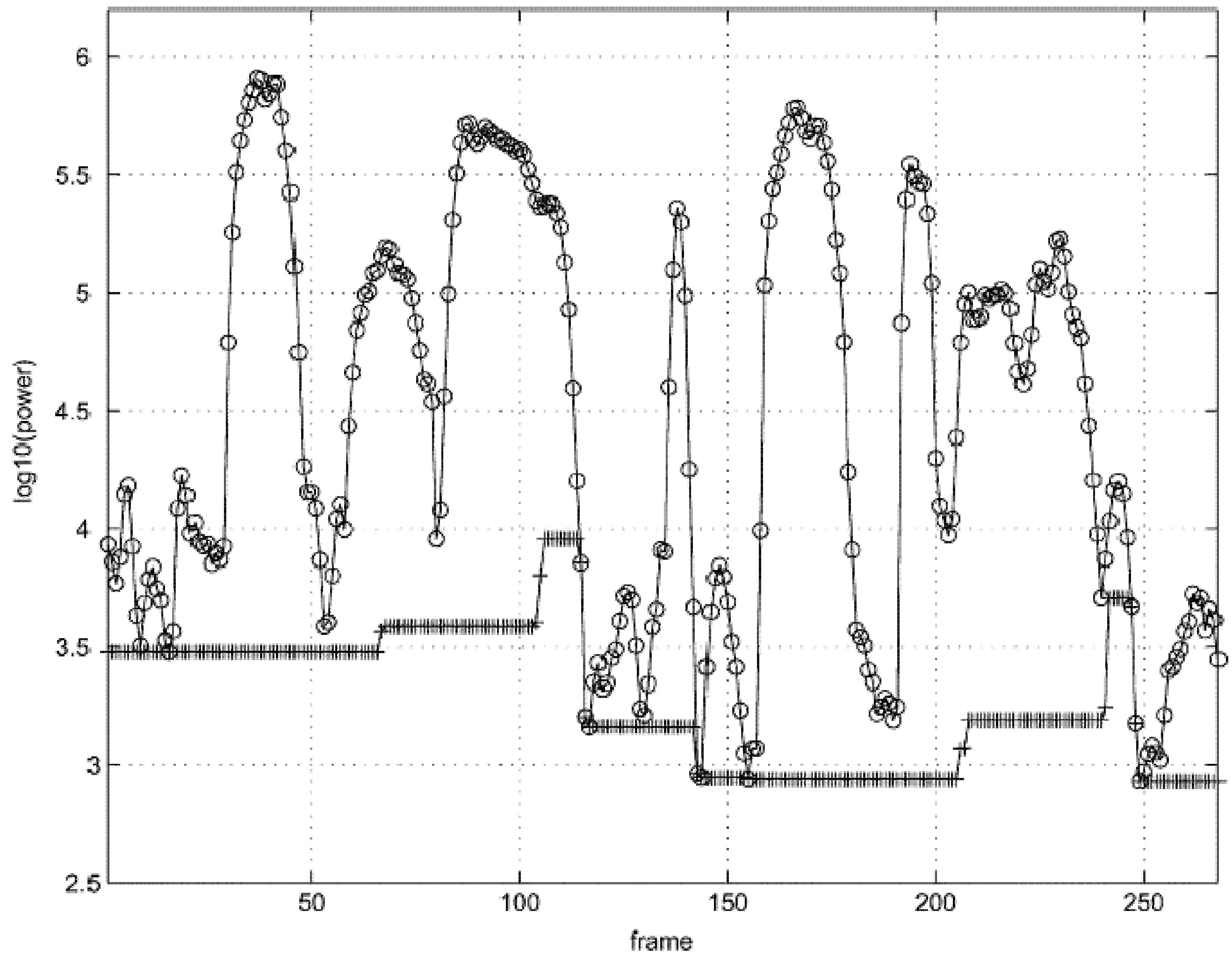

Fig. 1. Noise power(+) estimated by minimum statistics tracking from the noisy speech power spectrum (o) for the utterance "43o6571." The utterance is labeled 15 dB SNR in the Aurora 2 database.

Utterance SNR estimation (2/2)

- 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.

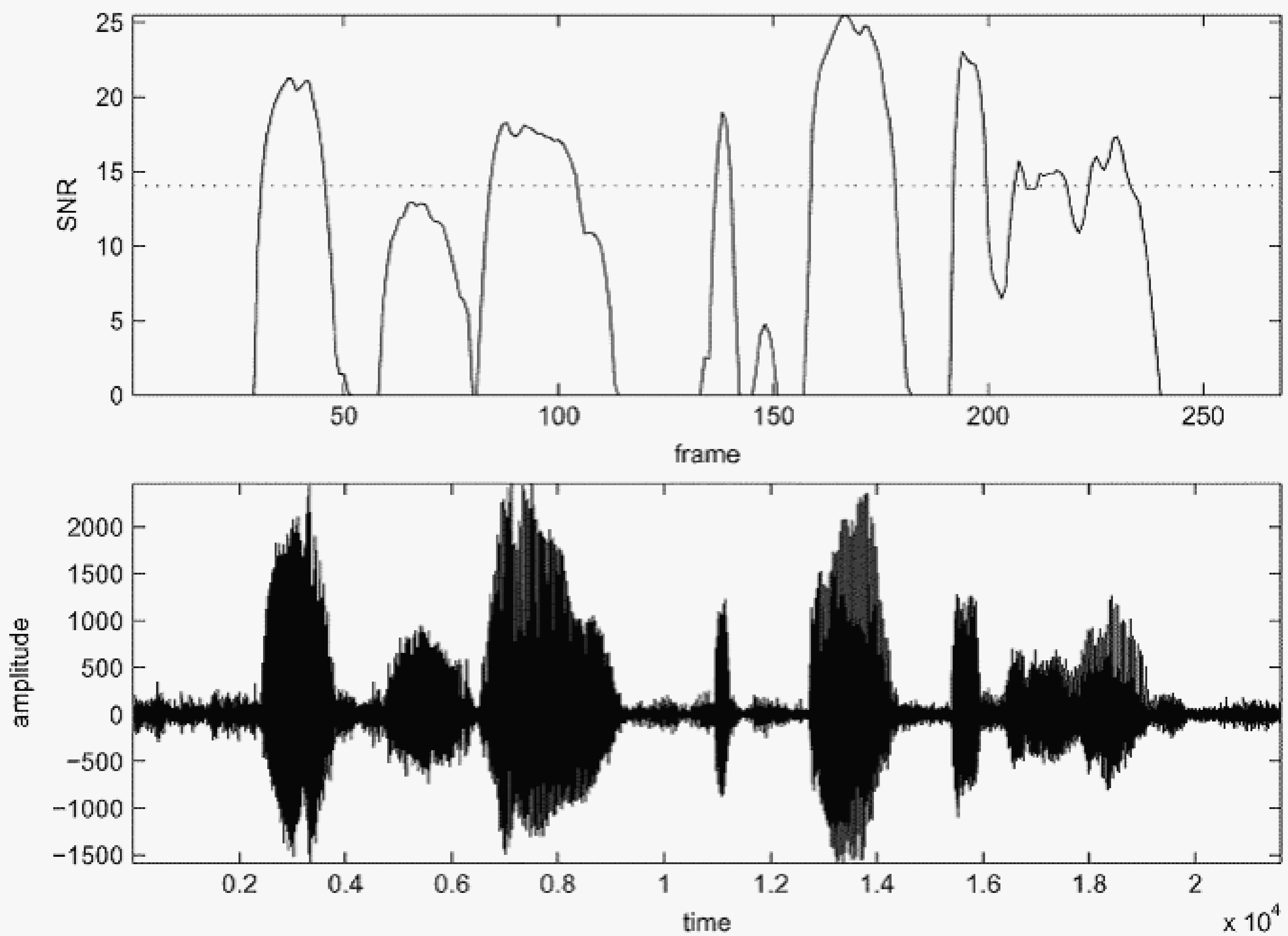

Fig. 2. Estimated frame-wise SNR (solid line in top panel), estimated utterance SNR (dotted line in top panel) and waveform (bottom panel) of the utterance "43o6571" labeled 15 dB SNR in the Aurora 2 database.

Experimental result

- Aurora 2
- HMMs
- All the Gaussian mixtures have diagonal covariance matrices.

Distribution of utterance SNRs

- Fig.4 and Fig.5.
- Small variances (2–3 dB) are observed in the estimated utterance SNRs for each condition in those figures.

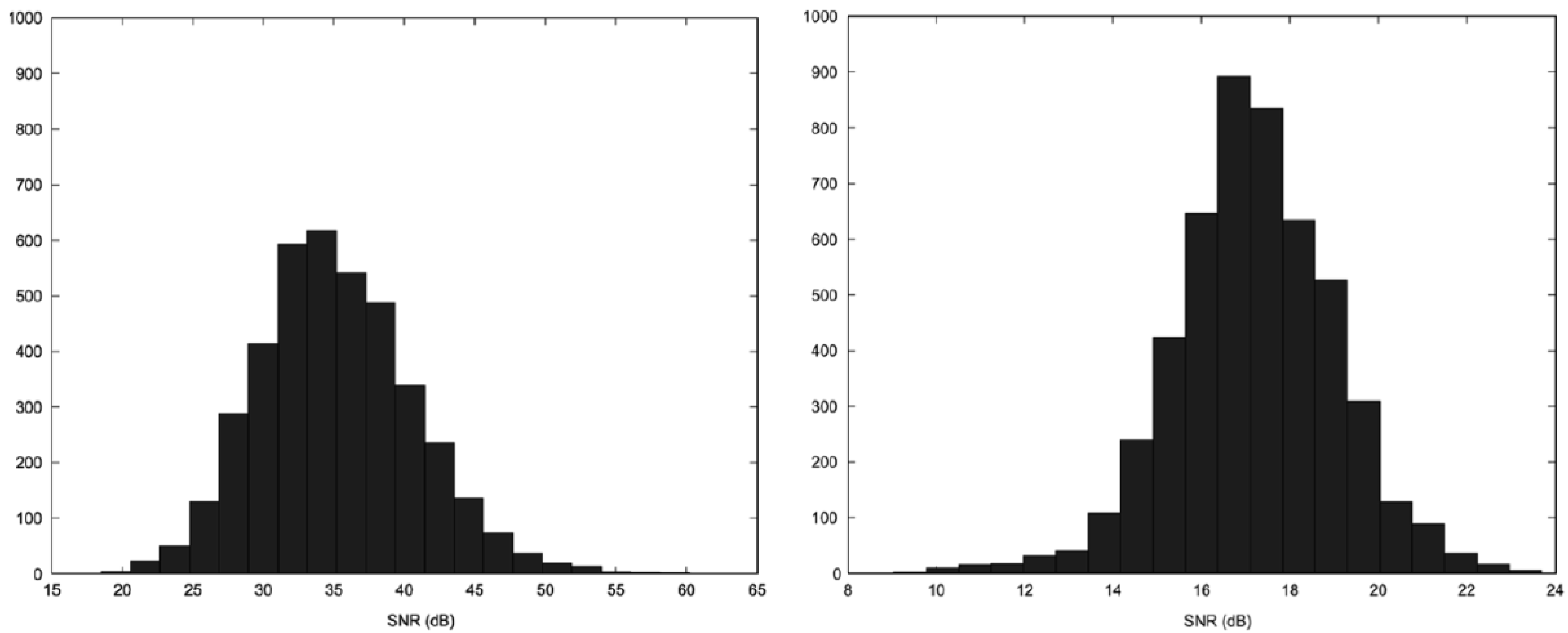

Fig. 4. Histograms of estimated utterance SNRs labeled as clean (left) and 20 dB SNR (right) data in Set A of Aurora 2 database.

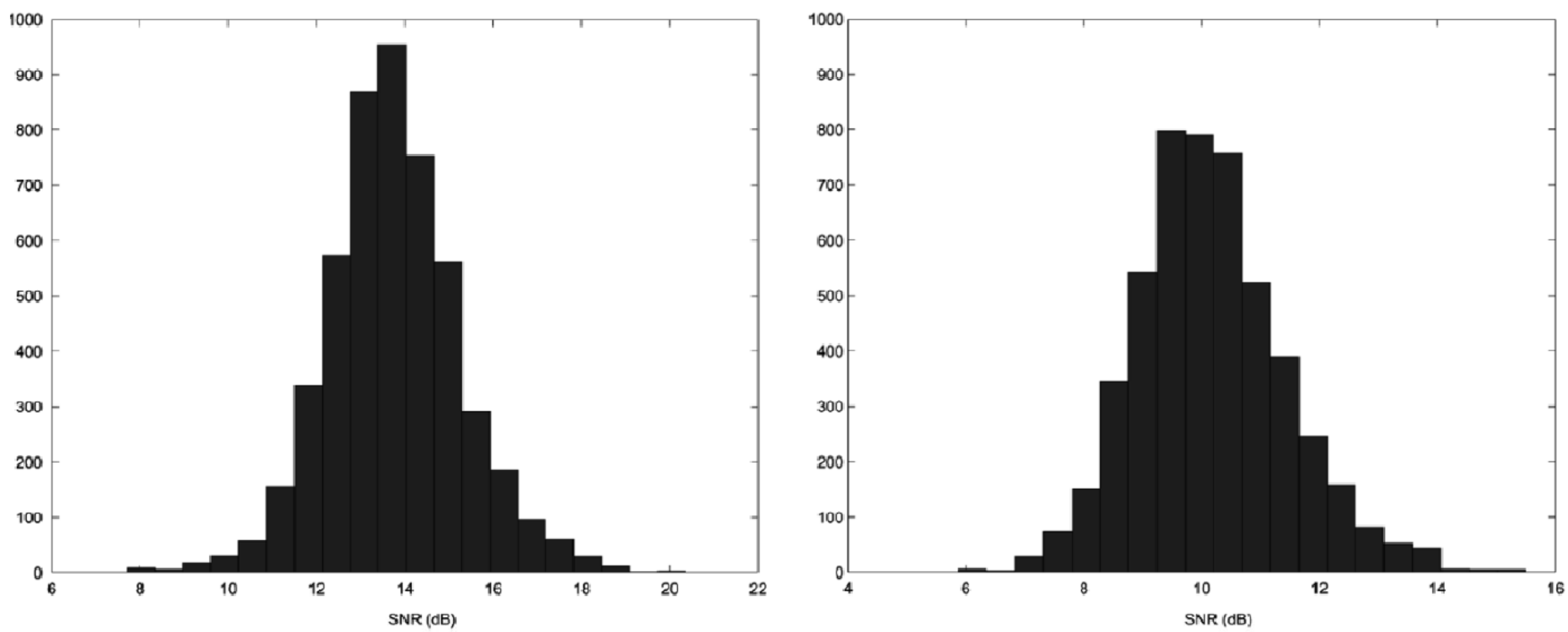

Fig. 5. Histograms of estimated utterance SNRs labeled as 15 dB (left) and 10 dB SNR (right) data in Set A of Aurora 2 database.

Regression polynomial orders

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

Estimated regression polynomials

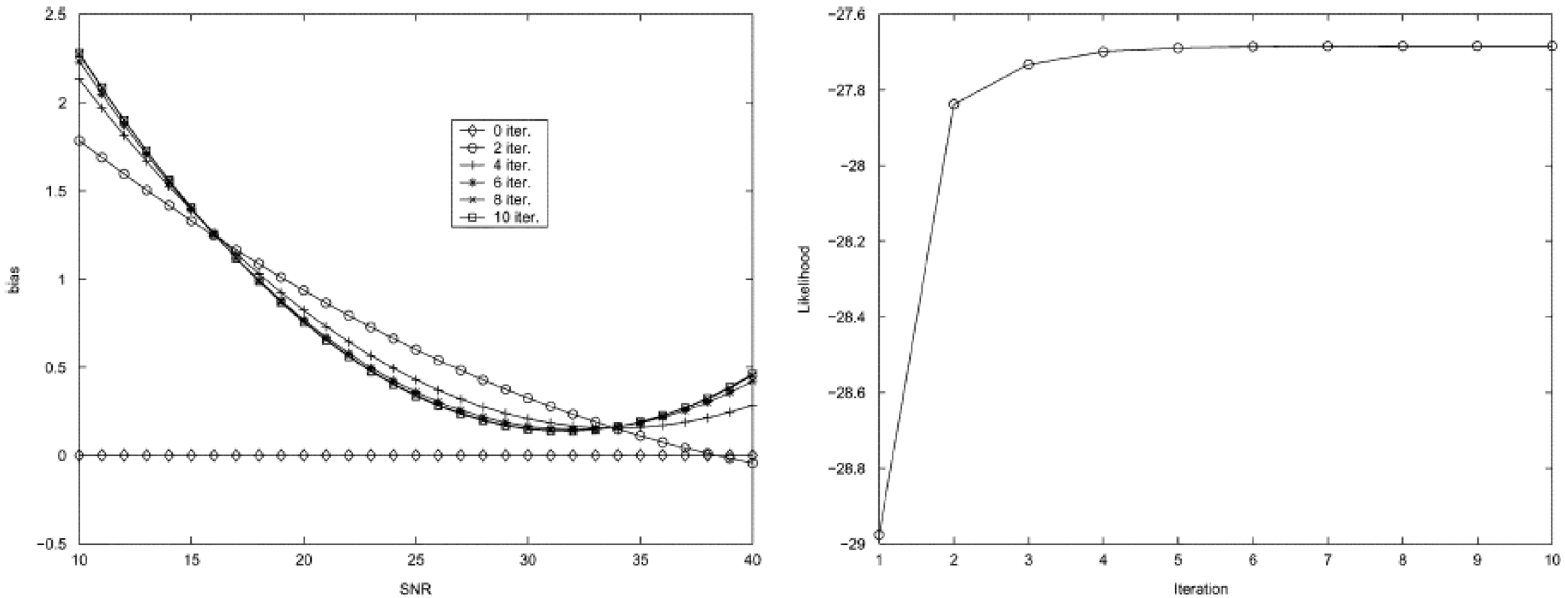

Fig. 9. Left panel shows estimated global polynomials as a function of SNR and iteration number. The right panel shows average likelihood of 50 utterances as a function of the number of EM iterations. Both panels use the energy feature component (E) for the airport noise data.

Estimated regression polynomials

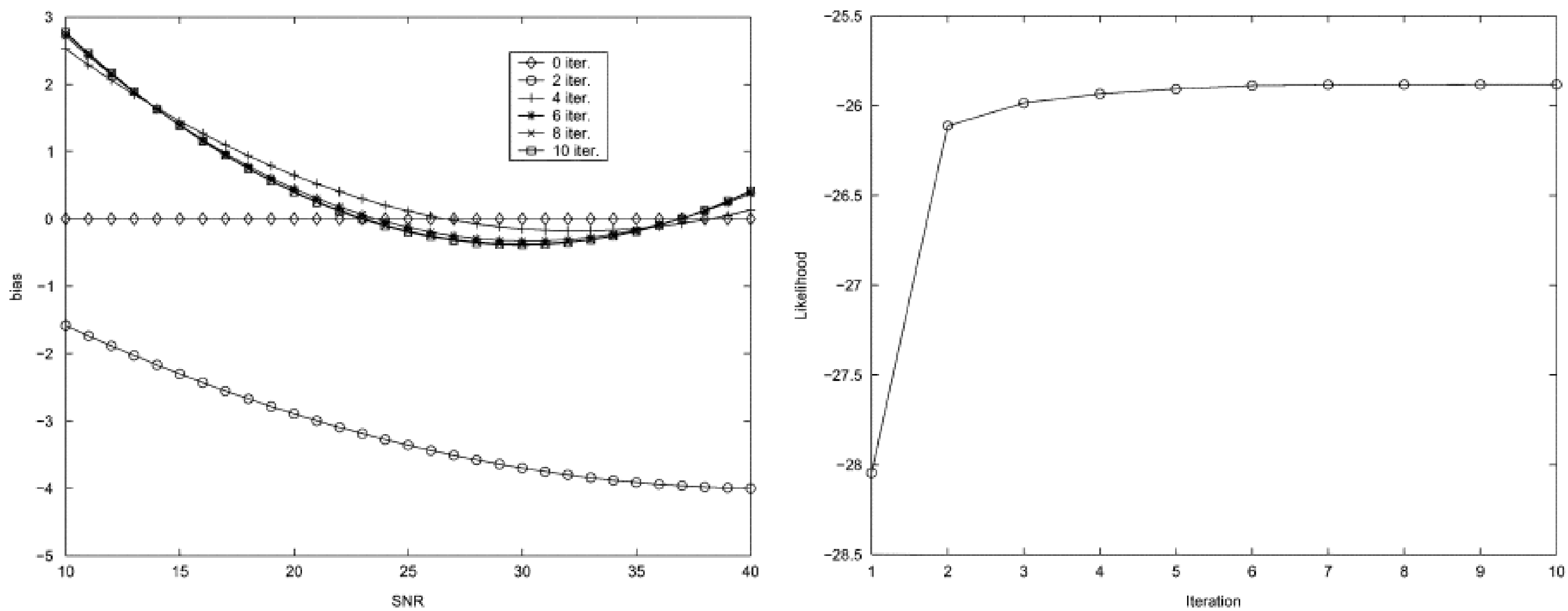

Fig. 10. Left panel shows estimated global polynomials as a function of SNR and iteration number. The right panel shows average likelihood of 50 utterances as a function of the number of EM iterations. Both panels use the energy feature component (E) for the station noise data.

Estimated regression polynomials

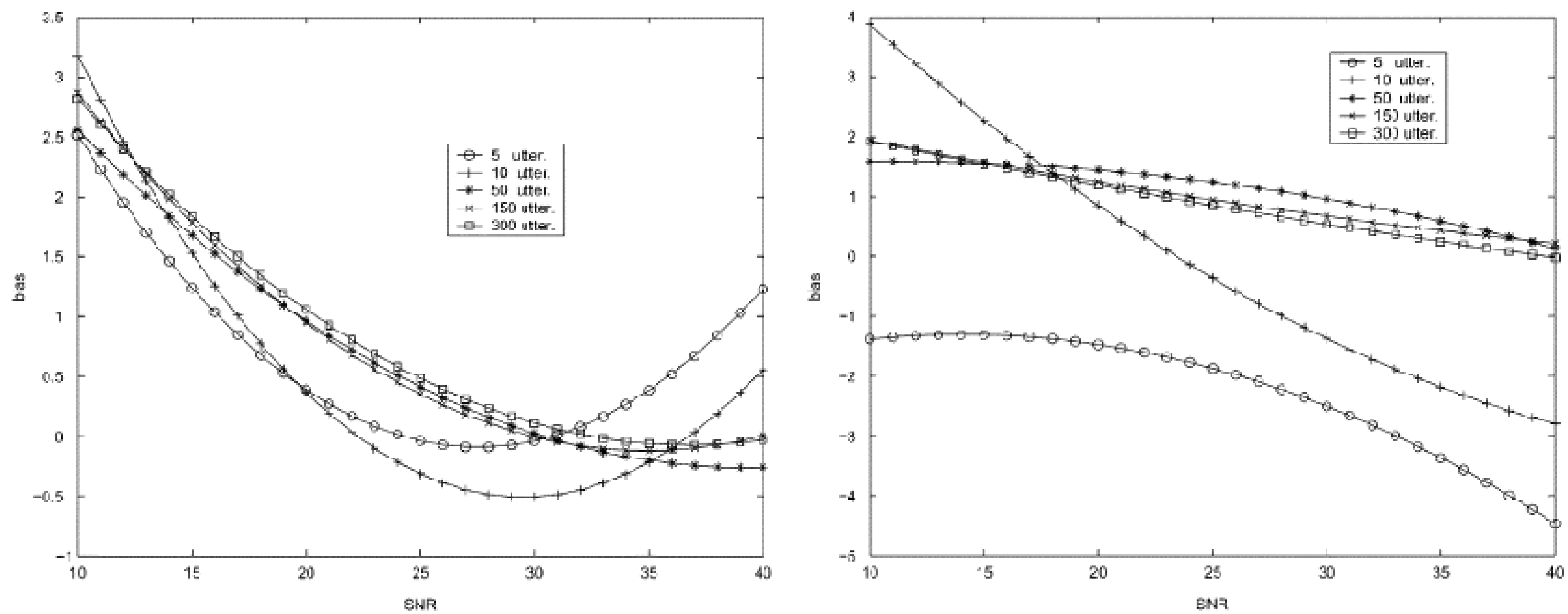

Fig. 11. Left and right panels show estimated global polynomials as a function of SNR and number of utterances for energy (E) and the first cepstral coefficient (C1), respectively, under car noise. The number of EM iterations is fixed at six.

Estimated regression polynomials

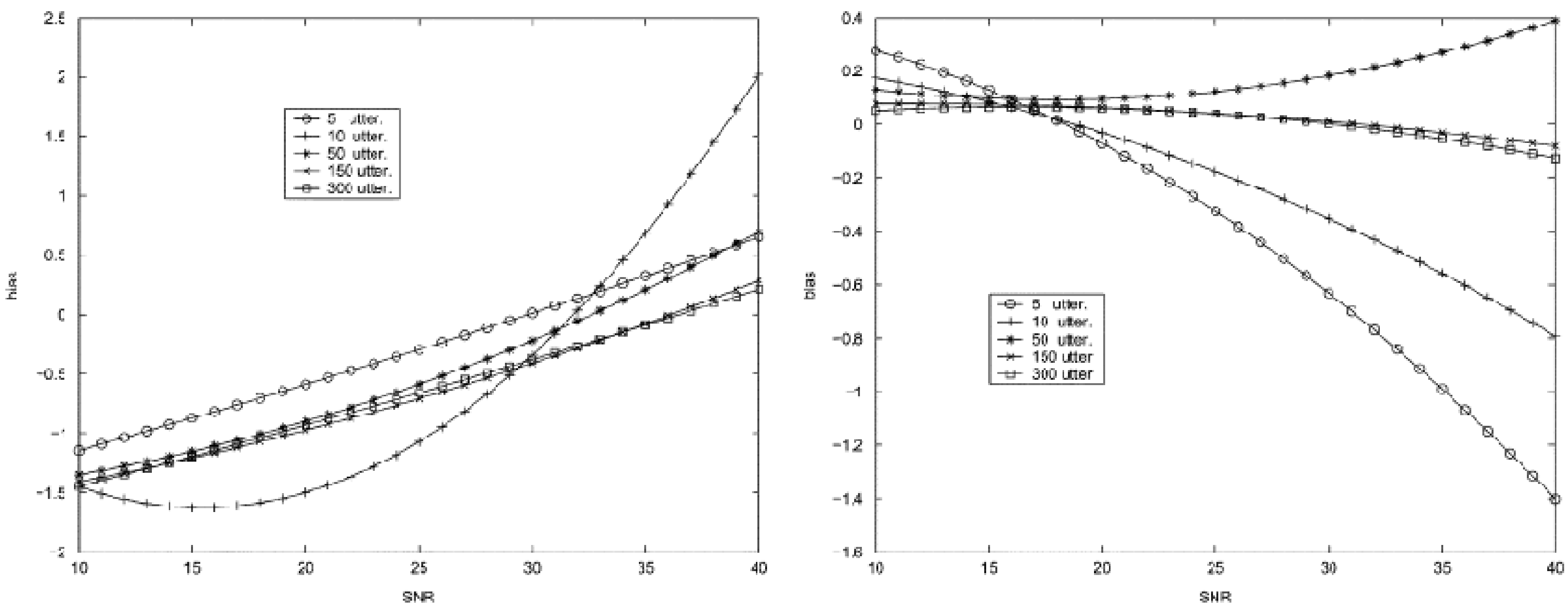

Fig. 12. Left and right panels show estimated global polynomials as a function of SNR and number of utterances for the sixth (C6) and tenth (C10) cepstral coefficients, respectively, under subway noise. The number of EM iterations is fixed at six.

Estimated regression polynomials

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

Recognition results

- 5, 10, and 20 utterances, polynomials are tied globally.
- 50, 100, and 150 utterances, polynomials are tied within states.
- 200, 250, and 300 utterances, polynomials are Gaussian mixture specific.
- The regression polynomial order is set to two.

Recognition results

- 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.

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

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

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