LOG-DOMAIN SPEECH FEATURE ENHANCEMENT USING SEQUENTIAL MAP NOISE ESTIMATION AND A PHASE-SENSITIVE MODEL OF THE ACOUSTIC ENVIRONMENT

Li Deng, Jasha Droppo, and Alex Acero

Microsoft Research, One Microsoft Way, Redmond WA 98052, USA

ABSTRACT

In this paper we present an MMSE (minimum mean square error) speech feature enhancement algorithm, capitalizing on a new probabilistic, nonlinear environment model that effectively incorporates the phase relationship between the clean speech and the corrupting noise in acoustic distortion. The MMSE estimator based on this phase-sensitive model is derived and it achieves high efficiency by exploiting single-point Taylor series expansion to approximate the joint probability of clean and noisy speech as a multivariate Gaussian. As an integral component of the enhancement algorithm, we also present a new sequential MAP-based nonstationary noise estimator. Experimental results on the Aurora2 task demonstrate the importance of exploiting the phase relationship in the speech corruption process captured by the MMSE estimator. The phase-sensitive MMSE estimator reported in this paper performs significantly better than phase-insensitive spectral subtraction (54% error rate reduction), and also noticeably better than a phase-insensitive MMSE estimator as our previous state-of-the-art technique reported in [2] (7% error rate reduction), under otherwise identical experimental conditions of speech recognition.

1. INTRODUCTION

This paper addresses the problem of speech feature enhancement, and the associated problem of noise feature estimation, when the noisy speech features alone are available as the observational information. These are long-standing, unsolved problems, and are becoming increasingly important recently due to emerging commercial deployment of speech recognition technology which demands a high degree of noise robustness.Towards high-performance solutions to robust speech feature enhancement and accurate noise estimation, we recently developed a series of enhancement techniques capitalizing on the availability of stereo training data [3, 4, 5] or on a simplistic, phase-insensitive nonlinear model of the acoustic environment [7, 1, 6, 2], both discarding the phase relationship between the clean speech and the additive noise during the speech signal corruption process. To overcome some weaknesses of these techniques, such as the difficulty of acquiring well-matched stereo training data and the performance limit due to loss of the phase information, we in more recent research have developed a new technique requiring no stereo training data. It explicitly exploits the novel concept of phase sensitivity and it uses a new sequential MAP (maximum a posteriori) noise estimator to design the speech feature enhancement algorithm.

In Section 2 of this paper, we will outline the new phase-sensitive nonlinear model for the acoustic environment. The MMSE estimator for noise removal based on this model is derived in Section 3. The novel MAP noise tracking algorithm is presented in Section 4, which provides an essential quantity required by the MMSE estimator. Finally, in Section 5 we will present experimental evidence on the Aurora2 task for the superiority of the phase-sensitive MMSE estimator and of the MAP noise tracker over the respective baselines.

2. A PROBABILISTIC ENVIRONMENT MODEL INCORPORATING PHASE OF ACOUSTIC DISTORTION

Using the discrete-time, linear system model for the acoustic distortion in the linear frequency domain, we have the well-known relationship among the noisy speech (Y), clean speech (X), additive noise (N), and channel transfer function (H) of

$$Y[k] = X[k]H[k] + N[k], \tag{1}$$

where k is the frequency-bin index in DFT given a fixed-length time window (frame).

Eq.1 can be shown to be equivalent, in the domain of log channel energy ($\mathbf{y}, \mathbf{x}, \mathbf{n}$, and $\mathbf{h}$), to the following relationship among these log-domain quantities [2]:

$$\begin{aligned} \mathbf{y} &= \mathbf{x} + \mathbf{h} + \log[\mathbf{1} + e^{\mathbf{n}-\mathbf{x}-\mathbf{h}} + 2\boldsymbol{\alpha} e^{\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}}] \\ &\equiv \mathbf{y}(\mathbf{x}, \mathbf{n}, \mathbf{h}, \boldsymbol{\alpha}), \end{aligned} \tag{2}$$

where the individual vector component of the random variable $\boldsymbol{\alpha}$ is the inner product (proportional to cosine of the phase) between Mel-channel-energy vectors of noise and the channel-distorted clean speech, characterizing their phase relationship. Based on the central limit theorem and empirical evidence, $\boldsymbol{\alpha}$ is assumed to follow a zero-mean Gaussian: $p(\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\alpha}; \mathbf{0}, \boldsymbol{\Sigma}_\alpha)$. This makes the model become phase sensitive, in contrast to our earlier model in [2] where an entire term of $\boldsymbol{\alpha}/\cos(\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2})$ is assumed to be Gaussian and hence the phase information is seriously smeared.

From Eq.2, $\boldsymbol{\alpha}$ can be solved as a function of the remaining variables:

$$\boldsymbol{\alpha} = \frac{e^{\mathbf{y}-\mathbf{h}} - e^{\mathbf{n}-\mathbf{h}} - e^{\mathbf{x}}}{2\, e^{\frac{\mathbf{n}+\mathbf{x}-\mathbf{h}}{2}}}. \tag{3}$$

The nonlinear transformation from $\boldsymbol{\alpha}$ to $\mathbf{y}$ in Eq.2 (for fixed values of $\mathbf{x}$ and $\bar{\mathbf{n}}$) allows us to obtain the conditional PDF of

$$p(\mathbf{y}|\mathbf{x}, \bar{\mathbf{n}}) = \frac{p(\boldsymbol{\alpha})}{|\frac{\partial \mathbf{y}}{\partial \boldsymbol{\alpha}}|}, \tag{4}$$

which gives rise to a probabilistic model of acoustic distortion.

In Eq.4, it can be shown from Eq.2 that

$$\frac{\partial \mathbf{y}}{\partial \boldsymbol{\alpha}} = 2\, e^{\frac{\mathbf{n}+\mathbf{x}+\mathbf{h}}{2} - \mathbf{y}}.$$

Also, the Gaussian assumption for $\boldsymbol{\alpha}$ gives

$$p(\boldsymbol{\alpha}) = p\left[\boldsymbol{\alpha}(\mathbf{x}, \mathbf{n}, \mathbf{h}, \mathbf{y})\right] = \mathcal{N}\left[\boldsymbol{\alpha}(\mathbf{x}, \mathbf{n}, \mathbf{h}, \mathbf{y}); \mathbf{0}, \boldsymbol{\Sigma}_\alpha\right]. \tag{5}$$

For exposition simplicity, in the remaining of this paper we assume: 1) The log-domain noise vector $\mathbf{n} = \bar{\mathbf{n}}$ is deterministic, or $p(\mathbf{n}) = \delta(\mathbf{n} - \bar{\mathbf{n}})$ ($\bar{\mathbf{n}}$ is obtained by a point estimator described in Section 4); and 2) $\mathbf{h} = \mathbf{0}$; i.e., the channel distortion can be ignored. Further, the covariance matrix $\mathbf{\Sigma}_\alpha$ is assumed to be diagonal with nonzero elements denoted by σ_α^2's. Thus, we will present the scalar rather than vector derivation, without loss of generality, for speech feature enhancement next.

3. MMSE ESTIMATOR FOR CLEAN SPEECH

Given the log-domain noisy speech observation y, the MMSE estimator $\hat{x}$ for clean speech x is the conditional expectation:

$$\hat{x} = E[x|y] = \int x p(x|y) dx = \frac{\int x p_{\bar{n}}(y|x) p(x) dx}{p(y)}, \quad (6)$$

where $p_{\bar{n}}(y|x) = p(y|x, \bar{n})$ is determined by the probabilistic environment model just presented. The prior model for clean speech, $p(x)$, in Eq.6 is assumed to have the Gaussian mixture PDF:

$$p(x) = \sum_{m=1}^{M} c_m \underbrace{\mathcal{N}(x; \mu_m, \sigma_m^2)}_{p(x|m)}, \quad (7)$$

whose parameters are pre-trained from the log-domain clean speech data. This allows us to write Eq.6 as

$$\hat{x} = \frac{\sum_{m=1}^{M} c_m \int x \overbrace{p(x|m) p(y|x, \bar{n})}^{J_m(x)} dx}{p(y)}, \quad (8)$$

The main difficulty in computing $\hat{x}$ above is the non-Gaussian nature of $p(y|x, \bar{n})$ of Eq.4. To overcome this difficulty, we use the truncated second-order Taylor series expansion to approximate the exponent of

$$\begin{aligned} J_m(x) &= \mathcal{N}(x; \mu_m, \sigma_m^2) \times \frac{\mathcal{N}(\alpha(x, \bar{n}, y); 0, \sigma_\alpha^2)}{2\, e^{\frac{\bar{n}+x}{2} - y}} \\ &= \frac{C}{\sigma_m} e^{-0.5(x-\mu_m)^2/\sigma_m^2 - 0.5x - 0.5\alpha^2(x)/\sigma_\alpha^2}. \end{aligned}$$

That is, we approximate the function

$$b_m(x) = -0.5(x - \mu_m)^2/\sigma_m^2 - 0.5x - 0.5\, \alpha^2(x)/\sigma_\alpha^2$$

by

$$b_m(x) \approx b_m^{(0)}(x_0) + b_m^{(1)}(x_0)(x - x_0) + \frac{b_m^{(2)}(x_0)}{2}(x - x_0)^2. \quad (9)$$

In Eq.9, we used a single-point expansion point x_0 (i.e., x_0 does not depend on the mixture component m) to have significantly improved computational efficiency, and x_0 is iteratively updated to increase its accuracy to the true value of clean speech x. The Taylor series expansion coefficients have the following closed forms:

$$\begin{aligned} b_m^{(0)}(x_0) &= b_m(x)\,|_{x=x_0} \\ &= -\frac{(x_0 - \mu_m)^2}{2\sigma_m^2} - \frac{x_0}{2} - \frac{(e^y - e^{\bar{n}} - e^{x_0})^2}{8\sigma_\alpha^2 e^{\bar{n}+x_0}}, \\ b_m^{(1)}(x_0) &= \frac{\partial b_m(x)}{\partial x}\,|_{x=x_0} = -\frac{x_0 - \mu_m}{\sigma_m^2} - \frac{1}{2} + \\ &\quad \frac{e^{2y-\bar{n}-x_0} - 2e^{y-x_0} + e^{\bar{n}-x_0} - e^{x_0-\bar{n}}}{8\sigma_\alpha^2}, \\ b_m^{(2)}(x_0) &= \frac{\partial^2 b_m(x)}{\partial^2 x}\,|_{x=x_0} = -\frac{1}{\sigma_m^2} + \\ &\quad \frac{-e^{2y-\bar{n}-x_0} + 2e^{y-x_0} - e^{\bar{n}-x_0} - e^{x_0-\bar{n}}}{8\sigma_\alpha^2}. \end{aligned}$$

Fitting Eq.9 into a standard quadratic form, we obtain

$$b_m(x) \approx \frac{b_m^{(2)}(x_0)}{2}\left[x - (x_0 - \frac{b_m^{(1)}(x_0)}{b_m^{(2)}(x_0)})\right]^2 + w_m(x_0),$$

where

$$w_m(x_0) = b_m^{(0)}(x_0) + \frac{b_m^{(2)}(x_0)}{2}\left[x_0^2 - \frac{2b_m^{(1)}}{b_m^{(2)}}x_0 - (x_0 - \frac{b_m^{(1)}}{b_m^{(2)}})^2\right].$$

This then allows us to compute the integral of Eq.8 in a closed form:

$$\begin{aligned} I_m(x_0) &= \int x J_m(x) dx = \frac{C}{\sigma_m}\int x e^{b_m(x)} dx \quad (10) \\ &\approx \frac{C'}{\sigma_m\sqrt{b_m^{(2)}}} e^{w_m(x_0)} \times \left(x_0 - \frac{b_m^{(1)}(x_0)}{b_m^{(2)}(x_0)}\right). \end{aligned}$$

The denominator of Eq.8 is computed according to

$$\begin{aligned} p(y) &= \sum_{m=1}^{M} c_m \int J_m(x) dx = \sum_{m=1}^{M} c_m \frac{C}{\sigma_m}\int e^{b_m(x)} dx \\ &\approx \sum_{m=1}^{M} c_m \frac{C'}{\sigma_m\sqrt{b_m^{(2)}}} e^{w_m(x_0)}. \quad (11) \end{aligned}$$

Substituting Eqs.10 and 11 into Eq.8, we obtain the final MMSE estimator:

$$\hat{x} \approx \sum_{m=1}^{M} \gamma_m(x_0, \bar{n}) \left(x_0 - \frac{b_m^{(1)}(x_0)}{b_m^{(2)}(x_0)}\right), \quad (12)$$

where the weighting factors are

$$\gamma_m(x_0, \bar{n}) = \frac{\frac{c_m}{\sigma_m\sqrt{b_m^{(2)}}} e^{w_m(x_0)}}{\sum_{m=1}^{M} \frac{c_m}{\sigma_m\sqrt{b_m^{(2)}}} e^{w_m(x_0)}}.$$

Note that $\gamma_m, b_m^{(1)}(x_0)$, and $b_m^{(2)}(x_0)$ in Eq.12 are all dependent on the noise estimate $\bar{n}$.

4. SEQUENTIAL MAP ESTIMATOR OF NOISE

In this section, we present a sequential MAP estimator (tracker) for log-domain nonstationary noise $\bar{n}$, which is used in computing quantities $\gamma_m, b_m^{(1)}(x_0)$, and $b_m^{(2)}(x_0)$ in the iterative MMSE estimation of clean speech according to Eq.12. This algorithm is generalized from the earlier ML estimator but within the same recursive-EM framework presented in [1] based on a relatively simple phase-insensitive acoustic distortion model (not the phase-sensitive model described in Section 2).

4.1. E-step

In the E-step, we compute the MAP auxiliary function of

$$Q_{MAP}(n_t) = Q_{ML}(n_t) + \rho \log p(n_t),$$

where

$$\begin{aligned} Q_{ML}(n_t) &= E\big[\log p(y_1^t, \mathcal{M}_1^t | n_t) | y_1^t, n_1^{t-1}\big] \\ &= \sum_{\tau=1}^{t} \epsilon^{t-\tau} \sum_{m=1}^{M} \xi_\tau(m) \log\ p(y_\tau \mid m, n_t) \\ &= -\sum_{\tau=1}^{t} \epsilon^{t-\tau} \sum_{m=1}^{M} \xi_\tau(m) \frac{(y_\tau - \mu_m^y)^2}{2\,\Sigma_m^y}. \quad (13) \end{aligned}$$

In Eq.13, ϵ is the forgetting factor, $\mathcal{M}_1^t$ is the sequence of the speech model's mixture components up to frame t, and $\xi_\tau(m) = p(m|y_\tau, n_{\tau-1})$ is the posterior probability. It is computed using Bayes rule by computing the likelihood $p(y_\tau|m, n_{\tau-1})$. This is approximated by a Gaussian with mean and variance of

$$\begin{aligned} \mu_m^y &\approx \mu_m^x + g_m + [1 - G_m](n_t - n_0) \\ \Sigma_m^y &\approx (1 + G_m)^2 \Sigma_m^x + (1 - G_m)^2 \Sigma^n. \quad (14) \end{aligned}$$

where g_m and G_m are computable quantities used to approximate the linear relationship among noisy speech y, clean speech x, and noise n (all in the log-domain) [1]. Σ^n is the variance (hyper-parameter) of the prior noise PDF $p(n_t)$, which is assumed to be Gaussian (with mean μ_n). And n_0 is the Taylor series expansion point for the noise, which will be iteratively updated by the MAP estimate in the M-step described below.

4.2. M-step

In the M-step, we estimate n_t by setting $\frac{\partial Q_{MAP}(n_t)}{\partial n_t} = 0$. Noting from Eq.14 that μ_m^y is a linear function of n_t, we obtain

$$\sum_{\tau=1}^{t} \epsilon^{t-\tau} \sum_{m=1}^{M} \xi_\tau(m) \frac{(y_\tau - \mu_m^y)}{\Sigma_m^y} (1 - G_m) - \frac{\rho(n_t - \mu_n)}{\Sigma^n} = 0. \quad (15)$$

Substituting Eq.14 into Eq.15 and solving for n_t, we obtain the MAP estimate of noise

$$\hat{n}_t = \frac{s_t + \rho\mu_n/\Sigma^n + K_t n_0}{K_t + \rho/\Sigma^n},$$

where

$$s_t = \sum_{\tau=1}^{t} \epsilon^{t-\tau} \sum_{m=1}^{M} \xi_\tau(m)(y_\tau - \mu_m^x - g_m) \frac{(1 - G_m)}{\Sigma_m^y},$$

and

$$K_t = \sum_{\tau=1}^{t} \epsilon^{t-\tau} \sum_{m=1}^{M} \xi_\tau(m) \frac{(1 - G_m)^2}{\Sigma_m^y}.$$

The s_t and K_t above can be efficiently computed using recursions based on the previous computation for s_{t-1} and K_{t-1}, as in our earlier work for the recursive ML noise estimate [1].

5. ROBUST SPEECH RECOGNITION EXPERIMENTS

In applying the MMSE estimator Eq.12 to perform speech feature enhancement, we first use the result of another enhancement algorithm (published in [2]) to initialize x_0 at the right hand side of Eq.12. The estimated clean speech $\hat{x}$ is then used to update x_0 and the iteration continues until a fixed number of iterations is reached or convergence occurs.

The MMSE estimator for clean speech features and the sequential MAP noise estimate described in this paper have been evaluated on the Aurora2 database, using the standard recognition tasks designed for this database. The database consists of English connected digits recorded in clean environments. Three sets of digit utterances (sets A, B, and C) are prepared as the test material. These utterances are artificially contaminated by adding noise recorded under a number of conditions and for different noise levels (sets A, B, and C), and also by passing them through different distortion channels (for set C only). The HMMs used in our evaluation experiments are specified by the Aurora2 task and trained using the clean-speech training set.

5.1. Results using phase-removed vectors of true noise

In this set of experiments, we use the MFCCs and their inverse cosine transform computed from true noise (available in the Aurora2 database) as the deterministic noise $\bar{n}$ in Eq.12 to evaluate the effects of various factors on the MMSE estimator's performance for noise-robust speech recognition. Other objectives of these experiments are to set the upper limit for the possible performance, and to demonstrate the effectiveness of incorporating the phase information in the speech distortion process.

Table 1 shows percent accuracy results on the full set of Aurora2 test data, when clean-speech HMMs are used, as a function of the number of iterations (L) for the MMSE estimator of Eq.12. When $L = 0$, the initial clean-speech estimate, obtained from the algorithm published in [2] that largely discards the phase information in the speech corruption process, is used for recognition. When the MMSE estimator of Eq.12 is applied iteratively to update the initial estimate, dramatic performance improvement is observed consistently across all three data sets. Performance convergence occurs at around seven iterations.

Table 1. Effects of the total number of iterations (L) on the MMSE estimator's performance (percent accurate) for the Aurora2 task. Phase-removed MFCC vectors of true noise are used for $\bar{n}$ in Eq.12.

L	0	1	2	4	7	12
SetA	85.7	94.1	96.8	97.8	98.1	98.1
SetB	86.2	94.8	97.3	98.1	98.5	98.6
SetC	80.4	91.0	94.5	96.5	97.9	98.0
Ave.	84.8	93.8	96.5	97.7	98.2	98.3

To further demonstrate benefits of the MMSE estimator of Eq.12 in modeling the phase information, we use the same true noise for log-domain spectral subtraction (SS) and perform the same Aurora2 evaluation. The SS algorithm is obtained by setting $\boldsymbol{\alpha} = 0$ in Eq.2 (as well as $\mathbf{h} = 0$), which gives

$$\hat{x} = \log(e^y - e^n) = y + \log(1 - e^{n-y}).$$

To avoid the possibility of taking logarithm of negative values (when $n > y$ due to statistical variation), we introduce the floor parameter F according to:

$$\hat{x} = y + \log\left[\max(1 - e^{n-y}, F)\right], \ or \quad (16)$$

$$\hat{x} = y + \log\left[\max(|\ 1 - e^{n-y}\ |, F)\right]. \quad (17)$$

These two ways of using the floor, in combination of applying the SS in the domains of direct Mel-scaled log-channel energies and of MFCCs as smoothed log-channel energies, result in four versions of the SS algorithm. Their respective recognition accuracies (%) as a function of the floor level are listed in Table 2 for Set A of the Aurora2 test data. Note that the best accuracy, 95.9%, still contains 54% more errors than that achieved by the converged MMSE estimator (98.1% accuracy).

Table 2. Performance (percent accurate) for the Aurora2 task (Set-A only) using four versions of spectral subtraction (SS).

Floor	e^{-20}	e^{-10}	e^{-5}	e^{-3}	e^{-2}
SS1	93.57	94.26	95.90	92.18	90.00
SS2	12.50	44.00	65.46	88.69	84.44
SS3	88.52	89.26	93.19	90.75	88.00
SS4	10.00	42.50	63.08	87.41	84.26

In Table 2, Phase-removed, Mel-scaled log-channel energies (SS1 and SS2) or MFCCs (SS3 and SS4) are computed from true noise waveforms. SS1 and SS3 make use of Eq.17. SS2 and SS4 make use of Eq.16.

5.2. Results using ML and MAP noise estimators

In contrast to using the true noise vector as $\bar{n}$ in Eq.12 when applying the MMSE estimator to speech feature enhancement just described, in this section are presented the results using the estimated noise vectors. The best technique we have developed so far is the sequential MAP noise estimator described in Section 4, where the prior distribution of the noise is assumed to be diagonal Gaussian. In the current implementation and in the evaluation on the Aurora2 task, the mean and variance of the Gaussian change from utterance to utterance in the test data. They are fixed to be the sample mean and sample variance of the first 20 frames in each separate test utterance, which are assumed to be free of any speech material.

Applying the MAP noise estimator to the MMSE estimator (one iteration) for clean speech, we obtain the percent-accuracy performance results for all three sets of the Aurora2 test data. The results are shown in the last column of Table 3, using $\hat{x}$ in Eq.12 (with MAP-tracked noise as $\bar{n}$) to score the pre-trained clean-speech HMMs. This gives significant improvement over the baseline performance (established in the work of [2] and shown in Column 2 in Table 3), where the initial clean speech vector x_0 in Eq.12 (i.e., without using the MMSE estimator) is used to score the HMMs. Compared with the performance shown in Column 3 in Table 3, the MAP-tracked noise (as described in Section 4) also provides moderate improvement (7% error rate reduction) over the use of the sequential maximum likelihood (ML) noise estimator in the otherwise identical experimental setup (i.e., using $\hat{x}$ in Eq.12 with the ML-tracked noise as $\bar{n}$). The algorithm for computing the ML-tracked noise estimator can be found in [1], which gave the state-of-the-art performance in our earlier noise-robust recognition system [5].

6. SUMMARY AND CONCLUSION

The earlier log-domain environmental models for speech distortion either did not incorporate any random variation [7], or, if so, did not capture the phase relationship between the clean speech and noise during the process of speech distortion [2, 6]. The new environmental model presented in this paper explicitly represents this phase relationship by modeling the inner product of the clean speech and noise vectors (in the frequency domain) as a random variable. This model offers the advantage of automatically capturing the effects of the instantaneous SNR on speech distortion. Experimental results obtained from the Aurora2 task demonstrate the importance of exploiting the phase relationship. The phase-sensitive MMSE estimator based on this new model performs significantly better than spectral subtraction, which discards the phase information, using identical noise estimates. It also outperforms our earlier algorithm [2] which is largely phase-insensitive also.

To further improve the phase-sensitive modeling technique for speech feature enhancement, we are currently working on sequential updating of the noise prior for improved point estimate of noise, and on incorporating posterior noise distributions into a new version of the phase-sensitive MMSE estimator.

Table 3. MMSE estimator's performance (percent accurate) for the Aurora2 task using sequential ML and MAP noise estimates (instead of true noise).

	Baseline (x_0 in Eq.12)	ML-tracked noise	MAP-tracked noise
SetA	85.66	86.34	86.39
SetB	86.15	86.24	86.30
SetC	80.40	82.50	83.35
Ave.	84.80	85.53	85.74

7. REFERENCES

[1] L. Deng, J. Droppo, and A. Acero. "Recursive estimation of nonstationary noise using a nonlinear model with iterative stochastic approximation," *Proc. ASRU Workshop,* Trento, Italy, Dec. 2001, 4 pages.

[2] L. Deng, J. Droppo, and A. Acero. "A Bayesian approach to speech feature enhancement using the dynamic cepstral prior," *Proc. ICASSP*, Vol.I, Orlando, Florida, May 2002, pp. 829-832.

[3] L. Deng, A. Acero, M. Plumpe, and X.D. Huang. "Large-vocabulary speech recognition under adverse acoustic environments," *Proc. ICSLP*, Vol. 3, 2000, pp. 806-809.

[4] L. Deng, A. Acero, L. Jiang, J. Droppo, and XD Huang. "High-performance robust speech recognition using stereo training data," *Proc. ICASSP*, Vol.1, 2001, pp. 301-304.

[5] J. Droppo, L. Deng, and A. Acero. "Evaluation of the SPLICE algorithm on the Aurora2 database," *Proc. Eurospeech*, Sept. 2001, pp. 217-220.

[6] B. Frey, L. Deng, A. Acero, and T. Kristjansson. "ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition," *Proc. Eurospeech*, Sept. 2001, pp. 901-904.

[7] P. Moreno, B. Raj, and R. Stern. "A vector Taylor series approach for environment-independent speech recognition," *Proc. ICASSP*, Vol.1, 1996, pp. 733-736.