A BAYESIAN APPROACH TO SPEECH FEATURE ENHANCEMENT USING THE DYNAMIC CEPSTRAL PRIOR

Li Deng, Jasha Droppo, and Alex Acero

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

ABSTRACT

A new Bayesian estimation framework for statistical feature extraction in the form of cepstral enhancement is presented, in which the joint prior distribution is exploited for both static and frame-differential dynamic cepstral parameters in the clean speech model. The conditional minimum mean square error (MMSE) estimator for the clean speech feature is derived using the full posterior probability for clean speech given the noisy observation. The final form of the estimator (for each mixture component) is a weighted sum of the prior information using the static and the dynamic priors separately, and of the prediction using the acoustic distortion model in absence of any prior information. Comprehensive noise-robust speech recognition experiments using the Aurora2 database demonstrate significant improvement in accuracy by incorporating the joint prior, compared with using only the static or dynamic prior and with using no prior.

1. INTRODUCTION

One major problem that remains unsolved in speech recognition technology is noise robustness. Towards solving this problem, we recently have successfully developed a family of front-end speech feature enhancement algorithms that make use of the availability of stereo training data consisting of simultaneously collected clean and noisy speech under a variety of noisy conditions [4, 5, 7]. While high performance under severe noise distortion conditions is achievable, it is desirable to remove or reduce the need for the stereo training data, and to overcome the potential problem of unexpected mismatch between the acoustic environments for recognizer deployment and for stereo training. To this end, we have more recently focused on the development of a new family of statistical and parametric techniques for noise-robust speech recognition. In this paper we present a new algorithm for statistical feature extraction in the form of cepstral (or equivalent log-spectral) enhancement.

The new algorithm has been built upon a series of published work on parametric modeling of nonlinear acoustic distortion [11, 1, 6, 9, 10, 14], and on the use of speech prior [8, 13, 9, 2, 3], but it represents a significant extension or generalization of the earlier work. The main innovations of the algorithm presented in this paper are: 1) incorporation of the dynamic cepstral features in the Bayesian estimation framework; 2) a new conditional MMSE estimator that elegantly integrates the predictive information from the nonlinear acoustic distortion model and the prior information based on the dynamic as well as static clean speech's cepstral distributions; and 3) efficient implementation of the new algorithm (real time using a Matlab code).

This paper is organized as follows. In Section 2, we establish a statistical model for the acoustic environment which relates the log-spectral vectors of clean speech, noise, and noisy speech in a nonlinear manner. In Section 3, we describe "prior" models for both clean speech and noise. In Section 4, we use Bayes rule to derive the conditional MMSE for the clean speech cepstra or log spectra by combining the prior information and a linearized version of the statistical model for approximating the nonlinear acoustic environment, In Section 5, noise-robust experimental results are reported using Aurora2 database, which demonstrate the effectiveness of the new Bayesian approach, and in particular, of the use of dynamic cepstral features in the prior model.

2. A STATISTICAL MODEL FOR ACOUSTIC DISTORTION

Following the discrete-time, linear system model for the acoustic distortion in the time and linear frequency domain, and taking into account the random angle between the speech and noise spectral vectors, we have established the following relationship among the noise (n), noisy speech (y), and clean speech (x) log-spectral vectors for each filter bank:

$$\begin{aligned} y &= x + \log(1 + e^{n-x}) + \log[1 + \lambda \cosh(\frac{n-x}{2})] \\ &\approx x + \log(1 + e^{n-x}) + \lambda \cosh(\frac{n-x}{2}) \end{aligned} \quad (1)$$

where λ (for each filter bank) is

$$\lambda \equiv \frac{\sum_k W_k X[k] N[k] cos\theta_k}{\sqrt{\sum_k W_k |X[k]|^2} \sqrt{\sum_k W_k |N[k]|^2}}.$$

In the above, $X[k]$ and $N[k]$ are the spectra for a frame of clean speech and noise, respectively, W_k is the frequency transfer function for a filter in the filter bank, and θ_k is the (random) angle between the DFTs (complex variables) of noise and clean speech for each frequency bin k.

Due to the generally small values of the last term in Eq.1, the nonlinear acoustic environment model described by Eq.1 can be interpreted as a predictive mechanism for y, where the predictor is

$$\hat{y} = x + g(n - x),$$

in which

$$g(z) = \log\big(1 + e^z\big).$$

The small prediction residual in Eq.1:

$$r = \lambda \cosh(\frac{n-x}{2}), \quad (2)$$

is complicated to evaluate and to model. It is therefore represented by an "ignorance" model as a zero-mean, Gaussian random vector.

The covariance matrix for the prediction residual Eq.2 is clearly a function of the (instantaneous) SNR. But to avoid the implementation complexity associated with the SNR dependency, we used one fixed (diagonal) covariance matrix, Ψ, which is estimated by pooling the training data (Aurora2) with all available SNRs. This thus establishes a crude but efficient statistical model for the acoustic environment:

$$p(y|x,n) = \mathcal{N}(y; x + g(n-x), \Psi) = \mathcal{N}(y; \hat{y}, \Psi), \qquad (3)$$

which will be used as one component in speech feature enhancement. This type of model has also been used in other frameworks for the enhancement [9].

3. PRIOR MODELS

The prior model exploited in this work takes into account both the static and dynamic properties of clean speech, in the domain of log Mel-channel energy (or equivalently in the domain of cepstrum via a fixed, linear transformation). One simple way of capturing the dynamic property is to use the frame-differential, or "delta" feature, defined by

$$\Delta x_t \equiv x_t - x_{t-1},$$

where a one-step, backward time (frame) difference is used in this work.

The functional form of the probability distribution for both the static and delta features of clean speech is chosen, motivated by simplicity in the algorithm implementation, as a mixture of multivariate Gaussians, where in each Gaussian component the static and delta features are assumed to be uncorrelated with each other. This gives the joint distribution:

$$p(x_t, \Delta x_t) = \sum_{m=1}^{M} c_m \mathcal{N}(x_t; \mu_m^x, \Phi_m^x) \mathcal{N}(\Delta x_t; \mu_m^{\Delta x}, \Phi_m^{\Delta x}). \qquad (4)$$

In our speech feature enhancement system implementation, a standard EM algorithm is used to train the mean and covariance parameters $\mu_m^x, \Phi_m^x, \mu_m^{\Delta x}$, and $\Phi_m^{\Delta x}$ in the cepstral domain. Then the mean vectors in the log Mel-channel energy domain are obtained via the linear transform using the inverse cosine transformation matrix. The diagonal elements of the two covariance matrices in the log Mel-channel energy domain are computed also from those in the cepstral domain, using the inverse cosine transformation matrix and its transpose. After this training and the transformations, we now assume that all parameters in Eq.4 are known in the log Mel-channel energy domain.

In principle, in the Bayesian framework adopted in this work, it is also desirable to provide a prior distribution for the noise parameter n. Due to the fast changing nature of the noise in the database (Aurora2) which we evaluate our algorithm on, the noise distribution would need to be nonstationary or time-varying; that is, the noise distribution be a function of time frame t. Given only a limited amount of noisy speech training data available, even assuming a simple Gaussian model for the noise feature with a time-varying mean and variance, accurate estimation of these parameters is still very difficult. To overcome this difficulty, we in this work use the results from our earlier research where the noise feature is assumed to be deterministic and is tracked sequentially directly from the individual noisy test utterance [6]. This is equivalent to assuming the prior distribution for noise is a time-varying, vector-valued, delta function:

$$p(n_t) = \delta(n_t - \bar{n}_t). \qquad (5)$$

4. ALGORITHMS FOR SPEECH FEATURE ENHANCEMENT

Given the observation vector y, the MMSE estimator $\hat{x}$ for the random vector x is the conditional expectation:

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

where the last step used Bayes rule. The enhancement algorithms to be described below provide efficient ways of computing the right hand side of Eq.6.

4.1. Estimation with static prior only

To facilitate the derivation of the MMSE estimator with the prior speech model for joint static and dynamic features, we first derive the estimator with the static prior only. The result will be extended to the desired case in the next subsection.

In this derivation, the prior model for clean speech is a simplified version of the model in Eq.4:

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

Eq.6 can then be evaluated as

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

after using the deterministic prior noise model Eq.5.

Based on the statistical environment model of Eq.3, the integral in Eq.8 is computed as

$$I_m = \int x \mathcal{N}(x; \mu_m^x, \Phi_m^x) \mathcal{N}(y; x + g(\bar{n} - x), \Psi) dx, \qquad (9)$$

where y and $\bar{n}$ are treated as constants. This integral, unfortunately, does not have a closed-form result due to the nonlinear function of x in $g(\bar{n} - x)$. To overcome this, we linearize the nonlinearity using truncated Taylor series. The zero-th order Taylor series expansion on $g(\bullet)$ at $x = x_0$ gives the following simple approximation:

$$y \approx x + g(\bar{n} - x_0) + r,$$

or equivalently,

$$p(y|x,n) = \mathcal{N}(y; \underbrace{x + g(\bar{n} - x_0)}_{\hat{y}}, \Psi). \qquad (10)$$

This approximation leads to the closed form of

$$I_m \approx [w_1(m)\mu_m^x + w_2(m)(y - g_0)] N_m(y), \qquad (11)$$

where $g_0 = g(\bar{n} - x_0)$, and where we introduced the weights

$$w_1(m) = \frac{\Psi}{\Phi_m^x + \Psi}$$

and $w_2(m) = I - w_1(m)$. In the above,

$$N_m(y) = \mathcal{N}(y; \mu_m^x + g_0, \Phi_m^x + \Psi)$$

can be easily shown to be the likelihood of observation y given the m-th component in the clean speech model and under the zero-th order approximation made in Eq.10. That is,

$$p(y|m) \approx N_m(y).$$

Using the result of Eq.11, together with

$$p(y) = \sum_{m=1}^{M} c_m p(y|m) \approx \sum_{m=1}^{M} c_m N_m(y), \qquad (12)$$

we obtain the approximate MMSE estimator as

$$\begin{aligned} \hat{x} &= \frac{\sum_{m=1}^{M} c_m N_m(y)[w_1(m)\mu_m^x + w_2(m)(y - g_0)]}{\sum_{m=1}^{M} c_m N_m(y)} \\ &= \sum_{m=1}^{M} \gamma_m(y)[w_1(m)\mu_m^x + w_2(m)(y - g_0)], \qquad (13) \end{aligned}$$

where

$$\gamma_m(y) = \frac{c_m N_m(y)}{\sum_{m=1}^{M} c_m N_m(y)}$$

is the discrete posterior probability $p(m|y)$.

4.2. Estimation with joint static and dynamic prior

We now derive the (conditional) MMSE estimator using a more complex prior speech model Eq.4 with joint static and dynamic features.

Given the estimated clean speech feature in the immediately past frame, $\hat{x}_{t-1}$, the conditional MMSE estimator for the current frame t becomes

$$\hat{x}_{t|t-1} \equiv E[x|y, \hat{x}_{t-1}].$$

Following a similar derivation for Eq.8, the corresponding result is

$$\hat{x}_{t|t-1} \approx \frac{\sum_{m=1}^{M} c_m \int x_t p(x_t|m, x_{t-1}) p(y_t|x_t, \bar{n}_t) dx_t}{p(y_t)}. \qquad (14)$$

To compute the integral in Eq.14, we first evaluate the conditional prior of

$$\begin{aligned} p(x_t|m, x_{t-1}) &\propto p(x_t, x_t - x_{t-1}|m) \\ &= \mathcal{N}(x_t; \mu_m^x, \Phi_m^x)\mathcal{N}(\Delta x_t; \mu_m^{\Delta x}, \Phi_m^{\Delta x}) \qquad (15) \end{aligned}$$

Fitting the exponent in the product of the above two Gaussians into the standard quadratic form in x_t, we have

$$p(x_t|m, x_{t-1}) = \mathcal{N}(x_t; \mu_m, \Phi_m), \qquad (16)$$

where

$$\mu_m = \frac{\Phi_m^{\Delta x}}{(\Phi_m^x + \Phi_m^{\Delta x})}\mu_m^x + \frac{\Phi_m^x}{(\Phi_m^x + \Phi_m^{\Delta x})}(x_{t-1} + \mu_m^{\Delta x}), \qquad (17)$$

and

$$\Phi_m = \frac{\Phi_m^x \Phi_m^{\Delta x}}{\Phi_m^x + \Phi_m^{\Delta x}}. \qquad (18)$$

Using the same zero-th order approximation of Eq.10, and substituting Eqs. 16-18 into Eq.14, we obtain the final result for the approximate conditional MMSE estimator:

$$\begin{aligned} \hat{x}_{t|t-1} &= \sum_{m=1}^{M} \gamma_m(y)[v_1(m)\mu_m^x + v_2(m)(\hat{x}_{t-1} + \mu_m^{\Delta x}) + \\ &\quad + w_2(m)(y - g(\bar{n} - x_0))], \qquad (19) \end{aligned}$$

where

$$v_1(m) = w_1(m)\frac{\Phi_m^{\Delta x}}{\Phi_m^x + \Phi_m^{\Delta x}},$$

$$v_2(m) = w_1(m)\frac{\Phi_m^x}{\Phi_m^x + \Phi_m^{\Delta x}},$$

and where $x_{t-1} \approx \hat{x}_{t-1}$ is used.

Note that

$$v_1(m) + v_2(m) + w_2(m) = 1. \qquad \forall\, m$$

This provides a clear interpretation of Eq.19 — each summand in Eq.19, as a mixture-component (m) specific contribution to the final estimator, is a weighted sum of three terms. The unweighted first two terms are derived from the static and dynamic elements in the prior clean speech model, respectively. The unweighted third term is derived from the predictive mechanism based on the linearized acoustic distortion model in absence of any prior information.

Note also that under the limiting case where $\Phi_m^{\Delta x} \to \infty$, we have

$$v_1(m) \to w_1(m), \;\; and \;\; v_2(m) \to 0.$$

Then the conditional MMSE estimator Eq.19 reverts to the MMSE estimator Eq.13 when no prior for dynamic speech features is exploited. This shows a desirable property of Eq.19 since when $\Phi_m^{\Delta x} \to \infty$ the effect of using the prior for dynamic features should indeed be diminishing to null.

As the opposite limiting case, let $\Phi_m^{\Delta x} \to 0$. We then have

$$v_1(m) \to 0, \;\; and \;\; v_2(m) \to w_1(m).$$

That is, only the prior information for the dynamic speech features is used for speech feature enhancement.

5. SPEECH RECOGNITION EXPERIMENTS

5.1. Database and recognition task

The algorithm presented thus far for estimating clean speech feature vectors has been evaluated on the Aurora2 database, using the standard recognition tasks designed for this database [12]. 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 recognition system used in our evaluation experiments are based on continuous HMMs, and one HMM is trained for each digit under clean condition. Both training and recognition phases are performed using the HTK scripts provided by the Aurora2 database. The speech feature used for the reference experiments to evaluate the new denoising algorithm is the standard MFCCs. The new algorithm is used only as the front-end.

5.2. Aurora2 results

Table 1 summarizes the results for all three sets of the test data in the Aurora2 database. The HMM systems with four different front-ends are compared: (1) use of the algorithm in Eq.19 to implement the conditional MMSE estimator, with the prior speech model consisting of both static and dynamic cepstra; (2) use of the same estimator except with the prior speech model consisting essentially of only the static cepstra (by setting $\Phi_m^{\Delta x} \to \infty$); (3) use of the same estimator except with the prior speech model consisting of only the dynamic cepstra (by setting $\Phi_m^{\Delta x} \to 0$); (4) no speech prior is used directly (prediction term only); and (5) a slight modification of Aurora2-supplied standard reference MFCCs with no denoising. [1] The HMMs used in the four systems are the same. They are trained using the same clean-speech training set.

Comparisons in Table 1 show that the conditional MMSE estimator that fully utilizes both the static and dynamic cepstral distributions (front-end (1)) performs significantly better than the same estimator which utilizes only the partial information ((2) and (3)) or using prediction only (4). They are, however, all significantly and consistently better than the standard MFCCs supplied by the Aurora task using no robust preprocessing to enhance speech features (front-end (5)). The relative word error rate reduction using front-end (1) is 64.54% compared with the results with standard MFCCs. These results are statistically significant, based on a total of $1001 \times 10 \times 5 = 50050$ test utterances from all set A, B, and C, among which there are 8008 distinct digit sequences corrupted under various distortion conditions.

It is worth mentioning that the front-end (4) as a degenerate, special case of the current algorithm, where $v_1(m) = v_2(m) = 0$, is similar to the VTS technique [11], with the difference that only one Taylor series expansion point is used and this point is iterated. With this degenerate case, our algorithm yielded similar results of the VTS as reported in [14] where the identical Aurora2 evaluation task was used.

Priors for Denoising	Set A	Set B	Set C	Overall
(1) Static-Dynamic Cepstra	86.72	87.03	81.70	**85.84**
(2) Static Cepstra only	84.74	85.19	78.87	83.74
(3) Dynamic Cepstra only	79.96	78.91	76.72	78.89
(4) Prediction only	77.72	77.30	75.40	77.08
(5) No Denoising (ref.)	61.34	55.75	66.14	60.06

Table 1. Comparisons of Aurora2 recognition rates (%) for the HMM systems using four different front-ends for all sets of the Aurora2 test data. Clean HMM training with cepstral mean normalization is used. Front-end (1) uses the algorithm described in Section 4.2. Front-ends (2) and (3) correspond to the two limiting cases discussed at the end of Section 4.2.

6. SUMMARY AND CONCLUSIONS

In this paper, a novel algorithm with its derivation, implementation, and evaluation is presented for statistical speech feature enhancement in the cepstral domain. It incorporates the joint static and dynamic cepstral features in the prior speech model within the Bayesian framework for optimal estimation of clean speech features. The estimator is based on the full posterior computation, and it elegantly integrates the predictive information from a statistical nonlinear acoustic distortion model, the prior information based on the static prior, and the prior based on the frame-differential dynamic prior. We have efficiently implemented this algorithm, which is used in the Aurora2 noise-robust speech recognition. The results demonstrate significant improvement in the recognition accuracy by incorporating the joint static/dynamic prior, compared with using only the static or dynamic prior and with using no prior.

The optimal estimator presented in Section 4 can be easily extended to include the conditional variance estimation. Given both the mean and variance estimates for the enhanced speech features, our future work will aim at a tight integration between the front-end denoising and the back-end speech recognition.

[1] The standard uses log-magnitude spectra, and we modified it to use log-magnitude squared spectra.

7. REFERENCES

[1] A. Acero, L. Deng, T. Kristjansson, and J. Zhang. “HMM adaptation using vector Taylor series for noisy speech recognition,” *Proc. ICSLP*, Vol.3, 2000, pp. 869-872.

[2] H. Attias, J. Platt, A. Acero, and L. Deng. “Speech denoising and dereverberation using probabilistic models,” *Advances in NIPS*, Vol. 13, 2000, pp. 758-764.

[3] H. Attias, L. Deng, A. Acero, and J. Platt. “A new method for speech denoising and robust speech recognition using probabilistic models for clean speech and for souse,” *Proc. Eurospeech*, 2001, pp. 1903-1906.

[4] 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.

[5] 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.

[6] L. Deng, J. Droppo, and A. Acero. “Recursive estimation of nonstationary noise using a nonlinear model with iterative stochastic approximation,” *Proc. ASRU Workshop*, Dec. 2001, Italy.

[7] J. Droppo, L. Deng, and A. Acero. “Evaluation of the SPLICE algorithm on the Aurora2 database,” *Proc. Eurospeech*, Sept 2001, Aalborg, Denmark.

[8] Y. Ephraim. “Statistical-model-based speech enhancement systems ,” *Proceedings of the IEEE* , Vol. 80, No. 10 , Oct. 1992, pp. 1526 -1555.

[9] 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, Aalborg, Denmark.

[10] N.S. Kim. “Nonstationary environment compensation based on sequential estimation,” *IEEE Sig. Proc. Letters*, Vol.5, 1998, pp. 57-60.

[11] 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.

[12] H. Hirsch and D. Pearce. “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions,” *Proc. ISCA ITRW ASR2000 on ASR*, Sept 2000, Paris, France.

[13] H. Sameti, H. Sheikhzadeh, L. Deng, and R. Brennan. “HMM-based strategies for enhancement of speech embedded in nonstationary noise,” *IEEE Trans. Speech and Audio Processing*, Vol.6, No.5, Sept 1998, pp. 445-455.

[14] J. Segura, A. Torre, M. Benitez, and A. Peinado. “Model-based compensation of the additive noise for continuous speech recognition,” *Proc. Eurospeech*, Sept 2001, Denmark.