ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

CONSTRAINED MAXIMUM A POSTERIORI APPROACH FOR SPEECH ENHANCEMENT

Chuan JIA, Jian ZHANG, Bo XU

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing
cjia@nlpr.ia.ac.cn

ABSTRACT

The maximum a *posteriori* estimator based on HMM's is successful to some degree because of the incorporation of prior knowledge of speech and markovian properties of the models. The enhanced speech quality is, however, not satisfying at low input SNR. In order to improve speech quality at low input SNR, this paper proposes a method that incorporates codebook constrained Wiener filter into MAP framework to impose spectral constraints on estimated speech signals. The objective measures, global SNR and Itakura-Saito distortion measure, verified the quality improvement of the proposed method.

1. INTRODUCTION

The presence of background noise can seriously degrade the performance of many speech-processing systems. As an example, an automatic speech recognizer trained in noise-free environment cannot work in noisy factory as well as in laboratory. Among various approaches to improve system robustness in noisy environments, speech enhancement is an important way. The main objective of speech enhancement is to improve perceptual aspects, such as overall quality and intelligibility for human or machines. Speech signals preprocessed by speech enhancement system become less sensitive to background noise so that existing speech-processing systems may be employed.

Speech enhancement has been a classic research topic for decades of years and a number of single-microphone speech enhancement algorithms have been proposed. These include variants of spectral subtraction; adaptive filtering, methods based on all-pole modeling and methods based on hidden markov models [1]. Among stochastic model based methods, HMM-based methods have been most successful. It combines knowledge of speech and noise by modeling both clean speech and noise and relaxes the assumption of noise stationary using an HMM with multiple states and mixtures for noise.

Ephraim proposed the maximum a *posteriori* (MAP) estimation using autoregressive HMM's in [2][3]. The MAP enhancement algorithm is a straightforward application of the EM algorithm for its iterative estimation of the clean speech waveform given the noisy speech. The algorithm starts from the noisy speech and generates a sequence of speech sample functions with non-decreasing likelihood values by maximizing an auxiliary function in each iteration [3]. His experiments indicated that the MAP enhancement system is very effective at the input SNR values above 5db and it is observed that it introduced some noticeable distortion and some additional residual noise into enhanced speech signals at the input SNR below 5db. One of the reasons is that at low input SNR the probability calculation especially in the first iteration is not so accurate as at higher input SNR. The derived filters that are harmonic mean of Wiener filters are not appropriate because of the inaccuracy. Furthermore, it cannot ensure that these frequency domain filters satisfy some spectral constraints on vocal tract spectrum suggested by Hansen and Clements [4]. Considering the above problems, it is valuable to try to incorporate spectral constraints into the MAP estimation system.

In the present work, we use codebook derived by unsupervised learning to constrain the frequency domain filters so as to suppress residual noise and reduce speech distortion. Besides imposing some spectral constraints on the estimated speech, it has implicitly other advantages for estimation process to combine codebook constrained Wiener filter into the original MAP framework. Experiments show that the performance of MAP estimator and that of codebook constrained Wiener filter are improved by each other. Moreover, as we know, the computational cost is raised along with the increase of the model size, but it is better to use more elaborate model. It is possible that the performance of using smaller speech model and codebook with enough size is similar to using large speech model.

In section 2, the background of the MAP enhancement algorithm is presented. Constrained MAP estimator is presented in section 3. Finally, the proposed enhancement algorithm is evaluated in section 4.

2. MAP ESTIMATOR BASED ON AR-HMM

2.1 Models for clean speech and noise process

The modeling problem in speech enhancement is different from in speech recognition. The objective is to train a single and compact model, which is to capture the global characteristics of speech and the temporal order of the states in the model, is not limited.

For autoregressive HMM [5], the observation vectors are drawn from an autoregression process. Consider the observation vector $\boldsymbol{O}=(x_0, x_1, x_2, \ldots x_{K-1})$. The components of $\boldsymbol{O}$ are assumed to satisfy the relationship

$$x_k = -\sum_{i=1}^{p} a_i x_{k-i} + e_k \qquad (1)$$

where e_k, $k=0,1,2,\ldots,K-1$ are Gaussian, independent and identically distributed random variables with zero mean and variance σ_e^2, and a_i, $i=1,2,\ldots p$ are the autoregression coefficients of order p. For large K, the density function for $\boldsymbol{O}$ is approximately

$$f(\boldsymbol{O}) = (2\pi\sigma_e^2)^{-K/2} \exp\{-\frac{1}{2\sigma_e^2}\delta(\boldsymbol{O},\boldsymbol{a})\} \tag{2}$$

where

$$\delta(\boldsymbol{O},\boldsymbol{a}) = r_a(0)r(0) + 2\sum_{i=1}^{p} r_a(i)r(i) \tag{3}$$

$$\boldsymbol{a} = [1, a_1, a_2, ..., a_p]' \tag{4}$$

$$r_a(i) = \sum_{n=0}^{p-i} a_n a_{n+i}, \qquad 1 \le i \le p \tag{5}$$

$$r(i) = \sum_{n=0}^{K-i-1} x_n x_{n+i}, \qquad 0 \le i \le p \tag{6}$$

$\delta(\boldsymbol{O}, \boldsymbol{a})$ is a form of residual energy resulting from inverse filtering the data x_i with an all-pole filter defined by $\boldsymbol{a}$.

The model parameters set with M states and L mixtures for each state is defined as $\lambda_s(\pi, A, C, H)$ for clean speech signals where $\pi=\{\pi_\beta\}$ is the set of initial state probabilities, $A=\{a_{\alpha\beta}\}$ is the set of state transition probabilities, $C=\{c_{\gamma|\beta}\}$ is the set of mixture weights and $H=\{h_{\gamma|\beta}\}$ with $h_{\gamma|\beta}$ being the AR parameter set corresponding to state and mixture pair (β, γ), $h_{\gamma|\beta}=\{1, h_{\gamma|\beta}(1), \cdots, h_{\gamma|\beta}(p), \sigma^2_{\gamma|\beta}\}$, $\sigma^2_{\gamma|\beta}$ being the variance (AR gain) for $\beta=1,\ldots,M$ and $\gamma=1,\ldots,L$. Given a K-dimension training sequence, a maximum likelihood estimation of the model parameter set λ_s is obtained through the Baum re-estimation algorithm. Alternately, the segmental K-means algorithm can be used to maximize the parameter set along the dominant state and mixture sequence [2][3].

The noise process considered in this paper is modeled by AR model. The noise modeling problem is to find the parameter set $\lambda_v=\{g_v, \sigma_v^2\}$ where $g_v=\{1, g_v(1),\ldots g_v(p_v)\}$ are the coefficients of AR process.

2.2 Maximum a posteriori estimator

Let y_t be a K-dimensional vector of the clean signal and v_t be a K-dimensional vector of the noise process. Assume that the noise is additive and statistically independent of the signal. Then the noisy speech of K-dimensional vector is $z_t = y_t + v_t$.

Let $y_0^t = \{y_\tau, \tau = 0,...,t\}$, $v_0^t = \{v_\tau, \tau = 0,...,t\}$ and $z_0^t = \{z_\tau, \tau = 0,...,t\}$. The causal MAP estimation of y_t given z_0^t is obtained by maximizing $p(y_t \mid z_0^t)$ over y_t. EM algorithm is applied for iterative local maximization of $p(y_t \mid z_0^t)$.

For each time frame, t, enhancement is done in frequency domain [3] using

$$y_{t,\theta}(k+1) = [\sum_{\beta=1}^{M}\sum_{\gamma=1}^{L} q_t(\beta,\gamma \mid y(k)) H^{-1}_{\gamma|\beta}(\theta)]^{-1} \cdot z_{t,\theta} \tag{7}$$

where k denotes the iteration index, $y_{t,\theta}(k+1)$ and $z_{t,\theta}$ are the Fourier transforms of y_t and z_t. $q_t(\beta, \gamma \mid y(k))$ is the conditional probability of being in state β and choosing mixture γ at time frame t given an estimation of the clean speech $y(k)$. $H_{\gamma|\beta}(\theta)$ is a Wiener filter associated with state β and mixture γ

$$H_{\gamma|\beta}(\theta) \overset{\Delta}{=} \frac{f_{\gamma|\beta}(\theta)}{f_{\gamma|\beta}(\theta) + f_v(\theta)} \tag{8}$$

where $f_{\gamma|\beta}(\theta)$ and $f_v(\theta)$ are the power spectral densities associated with the AR process corresponding to the state and mixture pair (β, γ) of clean speech model and the noise process respectively. From equation (7), a new estimate of the clean speech is obtained by filtering the noisy speech through a filter, which comprises the harmonic mean of Wiener filters.

2.3 Gain adaptation

The HMM-based speech enhancement approach uses the statistical properties of the speech signal from the training data. However, the energy contour cannot be reliably predicted from the training data and hence must be estimated from the given noisy signal. During training, ML estimation of the gain contour of the training data is used to train an HMM for gain-normalized clean speech signals. During enhancement, the HMM for gain-normalized clean signals is supplemented with ML estimation of the gain contour of the clean speech from the noisy speech [2].

For a given signal y_t and gain g_t at the kth iteration of gain-adaptation MAP enhancement, a new gain estimate at the $(k+1)$th iteration is calculated by

$$g_t^2(k+1) = \sum_{\beta=1}^{M}\sum_{\gamma=1}^{L} p(\beta,\gamma \mid y_t(k+1), g_t(k))\sigma^2_{t,\gamma|\beta} \tag{9}$$

where $\sigma^2_{t,\gamma|\beta}$ can be regarded as the average power of the residual signals, obtained from inverse filtering $y_t(k+1)$ using the AR process corresponding to the state and mixture pair (β, γ) of the HMM. The renewed gain is used in the calculation of power spectral density of speech in equation (8).

3. CONSTRAINED MAP

Since speech is assumed as AR process, we can express the harmonic mean of Wiener filters in equation (7) as a filter whose structure is similar to equation (8) and then the filtering process of equation (7) is like MAP estimation of speech formulated by Lim and Oppenheim [6]. Thus there is implicitly an estimate of vocal tract spectrum or AR model spectrum. Refer to [4], the vocal tract spectrum is required to satisfy some constraints: the all-pole model is stable; it possesses speech-like characteristics and the vocal tract characteristics do not vary too much between successive frames. The all-pole model spectrum of speech here, however, cannot ensure these constraints. Considering the deficiency, there is clearly a need to incorporate the constraints into the filtering process of equation (7).

Intra-frame constraints to LPC or LSP parameters [4] derived from acoustic-phonetic knowledge of speech are hard to be determined because of the variability among speakers. On the other hand, the codebook trained from clean speech data can cover the statistical knowledge of formant frequencies. In the present work, we combine codebook constraints into the MAP estimator to apply intra-frame constraints. Inter-frame constraints are neglected here because of the frame overlapping and markovian properties.

3.1 Framework

At each iteration, the MAP estimation of speech is firstly obtained by applying a filter that comprises the harmonic mean of Wiener filters to noisy speech. Then after gain adaptation of speech, choose codewords closest to enhanced speech derived from MAP estimator. Each frame of noisy speech is again filtered by a codebook constrained Wiener filter which is constituted by the corresponding optimal codeword and re-estimated gain of speech. Enhanced speech is input into MAP estimator again. The process is repeated until the enhanced speech quality reaches the "optimum". Noisy speech is assumed as initial estimation. The convergence is determined by some objective measures such as output SNR and Itakura-Saito distortion measure. The "optimal" enhanced speech is obtained after MAP estimation or filtering by codebook constrained Wiener filter at the best iteration number. The whole process is described in Fig. 1.

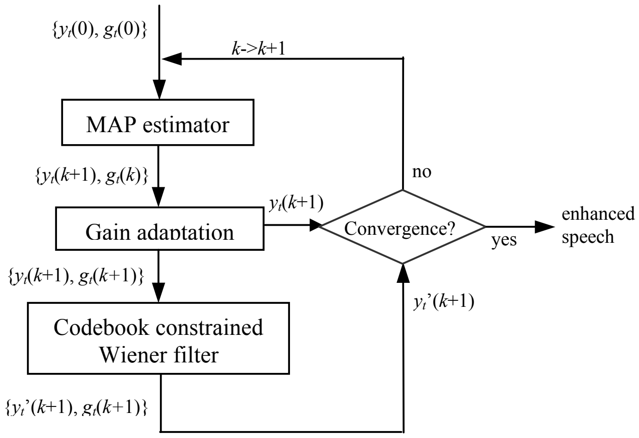

Fig. 1: A block diagram for constrained MAP

3.2 Codebook constrained Wiener filter

The perceptual distance between two p^{th} order LPC vectors is defined by Itakura-Saito (IS) measure

$$d_{IS}(S,S')=\int_{-\pi}^{\pi}[\log\frac{S'(\theta)}{S(\theta)}+\frac{S(\theta)}{S'(\theta)}-1]\frac{d\theta}{2\pi} \tag{10}$$

where $\boldsymbol{S}'(\theta)$ and $\boldsymbol{S}(\theta)$ is replaced by their corresponding p^{th} order all-pole model spectra $\sigma_p^{2'}/|A_p^{'}|^2$ and $\sigma_p^2/|A_p|^2$.

The codebook is obtained by unsupervised learning from a set of LPC vectors of clean speech. The LPC vectors are clustered by iterative K-means algorithm using IS distance. Over the iterations, the sum of the distance between centroid of each cluster and LPC vectors in this cluster is minimal. The centroid of each cluster is the average of its LPC vectors. The obtained centroids represent the most often occurring spectra of speech.

At each iteration during enhancement, after MAP estimation and gain adaptation, the LPC vector $\boldsymbol{a}$ of enhanced speech $y_t(k+1)$ derived from MAP estimator is calculated and the codeword c_i closest to $\boldsymbol{a}$ is chosen from codebook

$$d(a,c_i)\le d(a,c_k), \qquad \forall k \tag{11}$$

The power spectral density of clean speech is re-estimated by equation (12) and the codebook constrained Wiener filter for the present frame t is formed by equation (13)

$$f_{t,\theta}(k+1)=\frac{g_t^2(k+1)}{|1+\sum_{n=1}^{p}c_i(n)e^{-j\theta n}|^2} \tag{12}$$

$$H'_{t,\theta}(k+1)=\frac{f_{t,\theta}(k+1)}{f_{t,\theta}(k+1)+f_v(\theta)} \tag{13}$$

where $g_t(k+1)$ is the gain of speech derived from the process of gain adaptation. Since c_i has been estimated from clean speech, the corresponding Wiener filter imposes some spectral constraints on the filtering process. Thus, the enhanced speech in frequency domain is obtained by

$$y'_{t,\theta}(k+1)=H'_{t,\theta}(k+1)\cdot z_{t,\theta} \tag{14}$$

4. EXPERIMENTS AND RESULTS

4.1 Objective evaluation

Two types of measures of speech quality were used as the objective evaluation criterion. The first is the global output SNR

$$SNR=10\log\{\sum_{n=1}^{N}y^2(n)/\sum_{n=1}^{N}[y(n)-y'(n)]^2\} \tag{15}$$

where N is speech length, $y(n)$ and $y'(n)$ is clean speech and enhanced speech respectively. The second is Itakura-Saito distortion measure between enhanced speech and clean speech.

4.2 Basic MAP estimation

In all of our experiments, training of HMM for clean speech was performed using non-overlapping vectors of speech signals. The training data consisted of 15 minutes data, 16kHz sampling frequency, from 6 speakers, 3 males and 3 females. A special state was assigned to represent very low energy segments of clean speech and it had multiple mixtures in order to ensure that perceptually important weak speech signals will be well represented. The order of the AR-HMM for clean speech was p=12. The AR model for noise was trained from an initial segment of each noisy speech, which contains only noise and the order for noise was 10.

In the present experiments, 60 clean utterances spoken by 3 male speakers and 3 female speakers were degraded by additive Gaussian white noise at 0-15db input SNR for test. Neither the speakers nor the speech data used in test were in the training set. During enhancement, frames were 50% overlapped. A hanning window was applied to the speech frames during training and enhancement. The synthesis of the enhanced speech from the individually processed frame was done by the standard short time Fourier transform overlap and add algorithm.

In order to determine the number of states and mixture components for each state for speech HMM, we examined the experiment results obtained using different model size at input

SNR of 10db. Table 1 shows results of SNR improvement. According to the results, the model of 8 states and 12 mixture components for each state is the best here and was chosen for all following experiments. Table 2 shows results of SNR improvement at different input SNR using the model (8/12). The convergence occurred at the third iteration for almost all utterances of test set.

State/mixture	*5/5*	*6/16*	*6/8*	*8/12*	*8/6*	*8/2*
Iterate 1	3.45	3.57	3.51	3.53	3.47	3.19
Iterate 2	4.33	4.54	4.46	4.55	4.49	4.32
Iterate 3	4.64	4.88	4.78	5.09	4.86	4.58
Iterate 4	3.96	4.37	4.25	4.42	3.92	4.07

Table 1: Enhancement results (db) for different number of states and mixtures at 10db input SNR (female data).

Input SNR		*0db*	*5db*	*10db*	*15db*
Female	*Iterate 3*	7.54	6.27	5.09	4.08
Male	*Iterate 3*	7.85	6.47	5.26	4.22

Table 2: Enhancement results (db) at 0-15db input SNR using model 8/12.

4.3 Constrained MAP estimation

The clean speech set for training LPC codebook was the same as the speech set used for training HMM's. The LPC vectors, order=12, were extracted with consecutive frames overlapping. Since our aim is to improve quality of enhanced speech at low input SNR, the following experiments were done at 0db and 5db input SNR. Here the utterances for test, spoken by 3 female speakers and 3 male speakers, were selected randomly from above test set.

Table 3 and Table 4 shows the comparison of enhancement results between the basic MAP estimation and the proposed constrained MAP estimation. It can be seen that the constrained MAP estimation results in improvement in global SNR except that input SNR=0db and speaker is female. Moreover, the IS distance between enhanced speech and original clean speech is reduced by using constrained MAP. In all experiments, the utterances corresponding to the male speakers were better enhanced than the utterances corresponding to the female speakers. A possible reason for this difference is that the AR modeling approach is better for male voice signals.

		MAP	*CMAP*
Female	*SNR (db)*	*6.96*	*6.96*
	IS	*1.89*	*1.74*
Male	*SNR (db)*	*7.97*	*8.21*
	IS	*1.79*	*1.27*

Table 3: Performance comparison of output SNR and IS measure between MAP and CMAP at 0db input SNR

		MAP	*CMAP*
Female	*SNR (db)*	*10.49*	*10.62*
	IS	*1.11*	*0.95*
Male	*SNR (db)*	*11.41*	*11.62*
	IS	*1.08*	*0.78*

Table 4: Performance comparison of output SNR and IS measure between MAP and CMAP at 5db input SNR

Finally, in order to explore the theoretical performance bounds within the proposed framework and the mutual enhancement between basic MAP estimator and codebook constrained Wiener filter, the proposed method was compared with codebook constrained iterative Wiener filtering method [7]. Clean speech was used for selecting codeword in both methods. Table 5 indicates that the upper limit of CMAP is over IWF. One of reason is that the gain calculation is more accurate in CMAP method by equation (9). Moreover, the incorporation of MAP estimator and codebook constrained Wiener filter results in better input for both parts when encoding by enhanced speech.

		CL_CMAP	*CL_IWF*
0db	*SNR (db)*	*7.93*	*7.25*
	IS	*0.54*	*0.84*
5db	*SNR (db)*	*11.43*	*10.47*
	IS	*0.39*	*0.56*

Table 5: Performance comparison of output SNR and IS measure between MAP and CMAP at 0db and 5db input SNR encoding by clean speech

5. CONCLUSION

A further developed MAP enhancement system has been presented, which integrates codebook constrained Wiener filter into original MAP framework. As shown in experiments, the proposed system can improve quality of enhanced speech on global SNR and I-S distortion measure at low input SNR

6. ACKNOWLEDGEMENT

This research work was supported by National Key Fundamental Research Program of China (G19980300504) and National Natural Science Foundation of China (69835003).

7. REFERENCES

[1] Gong, Y., "Speech recognition in noisy environments: A survey", *Speech communication*, 16:261-291.

[2] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden markov models", *IEEE Trans. on Speech Processing*. 40(4): 725-735, 1992

[3] Y. Ephraim, "On the application of hidden markov models for enhancing noisy speech", *IEEE Trans. On Acoustics, Speech and Signal Processing,* 37(12): 1846-1856, 1989

[4] John H. L. Hansen, Mark A. Clements, "Constrained Iterative speech enhancement with application to speech recognition", IEEE Trans. On Signal Processing, 39(4): 795-805, 1991

[5] Lawrence R. Rabiner, Biing Hwang Juang, "Fundamentals of Speech Recognition", Prentice-hall PTR.

[6] Jae S.Lim, Alan V. Oppenheim, "All-pole modeling of degraded speech", *IEEE Trans. On Acoustics, Speech and Signal Processing*, 26(3): 197-210, 1978.

[7] T. V. Sreenivas, Pradeep Kirnapure, "Codebook constrained Wiener filtering for speech enhancement", *IEEE Trans. On Speech and Audio Processing*, 4(5): 383-389, 1996