SPEECH ENHANCEMENT BY PERCEPTUAL FILTER WITH SEQUENTIAL NOISE PARAMETER ESTIMATION

Te-Won Lee and Kaisheng Yao⋆

Institute for Neural Computation, Univ. of California at San Diego
9500 Gilman Drive, La Jolla, CA 92093-0523
tewon@ucsd.edu, ⋆kyao@ucsd.edu

ABSTRACT

We report a work on speech enhancement that combines sequential noise estimation and perceptual filtering. The sequential estimation employs an extension of the sequential EM-type algorithm. In this algorithm, statistics of clean speech are modeled by hidden Markov models (HMM) and noise is assumed to be Gaussian distributed with a time-varying mean vector (the noise parameter) to be estimated. The estimation process uses a non-linear function that relates speech statistics, noise, and noisy observation. With the estimated noise parameter, subtraction-type algorithm for speech enhancement may be extended to non-stationary environments. In particular, a perceptual filter with frequency masking is constructed with a tradeoff between noise reduction and speech distortion considering the sensitivity of speech recognition systems to speech distortion. Our experiments in speech enhancement and speech recognition in non-stationary noise confirmed that this approach seems promising in improving performances compared to alternative speech enhancement algorithms.

1. INTRODUCTION

The goal of speech enhancement is to recover original speech signals from noisy observations, and has been greatly studied in the past decades [1]. Traditional methods [2][3] usually assume that the statistics of the contaminating noise is known to the enhancement system. In the simplest manner, the noise statistics can be modeled by a simple Gaussian density, which assumes that noise statistics is constant. More detailed modeling of noise statistics may be done by using Gaussian mixture models (GMM). This assumption requires sufficient amount of noise data to learn the noise statistics. Unfortunately, the assumption may not hold in realistic environments, where noise statistics may differ from those during training, thus limiting the performance of these methods.

More recently, researchers have started to investigate speech enhancement in time-varying noisy environments. Proposed methods assume a parametric function to relate speech and background noise, and use sequential methods, e.g., sequential Monte Carlo [4] and Bayesian inference [5]. These methods usually use HMM to model clean speech statistics and a simple noise model with parameters to be estimated from noisy speech.

This paper presents a method for speech enhancement within the above approach but makes the following contributions. First, for the purpose of sequential noise parameter estimation, it is beneficial to have algorithms with fast convergence rate and low computational requirements. Since the noise parameter estimation process involves estimation of hidden speech mixtures/states, (deterministic or stochastic) EM-type algorithms have to be used. This paper applies the sequential Kullback proximal algorithm (SKP) [6], which is a sequential version of the Kullback proximal algorithm [7]. The Kullback proximal algorithm can achieve faster convergence rate than the normal EM algorithm. Moreover, the computational requirement for the SKP algorithm is much less than some alternative methods [4][5]. Second contribution is a subtraction-type speech enhancement algorithm that makes use of the estimated noise statistics. The subtraction-type algorithm is designed to consider a tradeoff between noise reduction and speech distortion, as both have influences on speech recognition system performances. The tradeoff may be achieved by retaining a certain amount of residual noise in the enhanced speech signals. We suggest to employ human auditory properties [8] for the design of the subtraction-type algorithm. Although human auditory properties have been applied to some previous methods, e.g., [3], these previous methods may not be able to handle time-varying noise [3] as their underlying assumption of noise stationarity. With the sequential noise estimation in this paper, we may extend these previous works [3] to time-varying environments. We conducted experiments on speech enhancement and speech recognition in time-varying noisy environment to verify the algorithm and validate its applicability.

2. SPEECH ENHANCEMENT WITH SEQUENTIAL NOISE PARAMETER ESTIMATION

2.1. Time-varying Linear Filtering

Assume speech and noise are uncorrelated. In this context, the power spectrum of the input noisy signal at filter bin j ($1 \leq j \leq J$), $Y_j^{lin}(t)$, can be considered as the summation of the power spectrum from the clean speech signal and the noise, i.e.,

$$Y_j^{lin}(t) = X_j^{lin}(t) + N_j^{lin}(t) \tag{1}$$

where superscript lin denotes linear spectral domain.

The process of subtraction-type enhancement methods is equivalent to attenuating the above spectrum with a time-varying coefficient $H_j(t)$, i.e., $\hat{X}_j^{lin}(t) = H_j(t)X_j^{lin}(t) + H_j(t)N_j^{lin}(t)$. We consider two choices for speech enhancement because of their simplicity.

1. Wiener filter constructs the coefficient as $H_j(t) = \left|1 - \frac{\hat{N}_j^{lin}(t)}{Y_j^{lin}(t)}\right|$, where operator $|\cdot|$ means absolute value, and

$\hat{N}_j^{lin}(t)$ is the estimate of noise power spectrum.

2. Perceptual filter constructs $H_j(t)$ so that the filtered noise power spectrum $H_j(t) \cdot N_j^{lin}(t)$ below the masking threshold of the denoised speech, i.e.,

$$H_j(t) \cdot N_j^{lin}(t) \leq T_j(t) \tag{2}$$

where $T_j(t)$ is the masking threshold of the denoised speech signal, and it is a function of clean speech signal $X_j^{lin}(t)$. Since $X_j^{lin}(t)$ is not directly observed, $\hat{X}_j^{lin}(t) = Y_j^{lin}(t) - \hat{N}_j^{lin}(t)$ is used instead, which makes the masking threshold as a function of the estimated noise component $\hat{N}_j^{lin}(t)$.

Note that both of the filters require estimated noise power spectra.

2.2. Sequential Noise Parameter Estimation

We use superscript l to denote log-spectral domain. Log-spectral vectors of clean speech signals are modeled by the HMM Λ_X. Assuming that the variances of speech spectrum $x_j^l(t) = \log X_j^{lin}(t)$ and noise spectrum $n_j^l(t) = \log N_j^{lin}(t)$ are very small, the following equation may be used to relate speech mean vector μ_{ik}^l in mixture k of state i in Λ_X by

$$\hat{\mu}_{ik}^l(t) = \mu_{ik}^l + \log(1 + \exp(\mu_n^l(t) - \mu_{ik}^l)) \tag{3}$$

where $\mu_n^l(t)$ denotes time-varying mean vector of $(n_1^l(t) \cdots n_J^l(t))^T$. The mean vector is hereafter called as noise parameter. We denote the sequence of noise parameter by $\mathbf{\Lambda}_N(t) = (\mathbf{\Lambda}_N(t-1)\lambda_N(t))$, where $\lambda_N(t) = \mu_n^l(t)$. The transformed $\hat{\mu}_{ik}^l(t)$ represents speech mean vector at state i and mixture k for noisy speech observation more accurately than μ_{ik}^l. The likelihood function is then given as $\log b_{ik}(\mathbf{y}^l(t)|\mathbf{\Lambda}_N(1:t)) = \log P(\mathbf{y}^l(t)|i,k,\Lambda_X,\mathbf{\Lambda}_N(1:t)) = C - \frac{1}{2}(\mathbf{y}^l(t) - \hat{\mu}_{ik}^l(t))^T \mathbf{\Sigma}_{ik}^{-1}(\mathbf{y}^l(t) - \hat{\mu}_{ik}(t))$, where C is not a function of $\mu_n^l(t)$.

In the context of speech enhancement, noise parameters may be obtained by maximum likelihood estimation as

$$\hat{\lambda}_N(t) = \arg\max_{\lambda_N(t)} P(\mathbf{y}^l(1:t)|\mathbf{\Lambda}_N(1:t), \Lambda_X), \tag{4}$$

which involves Eq. (3) to construct the above likelihood function to relate noisy observation $\mathbf{y}^l(1:t)$, noise parameter $\mu_n^l(t)$ in $\mathbf{\Lambda}_N(1:t)$, and speech mean vector μ_{ik}^l in Λ_X. Since the estimation involves hidden speech state sequence $S(t)$ in HMM, EM type algorithm has to be applied.

E-step: Given $\hat{\mathbf{\Lambda}}_N(1:t-1) = (\hat{\lambda}_N(1), \cdots, \hat{\lambda}_N(t-1))$ as the previously estimated noise parameter sequence, calculate the posterior probability $P(S(t)|\mathbf{y}^l(1:t), \Lambda_X, (\hat{\mathbf{\Lambda}}_N(t-1), \hat{\lambda}_N(t-1)))$.

M-step: Obtain time-varying noise parameter by the following objective function,

$$\begin{aligned}\hat{\lambda}_N(t) = \arg\max_{\lambda_N(t)} & Q_t(\hat{\lambda}_N(t-1); \lambda_N(t)) \\ - & (\beta_t - 1) I(\hat{\lambda}_N(t-1); \lambda_N(t))\end{aligned} \tag{5}$$

where the auxiliary function is defined as

$$\begin{aligned}&Q_t(\hat{\lambda}_N(t-1); \lambda_N(t)) = \\ &\sum_{S(t)} P(S(t)|\mathbf{y}^l(1:t), \Lambda_X, (\hat{\mathbf{\Lambda}}_N(t-1), \hat{\lambda}_N(t-1))) \\ &\log\{P(\mathbf{y}^l(1:t), S(t)|\Lambda_X, (\hat{\mathbf{\Lambda}}_N(t-1), \lambda_N(t)))\},\end{aligned} \tag{6}$$

and the Kullback-Leibler (K-L)distance is

$$\begin{aligned}&I(\hat{\lambda}_N(t-1); \lambda_N(t)) = \\ &\sum_{S(t)} P(S(t)|\mathbf{y}^l(1:t), \Lambda_X, (\hat{\mathbf{\Lambda}}_N(t-1), \hat{\lambda}_N(t-1))) \\ &\log \frac{P(S(t)|\mathbf{y}^l(1:t), \Lambda_X, (\hat{\mathbf{\Lambda}}_N(t-1), \hat{\lambda}_N(t-1)))}{P(S(t)|\mathbf{y}^l(1:t), \Lambda_X, (\hat{\mathbf{\Lambda}}_N(t-1), \lambda_N(t)))}.\end{aligned} \tag{7}$$

In Eq. (5), $\beta_t \in R^+$ works as a relaxation factor. The algorithm is called as sequential Kullback proximal algorithm [6], and it may achieve faster convergence rate than the sequential EM algorithm[1].

After manipulations on the second-order Taylor series of Eq. (5), updating of $\lambda_N(t)$ is achieved as[2]

$$\hat{\lambda}_N(t) = \hat{\lambda}_N(t-1) - (\frac{\partial^2 F(t)}{\partial \lambda_N(t)^2})^{-1} (\frac{\partial F(t)}{\partial \lambda_N(t)})|_{\lambda_N(t) = \hat{\lambda}_N(t-1)} \tag{8}$$

where

$$\frac{\partial F(t)}{\partial \lambda_N(t)} = \sum_{ik} \gamma_{ik}(t) \frac{\partial \log b_{ik}(\mathbf{y}^l(t)|\lambda_N(t))}{\partial \lambda_N(t)} \tag{9}$$

and

$$\begin{aligned}\frac{\partial^2 F(t)}{\partial \lambda_N(t)^2} = \beta_t \{ & \rho \cdot C + \sum_{ik} \gamma_{ik}(t) \frac{\partial^2 \log b_{ik}(\mathbf{y}^l(t)|\lambda_N(t))}{\partial \lambda_N(t)^2} \} \\ & + (1 - \beta_t) \cdot \{ -(\frac{\partial F(t)}{\partial \lambda_N(t)})^2 + \sum_{ik} \gamma_{ik}(t) [(\frac{\partial \log b_{ik}(\mathbf{y}^l(t)|\lambda_N(t))}{\partial \lambda_N(t)})^2 \\ & + \frac{\partial^2 \log b_{ik}(\mathbf{y}^l(t)|\lambda_N(t))}{\partial \lambda_N(t)^2}] \}\end{aligned} \tag{10}$$

$C = \sum_{\tau=1}^{t-1} \rho^{t-1-\tau} \sum_{ik} \gamma_{ik}(\tau) \frac{\partial^2 \log b_{sk}(\mathbf{y}^l(\tau)|\lambda_N(\tau))}{\partial \lambda_N(\tau)^2}|_{\lambda_N(\tau) = \hat{\lambda}_N(\tau)}$, and $\gamma_{ik}(t) = P(ik|\mathbf{y}^l(1:t), (\hat{\Lambda}_N(1:t-1), \hat{\lambda}_N(t-1)))$ is given by the normalized likelihood for partial state sequence [6]. $\rho \in (0, 1]$ is forgetting factor. Differentials of log-likelihood w.r.t. noise parameters are $\frac{\partial \log b_{ik}(\mathbf{y}^l(t))}{\partial \lambda_N} = \mathbf{G}_{\lambda_N} \frac{\partial \mu_{ik}^l(t)}{\partial \lambda_N}$ and $\frac{\partial^2 \log b_{ik}(\mathbf{y}^l(t))}{\partial \lambda_N^2} = \mathbf{H}_{\lambda_N} (\frac{\partial \mu_{ik}^l(t)}{\partial \lambda_N})^2 + \mathbf{G}_{\lambda_N} \frac{\partial^2 \mu_{ik}^l(t)}{\partial \lambda_N^2}$, where the jjth element in diagonal matrices $\mathbf{G}_{\lambda_N}$ and $\mathbf{H}_{\lambda_N}$ are respectively given as $G_{\lambda_N jj} = \frac{(\mathbf{y}_j^l(t) - \hat{\mu}_{ikj}(t-1))}{\Sigma_{ikj}^2}$ and $H_{\lambda_N jj} = -\frac{1}{\Sigma_{ikj}^2}$. The jth element in $\frac{\partial \mu_{ik}^l(t)}{\partial \lambda_N}$ and $\frac{\partial^2 \mu_{ik}^l(t)}{\partial \lambda_N^2}$ are $\frac{\exp(\mu_{nj}^l(t) - \mu_{ikj}^l)}{1 + \exp(\mu_{nj}^l(t) - \mu_{ikj}^l)}$ and $\frac{\exp(\mu_{nj}^l(t) - \mu_{ikj}^l)}{(1 + \exp(\mu_{nj}^l(t) - \mu_{ikj}^l))^2}$, respectively.

The algorithm implemented in this paper differs from the work in [6] in that the operations are on features in log-spectral domain instead of cepstral domain as in [6]. Since the primary goal of the work is to enhance speech signals instead of speech recognition, the models Λ_X are trained from log-spectral features. As a result, the number of filter banks J increases from twenties in speech recognition to 65 in the work, and the differentials in the above paragraph have been modified from [6] for log-spectral observations.

[1] When $\beta_t = 1.0$, Eq. (5) corresponds to sequential EM algorithm.

[2] Some derivations are in [6].

2.3. Perceptual Filter

The perceptual filter works in the linear spectral domain. As described in Section 2.1, the perceptual filter employs human auditory property with masking. This is manifested by the different design function for Wiener filter and for the perceptual filter in Eq. (2). Furthermore, it is known that automatic speech recognition is sensitive to nonstationary noise. The more nonstationary the noise is, the easier an insertion error (e.g., false recognition of noise as speech event) may occur. We would have a further requirement on the "stationarity" of the output noise after filtering. Denote the actual noise power spectrum after filtering as $H_j(t)N_j^{lin}(t)$ and the desired noise power spectrum as $\xi^2 N_j^{lin}(t)$ [9]. We constrain the error between actual noise spectrum and desired noise spectrum within a masking threshold,

$$N_j^{lin}(t)(H_j(t) - \xi^2) \leq T_j(t) \tag{11}$$

Note that Eq. (11) is more general than the criterion in Eq. (2), which corresponds to setting $\xi = 0$. This is equivalent to saying that, the filter in Eq. (2) has no desired noise output, whereas setting $\xi > 0$ in Eq. (11) allows residual noise output. This retained residual noise component $\xi^2 N_j^{lin}(t)$ could be helpful to smooth output spectra, thus results in more stationary output noise. Solving Eq. (11) for $H_j(t)$ with constraint $H_j(t) \leq 1.0$ yields the spectral weighting function $H_j(t) = \min\{\frac{T_j(t)}{N_j^{lin}(t)} + \xi^2, 1.0\}$. Since $N_j^{lin}(t)$ has to be estimated, the above function is rewritten as,

$$H_j(t) = \min\{\frac{T_j(t)}{\hat{N}_j^{lin}(t)} + \xi^2, 1.0\} \tag{12}$$

Now, the estimated noise parameter $\hat{\mu}_n^l(t)$ presented in Section 2.2 is transformed into linear-spectral domain by $\hat{N}_j^{lin}(t) = \log \hat{\mu}_{nj}^l(t)$ for each filter bank j. Masking threshold $T_j(t)$ in the above equation is obtained from frequency masking model by [8]. Since the model requires clean speech power, we approximate the clean speech power by output spectrum from Wiener filter, i.e., $\hat{X}_j^{lin}(t) = \max\{Y_j^{lin}(t) - \hat{N}_j^{lin}(t), 0\}$. We thus combined the sequential noise estimation and the filtering function (12) to have a novel perceptual filter for speech enhancement.

2.4. Proposed Speech Enhancement Algorithm

The speech enhancement algorithm is thus a combination of sequential noise parameter estimation and speech enhancement method exploiting masking properties. At each frame t, the algorithm carries out noise parameter estimation in log-spectral domain and perceptual enhancement of noisy speech in time-domain. Noise parameter estimation works in the log-spectral domain with the objective function (5). Speech spectrum is first enhanced by a Wiener filter which is designed with the estimated noise spectrum. The enhanced speech spectrum is used to calculate the masking threshold. The perceptual filter is designed with the masking threshold and the estimated noise parameter by Eq. (12). The perceptual filter then does filtering of the noisy speech.

3. PERFORMANCE EVALUATION

Speech signals were taken from the Aurora 2 database. Speech model was trained from log-spectral speech power on 8840 clean speech utterances, and was a left-to-right HMM with 18 states and 8 Gaussian mixtures in each state. Noise model was a single Gaussian with time-varying mean vector. The window size was 25.0ms with a 10.0ms shift. Number of filter banks J was 65. Contaminating noise includes simulated nonstationary noise (generated by multiplying white noise signals with a time-varying factor in the time domain), White, Babble, and Restaurant noise.

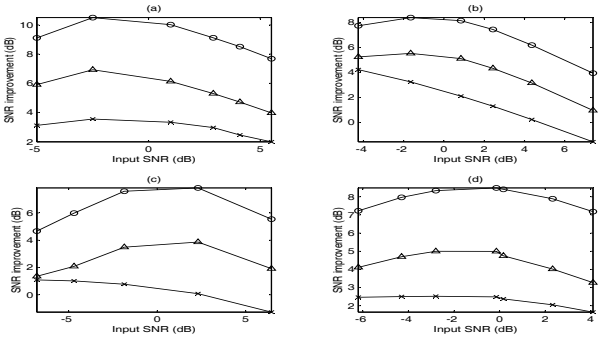

Fig. 1. SNR improvement for various noise types and input SNRs: (a) White noise; (b) Simulated non-stationary noise; (c) Babble noise; (d) Restaurant noise. The tested methods are as follows: $(-o-)$ Perceptual filter with sequential noise parameter estimation, $(-\triangle-)$ Wiener filter with sequential noise parameter estimation, and $(-\times-)$ the traditional Wiener filter.

We compared three systems. The first system, denoted as Baseline, was Wiener filtering implemented according to [2], in which voice activity detection (VAD) was used for decision of speech/noise segments in utterances. The second system, denoted as Known, differed from the first system in the way that the noise parameter was estimated by sequential estimation in Section 2.2. The third system, denoted as Perceptual, was the proposed speech enhancement algorithm in Section 2.4. Relaxation factor β_t in Eq. (5) was set to 0.9.

SNR Improvement: The amount of noise reduction is measured with the Segmental SNR (SegSNR) improvement. Perceptual filter had the flooring constant ξ in Eq. (12) set to zero. Fig. 1 shows the SegSNR improvement obtained from various noise types and at various noise levels. Positive SNR improvements are observed for system "Known" and system "Perceptual" with sequential noise parameter estimation. Such performance differences present the efficacy of sequential noise parameter estimation. Furthermore, system "Perceptual", which is a combination of auditory modeling and sequential noise parameter estimation, has larger SNR improvements than system "Known", which confirms observations by other researchers that incorporation of human auditory property is helpful to achieve improved noise reduction in low SNR conditions.

Speech Spectrograms: Since speech spectrograms provide the structure of the residual noise, we present spectrograms of speech signals after processing by these systems in Fig. 2. The simulated nonstationary noise had SNR at 0.8dB. It is observed that the noise power appeared after 0.4s (almost at the time when the speech segments was occurring). Fig. 2 (b) shows that "Baseline" cannot handle this kind of nonstationarity in the noise signal. The enhanced signal by the system still contains significant noise power in speech segments. On the contrary, with the sequen-

tial noise parameter estimation, the enhanced signal by "Perceptual" has reduced noise power in these speech segments (shown in Fig. 2 (c)).

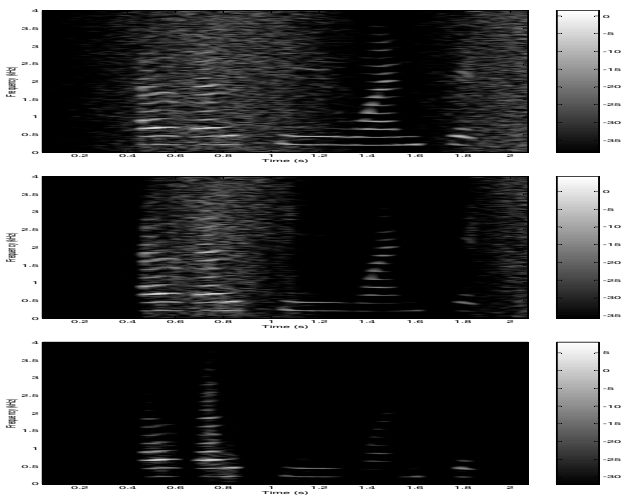

Fig. 2. Example of spectrogram of the signals. (a) Noisy signal (noise is non-stationary and the SNR is 0.8dB.) (b) Enhanced speech by Wiener filter. (c) Enhanced speech by perceptual filter with noise statistics estimated by the proposed algorithm.

Note that some higher frequency components of speech segments are attenuated more strongly than "Baseline" (see segments between 1.4s and 2.0s). Since Eq. (3) assumes small speech and noise variance, lack of speech energy in these higher frequency parts may cause unreliable estimation of noise parameter. As a result, the filter coefficient $H_j(t)$ in Eq. (11) may unnecessarily small.

Recognition Results: Since the speech recognizer is more sensitive to speech distortion, the proposed speech enhancement scheme utilizes a flooring scheme with a proper selection value for ξ in Eq. (12) (in this work, $\xi^2 = 0.05$). Speech models for recognition had 10 states and 3 Gaussian mixtures in each state. Training utterances were clean. For feature extraction, twenty-six filter banks were used. The features were 39-dimensional MFCC + C0 and its first- and second-order derivatives.

Recognition accuracy improvements obtained with the proposed algorithm is summarized in Table 1 along with relative error rate reduction over that achieved by Wiener filtering. In high SNR conditions (SNR $\geq$ 10), properly selected flooring parameter ξ in Eq. (12) in perceptual filter prevents speech from much distortion. This results in relatively higher recognition accuracy than that by traditional Wiener filter. For example, there was a 26% averaged relative error rate reduction (AERR) by perceptual filter over traditional Wiener filter in 20dB noise. With the decrease in SNR, this gain in perceptual filtering becomes marginal, reflecting the difficulty for sequential noise parameter estimation with low SNR environments. The benefit by the sequential noise estimation is apparent in slowly varying nonstationary noise such as Babble noise. Babble noise has an averaged spectrum similar to speech spectrum, making a normal VAD hard to discriminate it from speech. In the noise, "Perceptual" made 27% averaged error rate reduction (AERR) over traditional Wiener filtering. Performance in Car noise by the perceptual filter is also significant.

Table 1. Word Accuracy (in %) in Aurora 2 database achieved by the proposed enhancement system ($\beta_t = 0.9$ and $\rho = 0.995$) in comparison with a system, denoted as Baseline, without the noise parameter estimation method.

	Proposed algorithm				
	Subway	Babble	Car	Exhibition	AERR
Clean	99.39	99.28	99.55	99.02	65.5%
20 dB	98.29	98.57	99.03	98.30	26.3%
15 dB	96.88	97.84	98.06	96.99	14.7%
10 dB	94.07	94.93	95.57	92.85	10.1%
5 dB	88.13	88.00	89.65	85.64	13.2%
0 dB	75.43	72.47	77.13	73.20	11.3%
-5 dB	55.54	51.54	59.42	57.28	0.1%
AERR	15.7%	27.6%	37.3%	16.0%	

4. CONCLUSIONS

We have presented a speech enhancement algorithm that combines a sequential maximum likelihood estimation of the time-varying noise parameter (time-varying mean vector of the noise spectrum) and perceptual filtering. Estimated noise parameter is used to design a perceptual filter, which employs human auditory properties. The speech enhancement algorithm works under time-varying noise conditions. We have conducted experiments in varies noise and SNR situations to verify that the method can improve performances in speech signal enhancement and compared it to alternative approaches.

5. REFERENCES

[1] E. Weinstein, A. V. Oppenheim, M. Feder, and J. R. Buck, "Iterative and sequential algorithms for multisensor signal enhancement," *IEEE Trans. on Signal Processing*, vol. 42, no. 4, pp. 846–859, April 1994.

[2] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," Tech. Rep. ETSI ES 202 050, ETSI, 2002.

[3] N. Virag, "Single channel speech enhancement based on masking properties of the Human auditory system," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, 1999.

[4] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 3, pp. 173–185, 2002.

[5] T. Kristjansson, B. Frey, L. Deng, and A. Acero, "Joint estimation of noise and channel distortion in a generalized em framework," in *ASRU*, 2001.

[6] K. Yao, K. Paliwal, and S. Nakamura, "Noise adaptive speech recognition based on sequential noise parameter estimation," *to appear in Speech Communication, also in pp. 189-192 in ICASSP 2002*.

[7] S. Chrétien and A. O. Hero III, "Kullback proximal point algorithms for maximum–likelihood estimation," *IEEE. Trans. on Information Theory*, vol. 46, no. 5, pp. 1800–1810, August 2000.

[8] J. D. Johnston, "Estimation of perceptual entropy using noise masking," in *ICASSP*, 1988, pp. 2524–2527.

[9] S. Gustafsson, P. Jax, and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," in *ICASSP*, 1998, pp. 397–340.