SPEECH ENHANCEMENT USING MINIMUM MEAN-SQUARE ERROR ESTIMATION AND A POST-FILTER DERIVED FROM VECTOR QUANTIZATION OF CLEAN SPEECH

Jason Wung, Shigeki Miyabe, Biing-Hwang (Fred) Juang

Center for Signal and Image Processing
Georgia Institute of Technology, Atlanta, GA 30332
jason.wung@gatech.edu, {smiyabe3, juang}@ece.gatech.edu

ABSTRACT

In this paper, a novel post-filtering method applied after the logSTSA filter is proposed. Since the post-filter is derived from vector quantization of clean speech database, it has an equivalent effect of imposing clean source spectral constraints on the enhanced speech. When combined with the logSTSA filter, the additional filter can noticeably suppress residual artifacts by effectively lowering the residual white noise of decision-directed estimation as well as reducing the musical noise of maximum likelihood estimation. Compared to the logSTSA enhanced speech, the overall enhanced speech is able to raise the PESQ score by nearly half a point.

Index Terms— Speech enhancement, Minimum mean-square error (MMSE) estimation, Vector quantization

1. INTRODUCTION

The Ephraim-Malah [1] log-spectral short-time amplitude (logSTSA) filter is an estimator that minimizes the mean-square error of the log spectra for speech signal corrupted by Gaussian additive white noise. Although the logSTSA filter is capable of reducing most of the white noise in the noisy speech signal, it has two limitations: 1) the decision-directed (D-D) SNR estimator leaves colorless residual noise; 2) the maximum likelihood (ML) SNR estimator produces an enhanced signal with a SNR higher than the D-D estimator while introducing the well known annoying "musical noise." The musical noise is caused by the lack of spectral constraints during spectral amplitude estimation. Without sensible spectral constraints, spectral components in some frequency bins may be unduly boosted or eliminated, resulting in musical noise. Various approaches that aim at modifying the SNR estimator have been investigated in the past [2] [3]. However, these methods are still under the framework of the logSTSA filter without much additional benefit in alleviating the residual artifacts. To reduce these artifacts, whether residual white noise or musical noise, two approaches can be considered, namely, the modeling of artifacts or the incorporation of source (clean speech) information. Modeling of the artifacts is difficult in that different input SNR levels and different logSTSA filter parameters create a wide range of possibilities. Therefore, this paper focuses on utilizing the statistics of clean speech signals instead.

Since the spectrum of enhanced speech should follow the statistics of clean speech spectra, a vector quantized codebook that contains only clean speech spectra is utilized to impose spectral constraints. Each codeword is a clean model spectrum that can be represented by linear predictor coefficients. With clean model spectra available, the speech enhancement problem transforms into finding the best matching model spectrum given only noisy speech utterance and imposing the spectral constraints on the noisy signal.

Search of the optimal model spectrum is done by spectral distortion measurement as well as the idea drawn from mixture autoregressive hidden Markov model (ARHMM) [4], where each state in the HMM corresponds to each codeword. For the purpose of our investigation, only the state emission probability is utilized while the state transition probability is not imposed. This gives us the insight on testing this enhancement algorithm with a frame by frame analysis approach, which is capable of processing in real time.

The post-filtering method is motivated by the idea of speech signal synthesis based on linear predictive coding (LPC). Traditionally, the signal to be synthesized is passed through an inverse filter to obtain the residual signal or excitation signal. Whether through additional coding of the excitation or not, the excitation is then passed through a filter of the form $\sigma/A(z)$, where σ is the gain and $1/|A(e^{j\omega})|$ is the model spectrum of the signal, to synthesize the speech. Given only noisy signals, finding an optimal model spectral sequence is possible while obtaining a clean residual signal is nontrivial. Therefore, the noisy signal or the logSTSA enhanced signal itself is treated as the residual signal, with each frame of the signal normalized by its own gain in that frame. By post-filtering, the spectral peaks of the noisy signal or the signal estimated from logSTSA filter can be further enhanced, and in the meantime the post-filter helps mask the noise surrounding the spectral peaks.

In Section 2, we review the Ephraim-Malah logSTSA filter with ML and D-D approach. In Section 3, we present the proposed filtering method. Experimental results are presented in Section 4 and conclusion is given in Section 5.

2. MMSE LOG-SPECTRAL AMPLITUDE ESTIMATOR

Let $y(n\mathrm{T})$ denotes the noisy speech samples, where T is the sampling period and n is the sample index. Let $x(n\mathrm{T})$ and $d(n\mathrm{T})$ denote the clean speech and additive noise samples, respectively. Let $Y_k(m)$, $X_k(m)$ and $D_k(m)$ be the k^{th} spectral component, in the m^{th} analysis window, of the noisy signal $y(n\mathrm{T})$, the clean speech signal $x(n\mathrm{T})$ and the noise $d(n\mathrm{T})$, respectively.

Since the clean speech signal is unknown, an estimate of the clean speech spectral component $\hat{X}_k$ by the Ephraim-Malah MMSE logSTSA filter [1] is given by

$$|\hat{X}_k| = \frac{\xi_k}{1+\xi_k} \exp\left\{\frac{1}{2}\int_{\nu_k}^{\infty} \frac{e^{-t}}{t}\,\mathrm{d}t\right\} |Y_k| \tag{1}$$

where ν_k is defined by

$$\nu_k \equiv \frac{\xi_k}{1+\xi_k}\gamma_k \tag{2}$$

and ξ_k, γ_k are defined by

$$\xi_k \equiv \lambda_X(k)/\lambda_D(k), \qquad \textit{a priori} \text{ SNR} \tag{3}$$

$$\gamma_k \equiv |Y_k|^2/\lambda_D(k), \qquad \textit{a posteriori} \text{ SNR} \tag{4}$$

where $\lambda_X(k)$ and $\lambda_D(k)$ denote the variances of the k^{th} spectral components of the clean speech and the noise, respectively.

Since clean speech and noise variances are unknown, two approaches can be used to estimate the *a priori* SNR. The D-D *a priori* SNR estimation [5] is given by

$$\hat{\xi}_{k,D-D}(m) = \alpha \frac{|\hat{X}_k(m-1)|^2}{\lambda_D(k,m-1)} + (1-\alpha)\mathrm{P}\{\gamma_k(m)-1\} \tag{5}$$

where $\hat{X}_k(m-1)$ is the amplitude estimate of the k^{th} signal spectral component in the $(m-1)^{th}$ analysis frame, α is a weighting factor that is set as 0.98 and $\mathrm{P}\{\cdot\}$ is defined as

$$\mathrm{P}\{x\} \equiv \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The name "decision-directed" comes from the fact that the *a priori* SNR is updated based on the previous frame's amplitude estimation.

The ML estimation is based on estimation of signal variance by maximizing the joint conditional PDF, which is given by

$$\hat{\lambda}_{X,ML}(k) = \arg\max_{\lambda_X(k)} \left\{ p(Y_k(m)|\lambda_X(k), \lambda_D(k)) \right\}. \tag{7}$$

This estimator results in the following *a priori* SNR estimator

$$\hat{\xi}_{k,ML}(m) = \begin{cases} \frac{1}{L}\sum_{l=0}^{L-1} \gamma_k(m-l) - 1 & \text{if nonnegative} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where estimation is based on L consecutive frames $Y_k(m) \equiv \{Y_k(m), Y_k(m-1), ..., Y_k(m-L+1)\}$, which are assumed to be statistically independent. The actual implementation is a recursive average [5] given by

$$\bar{\gamma}_k(m) = \alpha\bar{\gamma}_k(m-1) + (1-\alpha)\frac{\gamma_k(m)}{\beta} \tag{9}$$

$$\hat{\xi}_k(m) = \mathrm{P}\{\bar{\gamma}_k(m)-1\} \tag{10}$$

where α and β are specified as 0.725 and 2, respectively.

3. PROPOSED FILTER

Post-filtering is done by passing the gain normalized noisy signal or the logSTSA enhanced signal through a filter $1/A(z)$. However, the spectral peaks of frames with higher SNR may be enhanced too much so that the enhanced speech signal may sound narrowband and additional artifacts may be introduced. To compensate for this undesirable effect, a smoother post-filter $1/A'(z)$ is used for mild suppression of noise as well as retaining the naturalness of the enhanced speech sound. This is obtained by taking the square root of the spectrum $1/A(e^{j\omega})$ in frequency domain, taking the inverse DFT to get the time domain signal and applying the LPC analysis on the time domain signal. This procedure generates the proposed filter $1/A'(z)$ that has a smoother magnitude response than $1/|A(e^{j\omega})|$ while leaving the positions of spectral peaks and valleys unchanged. Sample magnitude responses of the two filters are shown in Fig. 1.

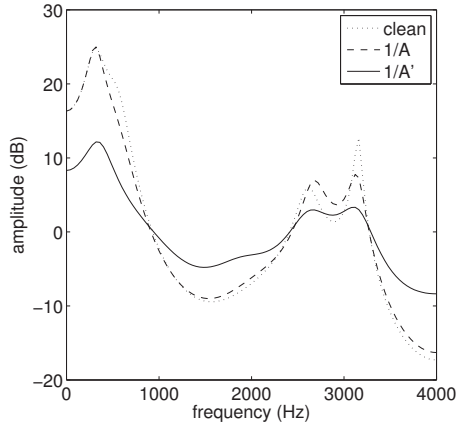

Fig. 1. Magnitude responses of the clean speech envelope, the best matching codeword $1/A(z)$ and the proposed filter $1/A'(z)$

In order to impose spectral constraints on the enhanced speech, a codebook that contains only clean model spectra are trained. Since only the shape of the model spectrum is utilized, a truncated cepstral distance, which is independent of signal gain level, is chosen, and is given by

$$d_c^2(L) = \sum_{n=1}^{L}(c_n - c_n')^2 \tag{11}$$

where c_n is the n^{th} cepstral coefficient and L is the order of the truncated cepstral coefficients.

With clean model spectra available, the enhancement problem narrows down to finding a sequence of best matching spectrum that is closest to the original clean speech spectrum, given a noisy speech signal. An iterative search based on repeatedly applying soft and hard decision estimation is proposed.

The soft decision method is motivated by the idea of HMM-based MMSE estimation [6]. Since the optimal filter for each frame is unknown, all filters are tried, and each filtered signal is assigned a weighting function that takes a form of state emission probability [4] given by

$$f(\mathbf{o},\mathbf{a}) = (2\pi)^{-N/2}\exp\left\{-\frac{1}{2}\delta(\mathbf{o},\mathbf{a})\right\} \tag{12}$$

where N is the frame length, $\mathbf{o}^{\mathrm{T}} = [o[0], o[1], ..., o[N-1]]$ are the observed samples in one frame, $\mathbf{a}^{\mathrm{T}} = [1, a_1, a_2, ..., a_p]$ are the linear predictor coefficients that are derived from the codebook, $\{\cdot\}^{\mathrm{T}}$ denotes matrix transposition, p is the LPC order and $\delta(\mathbf{o},\mathbf{a})$ [7] is given by

$$\delta(\mathbf{o},\mathbf{a}) \equiv r(0)r_a(0) + 2\sum_{n=1}^{p} r(n)r_a(n) \tag{13}$$

where $r_a(n)$ and $r(n)$ are defined as

$$r_a(n) \equiv \sum_{i=0}^{p-n} a_i a_{i+n} \tag{14}$$

$$r(n) \equiv \sum_{i=0}^{N-n-1} o[i]o[i+n]. \tag{15}$$

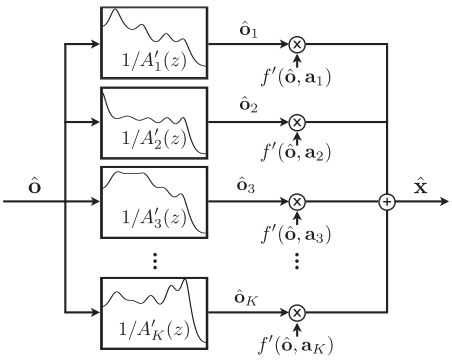

Fig. 2. A block diagram of soft estimation

The term $\delta(\mathbf{o}, \mathbf{a})$ takes the form of $d_{LR} + 1$, where d_{LR} is the likelihood ratio distortion measure, which is given by

$$d_{LR}(\frac{1}{|A(e^{j\omega})|^2}, \frac{1}{|A_p(e^{j\omega})|^2}) = \int_{-\pi}^{\pi} \frac{|A(e^{j\omega})|^2}{|A_p(e^{j\omega})|^2} \frac{\mathrm{d}\omega}{2\pi} - 1 = \frac{\mathbf{a}^t \mathbf{R}_p \mathbf{a}}{\sigma_p^2} - 1 \quad (16)$$

where $\mathbf{a}$ comes from the model spectrum $1/|A(e^{j\omega})|$, $\mathbf{R}_p/\sigma_p^2$ is the gain normalized p^{th} order autocorrelation matrix of the spectrum $1/|A_p(e^{j\omega})|$ that comes from the observed speech samples $\mathbf{o}$. Since $f(\mathbf{o}, \mathbf{a})$ is inversely proportional to d_{LR}, the larger the likelihood ratio distortion d_{LR}, the smaller the weight is applied.

A block diagram of soft estimation is illustrated in Fig. 2. Let $\hat{\mathbf{o}}_i$ denotes the signal enhanced by logSTSA and filtered by $1/A'_i(z)$, where i is the i^{th} entry in the codebook. The soft estimation is given by

$$\hat{\mathbf{x}} = \sum_{i=1}^{K} f'(\hat{\mathbf{o}}, \mathbf{a}_i)\hat{\mathbf{o}}_i \quad (17)$$

where K is the size of the codebook and $f'(\hat{\mathbf{o}}, \mathbf{a}_i)$ is a normalized version of $f(\hat{\mathbf{o}}, \mathbf{a}_i)$ such that $\sum_{i=1}^{K} f'(\hat{\mathbf{o}}, \mathbf{a}_i) = 1$.

The hard decision method, on the other hand, select one candidate codeword from the codebook that gives the lowest distortion measure between the noisy signal or the enhanced signal and the model spectrum based on the cepstral projection measure [7], which is given by

$$d(\mathbf{c}, \mathbf{c}_o) = \|\mathbf{c}_o\| - \mathbf{c}_o^{\mathrm{T}} \mathbf{c}/\|\mathbf{c}\| \quad (18)$$

where $\mathbf{c}_o$ is the noisy or enhanced cepstral coefficients, $\mathbf{c}$ is the codeword and $\|\cdot\|$ is the norm of a vector. This distortion measure has the advantage of being robust to additive noise in spectral comparison. Thus, it is expected to produce a more accurate estimate of the clean model spectral codeword.

A block diagram of the algorithm is shown in Fig. 3. The overall loop is implemented by post-filtering the noisy signal by hard decision, then the enhanced signal is passed through soft decision, and the whole process is iterated. The idea is motivated by two observations. Firstly, whenever a wrong codeword in the hard decision estimation is chosen, the wrong codeword may stay in the same state no matter how many iterations of hard decision is applied. On the other hand, once the correct codeword is chosen, the codeword obtained after additional iterations does not deviate too much from the ideal codeword. Therefore, by utilizing soft decision, additional flexibility is given such that there is a better chance for hard decision to find an optimal match. Secondly, the enhanced signal produced by soft decision is more natural sounding in that, unlike hard decision where a wrong sequence of filters can introduce severe spectral distortion, errors are distributed among several different filters.

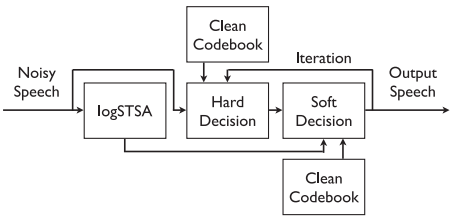

Fig. 3. A block diagram of the overall loop

4. EXPERIMENTAL RESULTS

The experiments were performed using the TIMIT database. Codebook training was performed using 4620 sentences of clean speech and testing was performed using 120 speech utterances. The speech database for testing were different from those used for training. Both male and female speakers were included. Gaussian white noise or pink noise was added to each testing utterance at signal-to-noise ratio (SNR) of -5, 0, 5, 10, 15 and 20 dB. The noise variance estimate of logSTSA was done by averaging sections with only noise.

In our experiments, all speech samples were downsampled to 8kHz prior to training and testing, and a frame size of 20ms with 50% overlap was used. A Hanning window was applied on each frame during training and testing. A 10th order LPC analysis was used and the order of truncated cepstral coefficients was set to be 20. Five codebooks of clean spectral shapes were trained using truncated cepstral distance with 32, 64, 128, 256 and 512 codewords. As the codebook size grows large, however, the enhancement results sounded similar. Thus, speech quality evaluation was performed using a codebook of size 32 for faster computation.

Fig. 4 shows the spectrograms of clean speech, noisy speech and enhanced speech. By comparing logSTSA D-D with logSTSA ML, we can see that logSTSA D-D has a higher background white noise than logSTSA ML, while logSTSA ML has isolated peaks in high frequency region that represents musical noise. It is clearly seen that the proposed method effectively lowers the noise floor in logSTSA D-D and greatly suppresses the intensity of isolated peaks in logSTSA ML, while leaving most speech information intact.

Fig. 5 and Fig. 6 show the PESQ scores of different enhancement methods under two noisy conditions. PESQ, Perceptual Evaluation of Speech Quality [8], is an objective measurement tool that predicts the results of mean opinion score (MOS) in subjective listening tests. The reason why PESQ score is chosen as opposed to SNR measurement is that the quality of speech cannot be directly reflected by SNR measurement. For example, the musical noise in ML estimator is usually considered more annoying than the D-D estimator and this can be observed in Fig. 5 and Fig. 6. However, the ML estimator is better than the D-D estimator is terms of SNR measurement, which does not fully take into account the perceptual quality of enhanced speech. As is observed, the post-filtering method proposed in this paper is able to increase the PESQ score of the logSTSA filter by nearly half a point, which is significant.

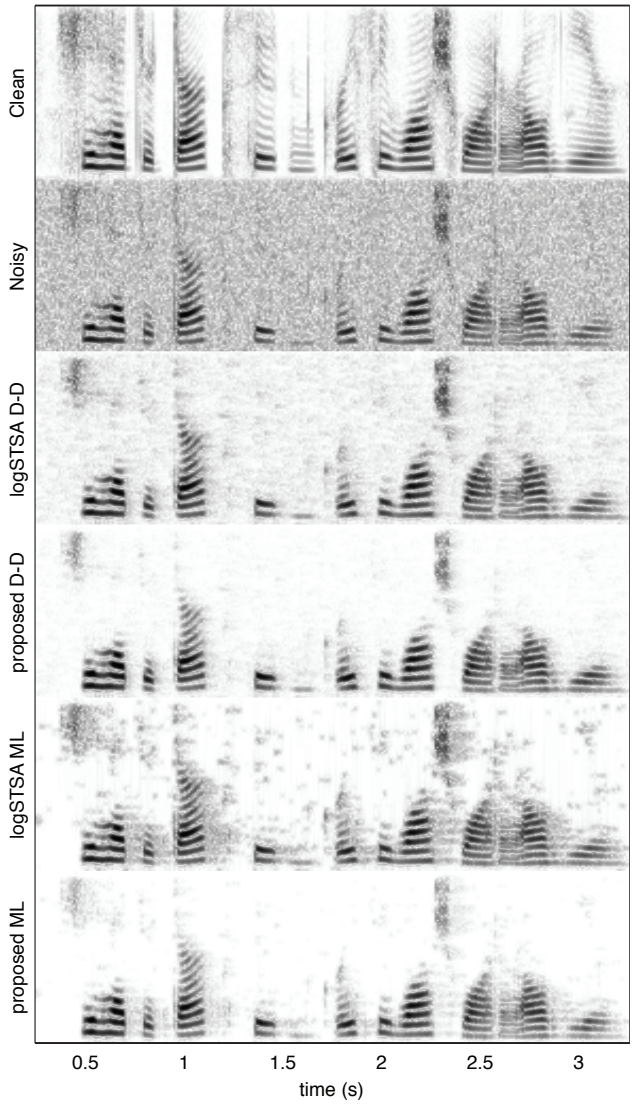

Fig. 4. Spectrograms of enhanced signal at 10 dB input SNR, where the noise type is Guassian white noise

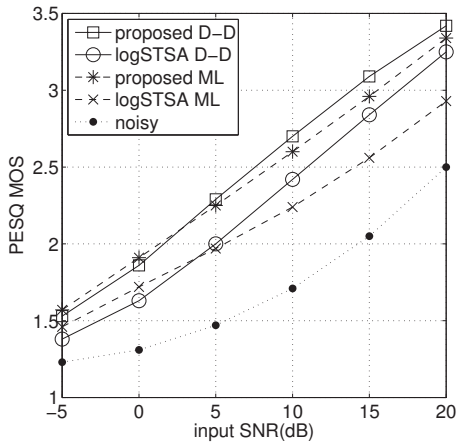

Fig. 5. PESQ scores of speech corrupted by Gaussian white noise

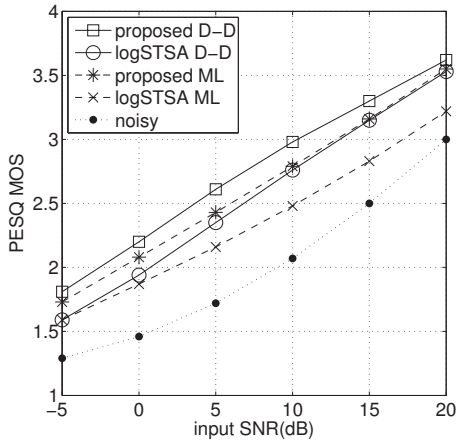

Fig. 6. PESQ scores of speech corrupted by pink noise

5. CONCLUSION

A two pass filtering technique based on logSTSA filter and post-filter for speech enhancement is discussed in this paper. The post-filter is based on vector quantization of the clean speech training database and is equivalent to imposing clean speech spectral constraints on the enhanced signal. Experimental results show that the use of post-filter can effectively reduce the residual white noise in D-D estimation and the musical noise in ML estimation. The results are confirmed by consistently higher PESQ scores.

6. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. ASSP*, vol. ASSP-33, no. 2, pp. 443–445, April 1985.

[2] Yao Ren and M.T. Johnson, "An improved SNR estimator for speech enhancement," *Proc. ICASSP, 2008*, pp. 4901–4904, April 2008.

[3] R. Gemello, F. Mana, and R. De Mori, "Automatic speech recognition with a modified Ephraim-Malah rule," *IEEE Signal Proc. Lett.*, vol. 13, no. 1, pp. 56–59, Janurary 2006.

[4] B.H. Juang and L.R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. ASSP*, vol. ASSP-33, no. 6, pp. 1404–1413, December 1985.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square short-time spectral amplitude estimator," *IEEE Trans. ASSP*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[6] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, October 1992.

[7] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, N.J., PTR Prentice Hall, April 1993.

[8] ITU-T P.862, *Perceptual Evaluation of Speech Quality*, ITU-T, February 2001.