Improved *a posteriori* Speech Presence Probability Estimation Based on Cepstro-Temporal Smoothing and Time-Frequency Correlation

Chao Li, and Wenju Liu

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

{cli,lwj}@nlpr.ia.ac.cn

Abstract

In this paper, we present a novel estimator for the SPP at each time-frequency point in the short-time Fourier transform (STFT) domain. Existing speech presence probability (SPP) estimators cannot perform quite reliably in nonstationary noise environment when applied to a speech enhancement task. To overcome this limitation, we propose a novel SPP estimation method. Firstly, the spectral outliers are eliminated by selectively smoothing the maximum likelihood estimate of *a priori* signal-noise ratio (SNR) in the cepstral domain. Furthermore, an adaptive tracking method for *a priori* SPP is derived by exploiting the strong correlation of speech presence in neighboring frequency bins of consecutive frames. The proposed approach outperforms the state-of-the-art approaches, resulting in less noise leakage and low speech distortions in both stationary and nonstationary noise environments.

Index Terms: speech presence probability, speech enhancement, cepstro-temporal smoothing, time-frequency correlation

1. Introduction

The singlechannel speech enhancement algorithm based on statistical model (e.g., log minimum mean square error (log-MMSE) estimator) can achieve a better performance by exploiting the speech presence probability (SPP) [1][2][3][4]. Additionally, in multichannel speech enhancement, SPP is used to reduce effectively spatially white and coherent additive noise [5].

The estimation of SPP at each time-frequency point in short-time Fourier transform (STFT) domain is a challenging task in adverse environments. The estimator as proposed in [1] brings only little speech distortion, but does not yield SPP estimates close to zero at time-frequency bins where speech is absent, e.g., between the harmonics of voiced speech or even in speech pauses. The estimator as proposed in [2] adopts hard- or soft- decision to improve the SPP estimation, but the SPP estimate in speech absence is still high, and some speech distortion is introduced. Estimators like [3][4] overcome these problems to some extent, but still exhibit severe noise leakage in nonstationary noise (e.g., babble noise). Moreover, the estimator in [4] based on fixed *prior*s usually introduces a large amount of outliers in the estimate of the SPP that are related to *musical noise*.

In this paper, we improve *a posteriori* SPP estimation using a novel *a priori* signal-noise ratio (SNR) estimation based on selective cepstro-temporal smoothing and a novel *a priori* SPP estimation based on time-frequency correlation. It has been shown in recent research [6] that a temporal smoothing of the cepstral representation of certain spectral quantities performs better than a smoothing in the frequency domain. Suitable smoothing factor enables the elimination of spectral outliers in original estimated *a priori* SNR caused by noise burst in nonstationary noise environment. Furthermore, three weighted parameters are computed by a soft-decision approach using the frequency distribution of estimated *a priori* SNR. The bias of *a priori* SPP estimation can be corrected via smoothing over frame using the correlation of speech presence in consecutive frames. Finally, the experimental results verify the effectiveness of the proposed method compared with the state-of-the-art methods.

The rest of this paper is organized as follows: Section 2 introduces generalized SPP estimator and reviews the state-of-the-art *a priori* SNR estimators and *a priori* SPP estimators. Section 3 presents the improved *a posteriori* SPP estimation. Section 4 describes the comparison experiments and discusses the results. Section 5 concludes this paper.

2. Generalized SPP Estimator and Review

In the STFT domain, we assume an additive mixture of clean speech $X_k(l)$ and the noise $D_k(l)$. Here k is the frequency bin index, l is the time frame index. Given two hypotheses, $H_0^k(l)$ and $H_1^k(l)$, which indicate speech absence and presence respectively [1],

$$\begin{aligned} H_0^k(l) &: Y_k(l) = D_k(l), \\ H_1^k(l) &: Y_k(l) = X_k(l) + D_k(l). \end{aligned} \tag{1}$$

We assume that the STFT coefficients, for both speech and noise, are complex Gaussian variables, and apply Bayes rule for the conditional SPP, one obtains *a posteriori* SPP [3]

$$\rho_k(l) \triangleq P(H_1^k(l) | Y_k(l)) = \frac{\Lambda_k(l)}{1+\Lambda_k(l)}, \tag{2}$$

where, $\Lambda_k(l)$ is the generalized likelihood ratio (GLR) defined by

$$\Lambda_k(l) = \frac{q_k(l)}{1-q_k(l)} \left(\frac{1}{1+\xi_k(l)}\right) \exp\left(\frac{\xi_k(l)\gamma_k(l)}{1+\xi_k(l)}\right), \tag{3}$$

where, $\xi_k(l)$ and $\gamma_k(l)$ represent the *a priori* and *a posteriori* SNRs [1], $q_k(l)$is the *a priori* SPP, i.e., $q_k(l) = P(H_1^k(l))$.

The *a posteriori* SNR can be easy calculated by

$$\gamma_k(l) = \frac{|Y_k(l)|^2}{\lambda_{d,k}(l)}, \tag{4}$$

where $\lambda_{d,k}(l)$ is the estimation of noise power, obtained during periods of silence.

Therefore, the SPP estimate depends mostly on the estimation of the *a priori* SNR $\xi_k(l)$ and the *a priori* SPP $q_k(l)$.

- **The estimation of the** *a priori* **SNR**. In existing approaches, the *a priori* SNR is mostly estimated using the *decision-directed* approach as proposed in [1]. However, since the *decision-directed* SNR estimator is sensitive to rising spectral amplitudes, besides of speech onsets, it also respond to noise bursts that are not tracked by noise estimation algorithm. Therefore, noise bursts will cause rising *a priori* SNR estimates, and thus produce outliers in the clean-speech estimate that are perceived as *musical noise* [6].
- **The estimation of the** *a priori* **SPP**. In [1], $q_k(l)$ is empirically set to 0.8 based on listening test. In running speech, however, we would expect $q_k(l)$ to vary with time and frequency, depending on the words spoken. In [2], a binary speech-presence decision is obtained based on a comparison of the *a posteriori* SNR against a threshold, and $q_k(l)$ is obtained by smoothing this binary decision over past frames. In [3], $q_k(l)$ is obtained by integrating two SPP estimates based on local and global spectral smoothing. However, these methods never exploit the strong correlation of speech presence in frequency and time simultaneously.

3. A Novel Estimation for *a posteriori* SPP

3.1. *A priori* SNR estimation

In order to avoid the annoying outliers in the estimate of *a priori* SNR, smooth processing is absolutely necessary. The approach taken to smooth the cepstral representation of certain spectral quantities is similar to the method proposed in [6].

From the maximum likelihood (ML) SNR estimate,

$$\xi_k^{\mathrm{ml}}(l) = \gamma_k(l) - 1, \tag{5}$$

we compute the speech power,

$$\lambda_{x,k}^{\mathrm{ml}}(l) = \lambda_{d,k}(l) \max\left\{\xi_k^{\mathrm{ml}}(l), \xi_{\min}^{\mathrm{ml}}\right\}, \tag{6}$$

where $\xi_{\min}^{\mathrm{ml}} > 0$ is a small lower bound. The speech power is transformed into the cepstral domain,

$$\lambda_{x,p}^{\mathrm{ml,ceps}}(l) = \mathrm{IDFT}\left\{\log\left(\lambda_{x,k}^{\mathrm{ml}}(l)\right)|_{k=0,\cdots,M-1}\right\}, \tag{7}$$

recursively smoothed over time,

$$\lambda_{x,p}^{\mathrm{ceps}}(l) = \alpha_p(l)\lambda_{x,p}^{\mathrm{ceps}}(l-1) + (1-\alpha_p(l))\lambda_{x,p}^{\mathrm{ml,ceps}}(l), \tag{8}$$

and transformed back into the frequency domain,

$$\bar{\lambda}_{x,k}(l) = \exp\left(\kappa + \mathrm{DFT}\left\{\lambda_{x,p}^{\mathrm{ceps}}(l)\,|_{p=0,\cdots,M-1}\right\}\right). \tag{9}$$

In (7)(8)(9), p is the cepstrum index, M is the length of discrete Fourier transform (DFT) and inverse DFT (IDFT), $\alpha_p(l)$ is the smoothing factor, κ is the correction constant for unbiased smoothing [6].

With the flooring, $\xi_{\min}$, the final *a priori* SNR estimate is computed as

$$\bar{\xi}_k(l) = \max\left\{\frac{\bar{\lambda}_{x,k}(l)}{\lambda_{d,k}(l)}, \xi_{\min}\right\}. \tag{10}$$

The smoothing factor $\alpha_p(l)$ in (8) should be chosen to be close to zero for speech related cepstral coefficients and close to one for the remaining coefficient, as follows:

$$\alpha_p(l) = \begin{cases} \alpha_{\mathrm{pitch}} & \text{if } p \in \mathrm{Q}_{\mathrm{pitch}}(l) \\ \beta\alpha_p(l-1) + (1-\beta)\alpha_p^{\mathrm{con}} & \text{else,} \end{cases} \tag{11}$$

where $\mathrm{Q}_{\mathrm{pitch}}(l)$ is a set of adjacent cepstral bins that are most likely to represent the fundamental frequency, $F_0(l)$, and α_{pitch} is the low smoothing constant for these bins, β is a forgetting factor, and the stationary values α_p^{con} is preset so that appropriate smoothing is adopted in each cepstral bin.

The cepstral index $p_{\mathrm{pitch}}(l)$ that most likely represents $F_0(l)$ is found via a maximum search in a given range between p_{low} and p_{high}[7],

$$p_{\mathrm{pitch}}(l) = \arg\max_p \left\{\lambda_{x,p}^{\mathrm{ml,ceps}}(l)\,|p_{\mathrm{low}} \le p \le p_{\mathrm{high}}\right\}. \tag{12}$$

In order to make sure that (12) yields meaningful results, we utilize three criteria to detect voiced speech sounds. Firstly, the found cepstral peak should be higher than a threshold, λ_x^{thr}. Secondly, the sum of cepstral coefficients in a small range, Δ_p around $p_{\mathrm{pitch}}(l)$, should be obviously higher than the surrounding counterpart, as voiced speech sounds have a relative high energy. Finally, the cepstral coefficients at $p = 1$ should be positive, as the voiced speech has more energy at low frequencies. Thus, $\mathrm{Q}_{\mathrm{pitch}}(l)$ can be gained as

$$\begin{array}{l} \text{if } \left\{\begin{array}{c} \lambda_{x,p_{\mathrm{pitch}}}^{\mathrm{ml,ceps}}(l) \ge \lambda_x^{\mathrm{thr}} \\ \sum_{m=-\Delta_p}^{\Delta_p} \lambda_{x,p_{\mathrm{pitch}}+m}^{\mathrm{ml,ceps}}(l) \ge \frac{1}{2} \sum_{m=-2\Delta_p}^{2\Delta_p} \lambda_{x,p_{\mathrm{pitch}}+m}^{\mathrm{ml,ceps}}(l) \\ \lambda_{x,1}^{\mathrm{ml,ceps}}(l) > 0 \end{array}\right. \\ \quad \mathrm{Q}_{\mathrm{pitch}}(l) = \mathrm{Q'}_{\mathrm{pitch}}(l) \quad \text{voiced speech} \\ \text{else} \\ \quad \mathrm{Q}_{\mathrm{pitch}}(l) = \emptyset \\ \text{end} \end{array} \tag{13}$$

where $\mathrm{Q'}_{\mathrm{pitch}}(l) = \{p_{\mathrm{pitch}}(l) - \Delta_p, \cdots, p_{\mathrm{pitch}}(l) + \Delta_p\}$ is the range of cepstral bins representing $F_0(l)$, and Δ_p is a small margin.

This temporal smoothing method of the cepstral representation of certain spectral quantities enables the elimination of spectral outliers in original estimated *a priori* SNR caused by noise bursts in nonstationary noise environment.

3.2. *A priori* SPP estimation

In order to take full advantage of the strong correlation of speech presence in frequency and time simultaneously, we combine soft-decision approach with the correlation of SPP in consecutive frames.

With the smoothed *a priori* SNR estimates, obtained by (10), we obtain the local, global and frame averages respectively,

$$\begin{aligned} \zeta_{\mathrm{local},k}(l) &= \underset{-j_{\mathrm{local}} \le m \le j_{\mathrm{local}}}{\mathrm{mean}} \left\{\bar{\xi}_{k+m}(l)\right\}, \\ \zeta_{\mathrm{global},k}(l) &= \underset{-j_{\mathrm{global}} \le m \le j_{\mathrm{global}}}{\mathrm{mean}} \left\{\bar{\xi}_{k+m}(l)\right\}, \\ \zeta_{\mathrm{frame}}(l) &= \underset{0 \le k \le M-1}{\mathrm{mean}} \left\{\bar{\xi}_k(l)\right\}, \end{aligned} \tag{14}$$

where j_{local} and j_{global} is the length parameter of averaging operation. Otherwise, we define three parameters, $P_{\mathrm{local},k}(l)$, $P_{\mathrm{global},k}(l)$ and $P_{\mathrm{frame}}(l)$ representing the relation of the above averages and the likelihood as follows [3]:

$$P_{\vartheta,k}(l) = \begin{cases} 0 & \text{if } \zeta_{\vartheta,k}(l) \le \zeta_{\min}^{\mathrm{SPP}} \\ 1 & \text{if } \zeta_{\vartheta,k}(l) \ge \zeta_{\max}^{\mathrm{SPP}} \\ \frac{\log\left(\zeta_{\vartheta,k}(l)/\zeta_{\min}^{\mathrm{SPP}}\right)}{\log\left(\zeta_{\max}^{\mathrm{SPP}}/\zeta_{\min}^{\mathrm{SPP}}\right)} & \text{else,} \end{cases} \tag{15}$$

where $\zeta_{\max}^{\text{SPP}}$ and $\zeta_{\min}^{\text{SPP}}$ are empirical constants, and the subscript ϑ designates either "local", "global" or "frame". We estimate the original *a priori* SPP using the following form:

$$\hat{q}_k(l) = P_{\text{local},k}^{w_{\text{local}}}(l) P_{\text{global},k}^{w_{\text{global}}}(l) P_{\text{frame}}^{w_{\text{frame}}}(l), \qquad (16)$$

where $0 < w_{\text{local}} < w_{\text{global}} < w_{\text{frame}} < 1$ are the weighted values of three parameters, which should be chosen to be close to zero for dominant parameter and close to one for supporting parameter.

We can finally determine the *a priori* SPP by smoothing the values with the *a posteriori* SPP in last frame,

$$\bar{q}_k(l) = \alpha_q \rho_k(l-1) + (1-\alpha_q)\hat{q}_k(l), \qquad (17)$$

where $0 < \alpha_q < 1$ is a smoothing constant.

Substituting $\bar{\xi}_k(l)$ and $\bar{q}_k(l)$ into (2) and (3), and after some algebraic manipulations, we express the *a posteriori* SPP estimator as:

$$\rho_k(l) = \frac{\bar{q}_k(l)}{\bar{q}_k(l) + (1-\bar{q}_k(l))(1+\bar{\xi}_k(l))e^{-\frac{\bar{\xi}_k(l)\gamma_k(l)}{1+\bar{\xi}_k(l)}}}. \qquad (18)$$

4. Implementation and Evaluation

To evaluate the performance of the proposed approach, the approaches presented in [2][3][4] are selected for comparison. For the evaluation we implement the *a posteriori* SPP estimator in the log-MMSE algorithm as proposed in [3]. The spectral gain function is

$$G_k(l) = \left(G_k^{LSA}(l)\right)^{\rho_k(l)} (G_{\min})^{1-\rho_k(l)}, \qquad (19)$$

where $G_k^{LSA}(l)$ is the original log-MMSE gain function, and $G_{\min}$ is a small value. Parameters values used in implementation of the proposed approach are summarized in Tab.1.

Table 1: *Parameter values for the proposed approach.*

Spectral analysis and synthesis		
$f_s = 16$ KHz	$M = 512$	$G_{\min} = -20$ dB
32 ms hamming window and 50% overlap		
A priori SNR estimation		
$\xi_{\min}^{\text{ml}} = -25$ dB	$\xi_{\min} = -27$ dB	$\kappa = 0.2886$
$\alpha_{\text{pitch}} = 0.2$	$\beta = 0.9$	$\lambda_x^{\text{thr}} = 0.25$
$p_{\text{low}} = 53$	$p_{\text{high}} = 228$	$\Delta_p = 2$
$\alpha_p^{\text{con}} = \begin{cases} 0.2 & \text{if } q \in \{0,\cdots,2\} \\ 0.65 & \text{if } q \in \{3,\cdots,23\} \\ 0.97 & \text{if } q \in \{24,\cdots,256\} \end{cases}$		
A priori SPP estimation		
$j_{\text{local}} = 0$	$j_{\text{global}} = 12$	
$\zeta_{\max}^{\text{SPP}} = -5$ dB	$\zeta_{\min}^{\text{SPP}} = -10$ dB	$\alpha_q = 0.5$
$w_{\text{local}} = 0.6$	$w_{\text{global}} = 0.9$	$w_{\text{frame}} = 1$

We process 30 utterances selected from TIMIT database [8](15 male, 15 female). The utterances are corrupted by three different noise types taken from Noisex92 [9], (white Gaussian noise, F16 cockpit noise, and babble noise), at five different input segmental SNRs (SegSNR) between -10 to 10 dB.

For evaluation, speech distortion (SD) and noise leakage (NL) measures as introduced in [4] are adopted to evaluate the SPP estimators. The SD measure indicates the percentage of the speech energy that the SPP estimator misses and is related to the miss-hit rate. The NL measure indicates the percentage of noise energy that is not attenuated by the SPP estimator and is related to the false-alarm rate. Additionally, the improvement of the segmental SNR (segSNRI) is adopted as the objective measure to denote the quality of enhanced speech [10].

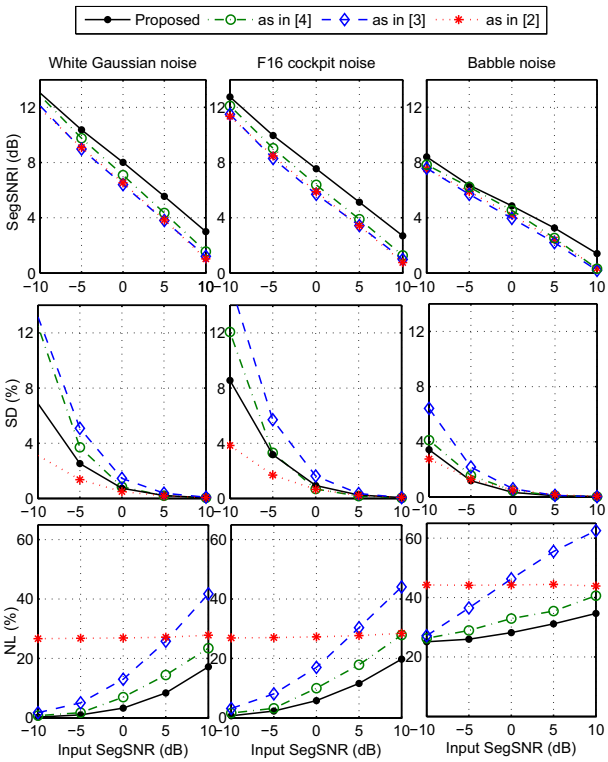

Figure 1: *Comparison results of various SPP estimators in terms of SegSNRI (top), SD (middle), and NL (bottom) averaged over 30 TIMIT sentences for white Gaussian noise (left), F16 cockpit noise (middle), and babble noise (right).*

Fig.1 shows the comparison results of SPP estimators in [2][3][4] and the proposed. The proposed approach yields the lowest NL, while yields similar or lower SD than estimators in [3][4]. The estimator in [2] exhibits lower SD and higher NL, because it cannot achieve values close to zero in speech absence. Simultaneously, especially under high input SegSNR condition, the proposed estimator improves segSNRI measures significantly in all noise, which confirms that the proposed method resulted in better speech quality.

In Fig.2, the resulting SPP estimates are shown for speech disturbed by additive mixed noise of white and babble noise at 0 dB input SegSNR. The estimator in [2] does not achieve SPP estimates close to zero in speech absence. The estimator in [3] exhibits a large miss-hit ratio and a large false-alarm ratio. These undesired behaviors are overcome by the estimator in [4] and the proposed estimator. The drawback of [4] is that a large amount of annoying outliers exist in the SPP estimate, which related to *musical noise*, especially in the speech-silent frames. The proposed estimator notably preserves plosives, vowels and envelop of fricatives. Simultaneously, the amount of outliers, due to narrowband bursts in nonstationary noise, is reduced significantly. Subjective tests also reveal that the proposed method improves intelligibility and comfort quality to some extent.

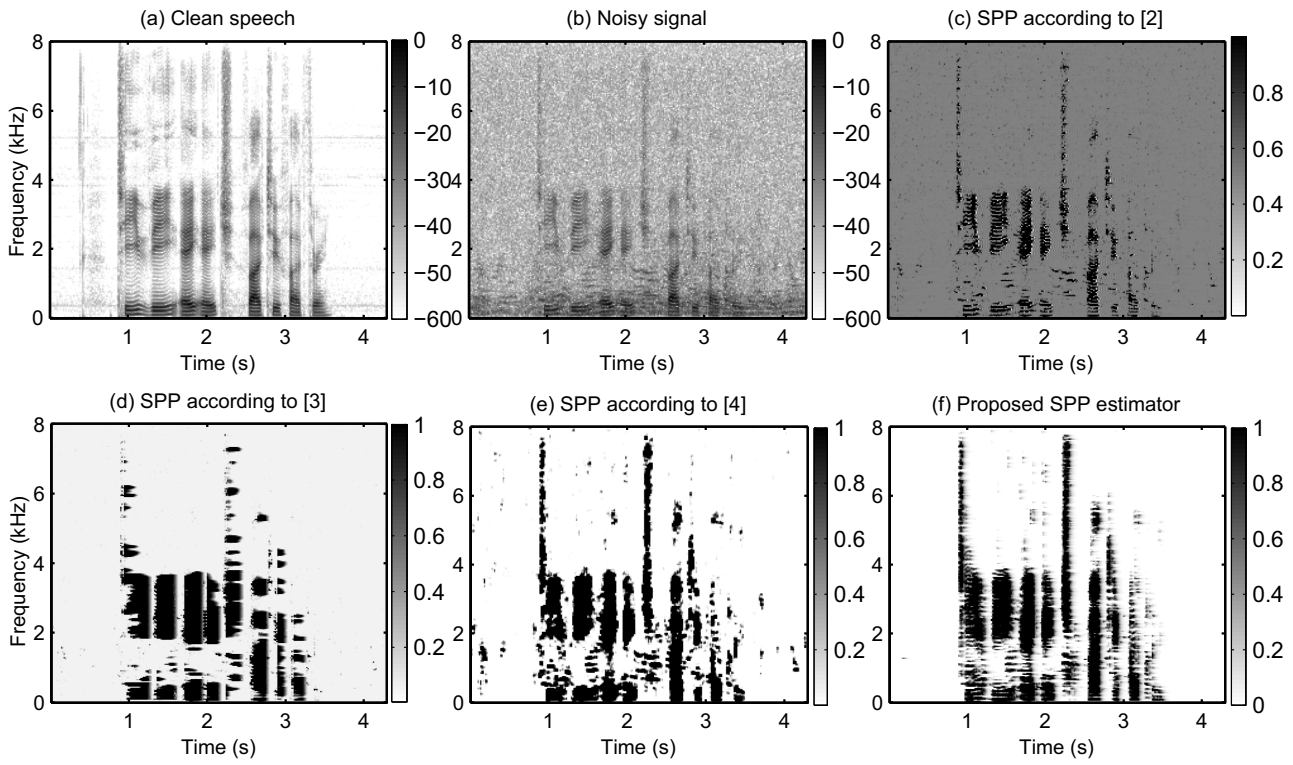

Figure 2: *Spectrograms of clean speech (a) and noisy signal (b), and the resulting SPP estimates (c)-(f) by approaches in [2][3][4] and the proposed, for speech from a male speaker disturbed by additive mixed noise of white and babble at 0 dB input SegSNR.*

5. Conclusions

In this paper, we improve *a posteriori* SPP estimation at each time-frequency point in the STFT domain, based on two essential algorithms: the *a priori* SNR estimation based on selective cepstro-temporal smoothing and the *a priori* SPP estimation based on the time-frequency correlation. Comparing to the existing SPP estimators, the proposed estimator yields lower miss-hit ratio and lower false-alarm ratio in nonstationary noise as well as stationary noise. Furthermore, the amount of spectral outliers due to narrowband noise bursts in nonstationary noise is reduced significantly. Experiment results indicate that the proposed approach can achieve lower or similar speech distortion and lower noise leakage comparing to existing estimators. Simultaneously, a consistent improvement of segmental SNR is achieved, when the proposed estimator is integrated into a speech enhancement system.

6. Acknowledgements

This work was supported in part by the National Nature Science Foundation of China (No.60675026, No.60121302, No.90820011), and the National Grand Fundamental Research 973 Program of China (No.2004CB318105).

7. References

[1] Ephraim. Y, and Malah. D, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", *IEEE Transactions on Audio, Speech, Signal Processing*, 23(2): 443-445, 1985.

[2] Soon. I, Koh. S, and Yeo. C, "Improved noise suppression filter using self-adaptive estimator of probability of speech absence", *Signal Processing*, 75: 151-159, 1999.

[3] Cohen. I, "Optimal speech enhancement under signal presence uncertainty using log-spectra amplitude estimator", *IEEE Signal Processing Letter*, 9(4): 113-116, 2002.

[4] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors", *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5): 910-919, 2008.

[5] M. Souden, J. Chen, J. Benesty, and S. Affes "Gaussian Model-Based Multichannel Speech Presence Probability", *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5): 1072-1077, 2010.

[6] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing", *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2008.

[7] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter", *IEEE Transactions on Audio, Speech, Signal Processing*, 28(2): 137-145, 1980.

[8] J.S. Garofolo, "DARPA TIMIT: an acoustic phonetic continuous speech database", *National Institude of Standards and Technology (NIST)*, 1988.

[9] A. Varga, H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to studythe ePect of additive noise on speech recognition systems", *Speech Communication*, 12(3): 247-251, 1993.

[10] Y. Hu, and P. C. Loizou, "Evaluation of objuective quality measures for speech enhancement ", *IEEE transactions on Speech Audio Processing*, 16: 229-238, 2008.