SPEECH ENHANCEMENT USING SPARSE CODE SHRINKAGE AND GLOBAL SOFT DECISION

Changkyu Choi, Seungho Choi, and Sang-Ryong Kim

Human & Computer Interaction Laboratory
Samsung Advanced Institute of Technology
San 14-1, Nongseo-ri, Kiheung-eup, Yongin-city, Kyonggi-do 449-712, Korea
Email: {flyers, shchoi, srkim}@sait.samsung.co.kr, http://hci.sait.samsung.co.kr/∼flyers

ABSTRACT

This paper relates to a method of enhancing speech quality by eliminating noise in speech presence intervals as well as in speech absence intervals based on speech absence probability. To determine the speech presence and absence intervals, we utilize the global soft decision. This decision makes the estimated statistical parameters of signal density models more reliable. Based on these parameters the noise suppressor equipped with sparse code shrinkage functions reduces noise considerably in real-time.

1. INTRODUCTION

The performance of a speech recognition system degrades when there is a mismatch between the training clean speech and the noisy input speech that is to be recognized. The situation is even worse in speech coding systems. The quality degradation gets worse in the speech processed by speech coding systems than in the noisy input speech. A conventional approach to alleviate this problem is the spectral enhancement technique. Spectral enhancement is used to estimate a noise spectrum in noise intervals where speech signals are not present, and in turn to improve a speech spectrum in a predetermined speech interval based on the noise spectrum estimate. Speech presence and absence intervals are determined from the uncorrelated statistical models of the spectra of clean speech and noise [1] [2].

In this paper, we try to lay a bridge between statistical speech processing for conventional speech enhancement and sparse code shrinkage which was originally considered for image de-noising [3]. There have been attempts to enhance noisy speech based on the sparse code shrinkage technique [4] [5]. However, both works pay little attention to the estimation of parameters needed for the calculation of shrinkage functions and in consequence they are proven unsuitable for on-line computation. Because any kind of optimal estimator cannot be obtained in closed-form for a generalized Gaussian density model, a closed-form solution of shrinkage function was obtained using a special kind of density model [3]. To make the problem at hand tractable, we adopt this shrinkage function as the noise suppressor for a generalized Gaussian density model. Then, we focus on the reliable estimation of statistical parameters based on global soft decision which decides whether the current frame is speech-absent or not. By doing so, the speech enhancement system works in real-time, and noise is considerably reduced.

This work was partly supported by the Critical Technology-21 Program of Korean Ministry of Science and Technology. The authors wish to thank Prof. Te-Won Lee for fruitful and helpful discussions in the course of this work.

2. SPEECH ENHANCEMENT

Referring to Fig. 1, the speech enhancement system involves a pre-processing step, a speech enhancement step and a post-processing step. In the pre-processing step, an input speech-plus-noise(noisy) signal in the time domain is pre-emphasized and subjected to an Independent Component Analysis Basis Function Transform(ICABFT). As a result, we get a noisy speech coefficient vector $\mathbf{Y}(m)$. In the speech enhancement step, the global speech absence probability (SAP) is calculated based on estimated noisy speech and noise parameters. The term 'global' comes from the fact that the decision, whether the speech is present or not, is performed globally using the coefficients of all the ICA basis functions in a given time frame. Noise parameters are updated only when the global SAP exceeds a predetermined threshold. Using predicted speech parameters and updated noise parameters we apply the shrinkage function to each component of $\mathbf{Y}(m)$ to enhance the noisy speech. This results in the enhanced speech coefficient vector $\mathbf{S}(m)$. In the post-processing step, $\mathbf{S}(m)$ undergoes a sequence of operations such as inverse ICABFT, overlap-and-add operation and de-emphasis, resulting in an enhanced speech signal in the time domain.

2.1. Pre-Processing and ICA basis functions

We assume that an input noisy speech signal is $y(n)$ and the signal of an m-th frame is $y_m(n)$, which is one of the frames obtained by segmentation of the signal $y(n)$. The signal $\hat{y}_m(n)$ and $\hat{y}_m(D+n)$, which is pre-emphasized and overlaps with the rear portion of the preceding frame by pre-emphasis, are given by

$$\begin{array}{ll} \hat{y}_m(n) & = \hat{y}_{m-1}(L+n), \quad 0 \le n < D \\ \hat{y}_m(D+n) & = y_m(n) - \zeta \cdot y_m(n-1), \quad 0 \le n < L \end{array}, \quad (1)$$

where D is the overlap length with the preceding frame, L is the length of frame shift and ζ is the pre-emphasis parameter. Then, prior to the ICABFT, the pre-emphasized input speech signal is subjected to the windowing given by

$$\tilde{y}_m(n) = \begin{cases} \hat{y}_m(n) sin^2(\frac{\pi(n+0.5)}{2D}), & 0 \le n < D \\ \hat{y}_m(n), & D \le n < L \\ \hat{y}_m(n) sin^2(\frac{\pi(n-L+D+0.5)}{2D}), & L \le n < M \end{cases}, \quad (2)$$

where $M = D + L$ is the size of ICABFT. The obtained signal $\tilde{y}_m(n)$ is converted into a signal in ICA basis domain by ICABFT given by

$$\mathbf{Y}(m) = \mathbf{A}_{oo}^T \cdot [\tilde{y}_m(0)\ \tilde{y}_m(1)\ \cdots\ \tilde{y}_m(M-1)]^T, \quad (3)$$

where $\mathbf{A}_{oo}$ is a 'frequency-ordered' and orthogonalized version of the matrix $\mathbf{A}$, columns of which are ICA basis functions.

ICA basis functions can be obtained by various algorithms [6], [7], [8] with the clean speech data pre-processed as described above. After estimating the ICA basis function matrix $\mathbf{A}$, we ordered the basis functions by the location of their power spectral densities, resulting in a frequency-ordered basis function matrix, $\mathbf{A}_o$. The term 'frequency-ordered' means that the basis functions having power spectral densities at lower frequency portions appear earlier in $\mathbf{A}_o$ than the ones at higher frequency portions. Then, we orthogonalize this by the following

$$\mathbf{A}_{oo} = \mathbf{A}_o(\mathbf{A}_o^T\mathbf{A}_o)^{-1/2}. \quad (4)$$

Because $\mathbf{A}_{oo}$ is orthogonal, the noise is still Gaussian in the ICA basis domain. Therefore, the ICABFT is used to obtain the M-dimensional coefficient vector $\mathbf{Y}(m)$, in which speech components are sparse while the statistical properties of noise components are preserved.

The pre-processing step involving overlapping segmentation, pre-emphasis and windowing seems to be needless in view of sparse coding. However, the pre-processing has an important meaning for speech signals which have both the inter-frame correlations in the time domain and the inter-frequency correlations in the frequency domain. In particular, a pre-emphasis of high frequencies is required to obtain similar spectral amplitude for all formants. This is because high frequency formants, although possessing relevant information, have smaller amplitude with respect to low frequency formants. Fig. 2 shows the plot of power spectral densities contained in frequency-ordered and orthogonalized ICA basis functions. The spectral components of each basis occupy a sub-band, which overlaps with neighboring sub-bands. This is conceptually very similar to the filter-bank approaches in speech signal processing. Therefore, the object of the ICABFT is to form independent signal channels, of which frequency contents are also independent.

2.2. Speech Enhancement in ICA basis function domain

As previously mentioned, the speech signal applied to the speech enhancement step is a noisy signal $\mathbf{Y}(m)$ which has undergone pre-emphasis, windowing, and the ICABFT. The output of this step is a noise suppressed speech signal $\mathbf{S}(m)$.

2.2.1. Hypotheses and Density Models

Assuming that the noisy speech observation $\mathbf{Y}(m)$ is a sum of clean speech $\mathbf{S}(m)$ and additive noise $\mathbf{N}(m)$, we consider the statistical model employing two 'global' hypotheses, H_0 and H_1, which indicate speech absence and presence at m-th frame, respectively.

$$\begin{array}{ll} H_0: & \mathbf{Y}(m) = \mathbf{N}(m), \\ H_1: & \mathbf{Y}(m) = \mathbf{S}(m) + \mathbf{N}(m) \end{array} \quad (5)$$

Moreover, since speech absence and presence arise independent component-wise, we further consider the statistical model employing two 'local' hypotheses, $H_{0,k}$ and $H_{1,k}$ for each independent component, which indicate speech absence and presence at k-th basis of the m-th frame, respectively.

$$\begin{array}{ll} H_{0,k}: & Y_k(m) = N_k(m), \\ H_{1,k}: & Y_k(m) = S_k(m) + N_k(m) \end{array} \quad (6)$$

It is also assumed that $Y_k(m)$ and $S_k(m)$ have zero-mean generalized Gaussian densities and $N_k(m)$ has a zero-mean Gaussian density.

$$p(Y_k(m)) = \frac{\nu_Y(k,m) \cdot \eta_Y(k,m)}{2 \cdot \Gamma(1/\nu_Y(k,m))} \cdot exp\{-[\eta_Y(k,m) \cdot |Y_k(m)|]^{\nu_Y(k,m)}\} \quad (7)$$

$$p(S_k(m)) = \frac{\nu_S(k,m) \cdot \eta_S(k,m)}{2 \cdot \Gamma(1/\nu_S(k,m))} \cdot exp\{-[\eta_S(k,m) \cdot |S_k(m)|]^{\nu_S(k,m)}\} \quad (8)$$

$$p(N_k(m)) = \frac{1}{\sqrt{2\pi\sigma_N^2(k,m)}} \cdot exp\left\{-\frac{N_k(m)^2}{2\sigma_N^2(k,m)}\right\}, \quad (9)$$

in which

$$\eta_X(k,m) = \frac{1}{\sqrt{\sigma_X^2(k,m)}} \cdot \left[\frac{\Gamma(3/\nu_X(k,m))}{\Gamma(1/\nu_X(k,m))}\right]^{1/2}, \quad (10)$$

where

$$\nu_X(k,m) = F\left(\frac{\sigma_{|X|}(k,m)}{\sqrt{\sigma_X^2(k,m)}}\right), \quad (11)$$

$$F(\nu) = \frac{\Gamma(2/\nu)}{\sqrt{\Gamma(1/\nu)\cdot\Gamma(3/\nu)}}, \quad (12)$$

and X denotes either Y or S.

The sparse density used in [3] does not fit the real density of the speech very well. As seen in Fig. 3, it fits the real density very well near the origin. However, there are significant deviations for larger values, in which the information about the speech signals reside. With this inaccurate sparse density, it is difficult to detect the speech absence intervals, and in turn, it will cause the noise variance to deviate from the real value. This is why we assumed that $Y_k(m)$ and $S_k(m)$ follow the generalized Gaussian densities.

2.2.2. Statistical Parameters Initialization

Statistical parameters are initialized for a predetermined number of initial frames to collect noisy speech, enhanced speech, and background noise information. These parameters are noisy speech power estimate, noisy speech magnitude estimate, enhanced speech power estimate, enhanced speech magnitude estimate and noise power estimate. For $m = 0$, the parameters are initialized by

$$\begin{array}{rcl} \sigma_Y^2(k,0) & = & Y_k(0)^2, \\ \sigma_{|Y|}(k,0) & = & |Y_k(0)|, \\ \sigma_S^2(k,0) & = & S_k(0)^2, \\ \sigma_{|S|}(k,0) & = & |S_k(0)|, \\ \sigma_N^2(k,0) & = & N_k(0)^2. \end{array} \quad (13)$$

and for $m <$*INIT-FRAMES*, the parameters are updated by

$$\sigma_Y^2(k,m) = \zeta_{Y^2}\sigma_Y^2(k,m-1) + (1-\zeta_{Y^2})Y_k(m)^2, \quad (14)$$

$$\sigma_{|Y|}(k,m) = \zeta_{|Y|}\sigma_{|Y|}(k,m-1) + (1-\zeta_{|Y|})|Y_k(m)|, \quad (15)$$

$$\sigma_S^2(k,m) = \zeta_{S^2}\sigma_S^2(k,m-1) + (1-\zeta_{S^2})S_k(m)^2, \quad (16)$$

$$\sigma_{|S|}(k,m) = \zeta_{|S|}\sigma_{|S|}(k,m-1) + (1-\zeta_{|S|})|S_k(m)|, \quad (17)$$

$$\sigma_N^2(k,m) = \zeta_{N^2}\sigma_N^2(k,m-1)) + (1-\zeta_{N^2})N_k(m)^2, \quad (18)$$

where ζ_{Y^2}, $\zeta_{|Y|}$, ζ_{S^2}, $\zeta_{|S|}$, and ζ_{N^2} are pre-defined constants in $[0,1]$.

Assuming that only noise is present at each k-th basis for the first *INIT-FRAMES* frames, each enhanced speech coefficient $S_k(m)$ is computed by

$$S_k(m) = GAIN_{MIN} \cdot Y_k(m), \quad (19)$$

where $GAIN_{MIN}$ is the minimum gain. The value of this is 0.2238, which corresponds to the one in the IS–127 standard used for North American CDMA digital PCS.

2.2.3. Global Soft Decision

After initialization, the frame index is incremented, and the signal of the corresponding frame (herein the m-th frame) is processed. The noisy speech power estimate $\sigma_Y^2(k,m)$ and the noisy speech magnitude estimate $\sigma_{|Y|}(k,m)$ are smoothed by (14) and (15) in consideration for the inter-frame correlation of the speech signal. Then, each generalized Gaussian exponent $\nu_Y(k,m)$ is computed by (11) and (12) using the method described in [9].

The global SAP, $p(H_0|\mathbf{Y}(m))$ of the m-th frame is computed by

$$\begin{array}{rcl} p(H_0|\mathbf{Y}(m)) & = & \dfrac{p(H_0,\mathbf{Y}(m))}{p(\mathbf{Y}(m))} \\ & = & \dfrac{1}{\prod_{k=1}^{M}\left[1+q_k\Lambda_k(m)\right]}, \end{array} \quad (20)$$

in which q_k is the ratio defined by

$$q_k = \frac{p(H_{1,k})}{p(H_{0,k})}, \quad (21)$$

and $\Lambda_k(m)$ is the likelihood ratio computed for the k-th basis of the m-th frame as

$$\Lambda_k(m) = \frac{p(Y_k(m)|H_{1,k})}{p(Y_k(m)|H_{0,k})}. \quad (22)$$

The computation of the right-hand side of (20) is possible because $Y_k(m)$'s are statistically independent due to the philosophy of the extraction algorithm of the ICA basis functions. Thus, in deriving (20), the following equations were utilized

$$p(H_0,\mathbf{Y}(m)) = \prod_{k=1}^{M}\left[p(Y_k(m)|H_{0,k})p(H_{0,k})\right], \quad (23)$$

and

$$\begin{array}{rcl} p(\mathbf{Y}(m)) & = \prod_{k=1}^{M} & p(Y_k(m)) \\ & = \prod_{k=1}^{M} & [p(Y_k(m)|H_{0,k})p(H_{0,k}) \\ & & +p(Y_k(m)|H_{1,k})p(H_{1,k})]. \end{array} \quad (24)$$

We compare the global SAP with a threshold that can be set by the user. If the global SAP exceeds the threshold, the noise power estimate is updated by (18). If the global SAP does not exceed the threshold, the noise power estimate remains the same.

2.2.4. Speech Parameters Prediction

Regardless of the global SAP, prediction of the speech power estimate, $\sigma_S^2(k,m)$ and the speech magnitude estimate, $\sigma_{|S|}(k,m)$ are performed.

$$\sigma_S^2(k,m) = \zeta_{S^2}^{pred}\sigma_S^2(k,m-1) + (1-\zeta_{S^2}^{pred}) \cdot \frac{Y_k(m)^2}{1+\sigma_N^2(k,m)/\sigma_S^2(k,m-1)} \quad (25)$$

$$\sigma_{|S|}(k,m) = \zeta_{|S|}^{pred}\sigma_{|S|}(k,m-1) + (1-\zeta_{|S|}^{pred}) \cdot \sqrt{\frac{Y_k(m)^2}{1+\sigma_N^2(k,m)/\sigma_S^2(k,m-1)}} \quad (26)$$

This prediction comes from the Wiener filter. In most cases, this is not a crucial step in affecting enhanced speech quality. However, the spectrogram of the enhanced speech with this step included looks sharper.

2.2.5. Sparse Code Shrinkage and Parameters Update

The enhanced speech coefficient $S_k(m)$ of the k-th basis of the m-th frame is computed with the updated and predicted parameters. Although we assumed different density models from the sparse densities used in the sparse code shrinkage technique, the shrinkage functions are adopted as noise suppressors, because the shapes of shrinkage functions of these two different density models are close to each other. Moreover, there is an advantage that the shrinkage functions can be expressed in closed-forms.

There are two models to compute $S_k(m)$ [3]. If

$$\sqrt{\sigma_S^2(k,m)}p(S_k(m)=0) < \frac{1}{\sqrt{2}}, \quad (27)$$

then $S_k(m)$ is obtained by using (28) through (30)

$$S_k(m) = \frac{1}{1+\sigma_N^2(k,m)a} \cdot sign(Y_k(m)) \cdot max(0, |Y_k(m)| - b\sigma_N^2(k,m)), \quad (28)$$

where

$$b = \frac{2p(S_k(m)=0)\sigma_S^2(k,m) - \sigma_{|S|}(k,m)}{\sigma_S^2(k,m) - \sigma_{|S|}(k,m)^2}, \quad (29)$$

$$a = \frac{1}{\sigma_S^2(k,m)}[1-\sigma_{|S|}(k,m)b]. \quad (30)$$

If (27) is not satisfied, then $S_k(m)$ is obtained by using (31) through (35)

$$S_k(m) = sign(Y_k(m)) \cdot max\left(0, \frac{|Y_k(m)|-ad}{2} + \frac{1}{2}\sqrt{(|Y_k(m)|+ad)^2 - 4\sigma_N^2(k,m)(\alpha+3)}\right), \quad (31)$$

where

$$d = \sqrt{\sigma_S^2(k,m)}, \quad (32)$$

$$k = d^2 p(S_k(m)=0)^2, \quad (33)$$

$$\alpha = \frac{2-k+\sqrt{k(k+4)}}{2k-1}, \quad (34)$$

$$a = \sqrt{\alpha(\alpha+1)/2}. \quad (35)$$

In calculating $S_k(m)$ we need to compute $p(S_k(m)=0)$. $S_k(m)$ also has the zero-mean generalized Gaussian density. Thus,

$$p(S_k(m)=0) = \frac{\nu_S(k,m)\cdot\eta_S(k,m)}{2\cdot\Gamma(1/\nu_S(k,m))}. \quad (36)$$

The computation of $\nu_S(k,m)$ may not be necessary for each frame if we already have the values of $\nu_S(k,m)$ from the off-line calculation. However, these values depend on a training database.

If $S_k(m)$, computed from the model selected by (27), is less than $GAIN_{MIN}Y_k(m)$, then $S_k(m)$ should be set to $GAIN_{MIN}Y_k(m)$. This prevents the noise suppressor from over-shrinking.

$$S_k(m) = max(S_k(m), GAIN_{MIN}Y_k(m)) \quad (37)$$

Unless speech enhancement is performed on all of the frames, the parameters are updated for the next frame. The noise power estimate is maintained for the next frame as

$$\sigma_N^2(k,m+1) = \sigma_N^2(k,m), \quad 1 \le k \le M. \quad (38)$$

The speech power estimate $\sigma_S^2(k,m)$ and the speech magnitude estimate $\sigma_{|S|}(k,m)$ are corrected by (16) and (17) using the enhanced speech coefficients.

After the parameters are updated for the next frame, the frame index is incremented to perform speech enhancement for all the frames.

2.3. Post-Processing

In post-processing, the enhanced signal $\mathbf{S}(m)$ is converted back into a signal of the time domain by an Inverse ICABFT given by (39), then de-emphasized.

$$\tilde{s}_m = \mathbf{A}_{oo}\mathbf{S}(m) \quad (39)$$

Prior to the de-emphasis, the signal obtained through the Inverse ICABFT is subjected to an overlap-and-add operation.

$$\hat{s}_m(n) = \begin{cases} \tilde{s}_m(n) + \tilde{s}_{m-1}(L+n), & 0 \le n < D \\ \tilde{s}_m(n), & D \le n < L \end{cases} \quad (40)$$

Then, the de-emphasis is performed to compute the speech signal $s_m(n)$ of the m-th frame in the time domain.

$$s_m(n) = \hat{s}_m(n) + \zeta \cdot s_m(n-1), \quad 0 \le n < L \quad (41)$$

Note that the s_m's are of length, L and non-overlapping.

3. EXPERIMENTAL RESULTS AND DISCUSSION

To verify the effect of the proposed speech enhancement method using sparse code shrinkage and global soft decision, we performed an experiment on the ITU Korean database. This database consists of 96 phonetically balanced Korean sentence pairs from four male and four female speakers. These 16 bit/16 kHz sampled clean speech data were downsampled to produce 16 bit/8 kHz sampled data.

72 sentence pairs uttered by three male and three female speakers were used for learning the ICA basis function matrix, **A**. In this experiment the ICA basis functions were extracted directly by the algorithm described in [8]. The speech signals were 16 bit/8 kHz sampled monaural data. The size of overlapping, D, frame shift, L and ICABFT, M were 16, 48, and 64, respectively. These correspond to 2 msec. of overlapping, 6 msec. of frame shift(or non-overlapping frame size at the output), and 8 msec. of ICABFT(or overlapping frame size at the input). The parameter, ζ used in pre-emphasis and de-emphasis was 0.95. The statistical learning parameters, ζ_{Y^2}, $\zeta_{|Y|}$, ζ_{S^2}, $\zeta_{|S|}$, $\zeta_{S^2}^{pred}$, $\zeta_{|S|}^{pred}$, and ζ_{N^2} were set to 0.5, 0.5, 0.5, 0.5, 0.8, 0.8, and 0.98, respectively. The number of initial frames, *INIT-FRAMES* was 10. The hypotheses ratio, q_k was 10^{-4} for all the independent components. The threshold value which determines whether the current frame is speech-absent was set to 0.95. Speech parameters, $\nu_S(k,m)$ are estimated frame by frame.

The remaining 24 sentence pairs from a male and a female speaker were prepared for testing. The signal-to-noise ratio(SNR) of each of the 24 sentence pairs was varied using three types of noise, white Gaussian, car, and babble noise on the basis of NOISEX-92 database. According to the SNR, noises were simply added sample by sample after adjusting the signal levels by the method described in the ITU-T recommendation P.830.

Figure 4 shows an experimental result of the proposed speech enhancement system for a test speech along with the clean and noisy speech. As expected, the enhanced speech reduced noise significantly and effectively in real-time. The quality of the enhanced speech was almost compatible with the one by the method in [2], except that especially in speech presence intervals, there were some minuscule artifacts. When the parameters were not properly estimated, this artifact became a harsh sound. The artifacts were thought to be caused by a mismatch between the statistical density models used in parameter estimations and shrinkage functions.

For speech quality evaluation, segmental SNR was considered as an objective criterion.

$$SNR(m) = 10log_{10}\frac{\sum_{i=0}^{L-1} s^2(mL+i)}{\sum_{i=0}^{L-1}\left[s(mL+i)-s_m(i)\right]^2} \quad (42)$$

This is believed to be a more adequate measure for speech quality evaluation, because it considers the difference between clean speech and the output of the speech enhancement system as the noise signal. Non-overlapping frames of 128 samples were used. Table 1 shows the objective test results for two different input SNRs and for three different noise types. For noisy and enhanced speech, the mean value of each segmental SNR was calculated for all the frames of all the test sentences. To show the noise suppression effect, the difference between average segmental SNRs of noisy and enhanced speech was also indicated in boldface figures. These figures represent the amount of noise actually suppressed on the average. In spite of the assumption that the noise density is Gaussian, noise reduction for colored noises (car and babble) were very effective.

Table 1. Averages of segmental SNRs.

SNR	10 dB		20 dB	
segmental SNR	noisy	enhanced	noisy	enhanced
	enhanced - noisy		**enhanced - noisy**	
white	-13.64	-6.00	-3.62	1.53
	7.64		**5.15**	
car	-13.42	-7.89	-3.39	0.73
	5.53		**4.12**	
babble	-13.41	-7.99	-3.38	0.62
	5.42		**4.00**	

4. REFERENCES

[1] Nam Soo Kim and Joon-Hyuk Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Letters*, vol. 7, no. 5, pp. 108–110, 2000.

[2] Vladimir I. Shin and Doh-Suk Kim, "Speech enhancement using improved global soft decision," in *Proc. Europ. Conf. on Speech Communication and Technology*, 2001.

[3] Aapo Hyvärinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Computation*, vol. 11, no. 7, pp. 1739–1768, 1999.

[4] Jong-Hwan Lee, Ho-Young Jung, Te-Won Lee, and Soo-Young Lee, "Speech coding and noise reduction using ica-based speech features," in *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation*, 2000, pp. 417–421.

[5] I. Potamitis, N. Fakotakis, and G. Kokkinakis, "Speech enhancement using the sparse code shrinkage technique," in *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, 2001.

[6] Aapo Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[7] Anthony J. Bell and Terrence J. Sejnowski, "An information-maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[8] Michael S. Lewicki and Terrence J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[9] Stephane G. Mallat, "Multifrequency channel decompositions of images and wavelet models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 12, pp. 2091–2110, 1989.

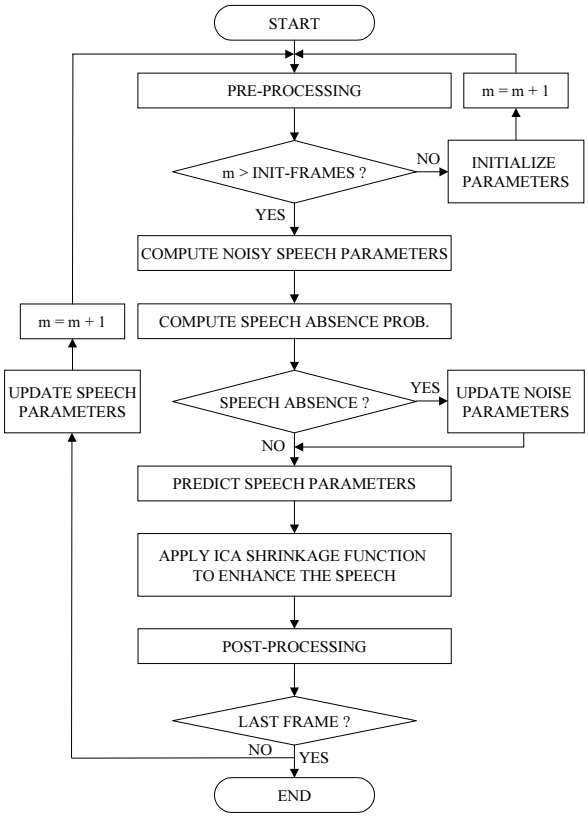

Fig. 1. A flowchart illustrating the speech enhancement method.

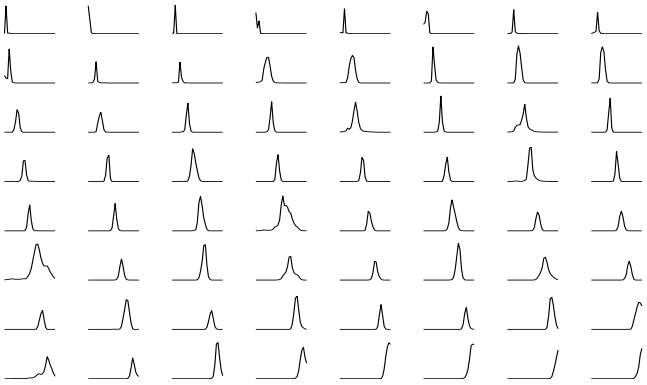

Fig. 2. Power spectral densities (0 to 4kHz) of the frequency-ordered and orthogonalized ICA basis function matrix, $\mathbf{A}_{oo}$.

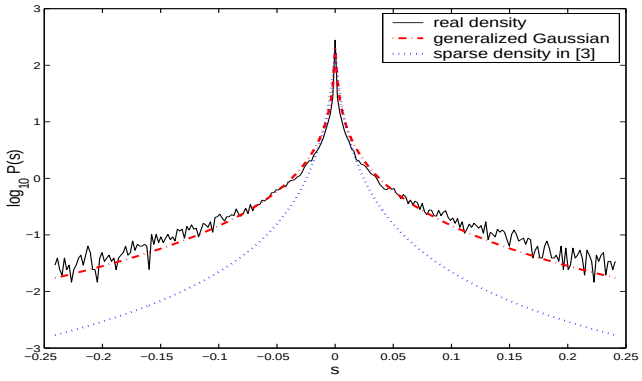

Fig. 3. Comparison of two estimated densities, generalized Gaussian density and sparse density used in [3]. Note log scale on y-axis.

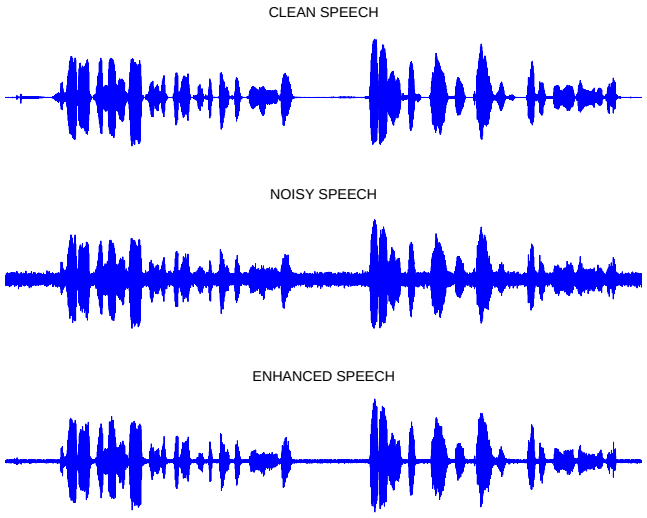

Fig. 4. An example of speech enhancement for a pair of test noisy sentences. A white Gaussian noise was used. SNR was 10dB.