BAYESIAN SINGLE CHANNEL SPEECH ENHANCEMENT EXPLOITING SPARSENESS IN THE ICA DOMAIN

Liang Hong, Justinian Rosca, Radu Balan

Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540

ABSTRACT

We propose a Bayesian single channel speech enhancement algorithm to exploit speech sparseness in the independent component analysis (ICA) domain. While recent literature considers the idea of denoising in the ICA domain, it relies on the unrealistic assumption of uncorelatedness of noise components in the ICA domain. Here we drop this limiting assumption and address the general case. The approach consists of two elements: (1) a *maximum a posteriori* (MAP) estimator for speech coefficients in the ICA domain, further used to estimate enhanced speech in the time domain, and (2) ICA domain transformation of data, learned from speech training data and then used in step (1). An implementation of the method shows considerable noise reduction capability in denoising speech keywords such as car navigation commands. Evaluation is based on objective measures of signal-to-noise ratio and distortion in enhanced signals versus the real-world speech and noise mixtures from car, street, office, industrial environments.

1. INTRODUCTION

Speech data admits sparse representations. Recent literature exploits this assumption informally by considering that speech features have Laplacian priors, or furthermore by using independent component analysis (ICA) to derive features for speech processing [1, 2, 3]. How can the sparseness assumption be exploited in the design of algorithms for speech enhacement?

Speech enhancement (SE) aims at suppressing noise and improving the perceptual quality and intelligibility of speech in speech-based human-machine interfaces [4]. Due to the random nature of noise and the inherent complexities of speech, the problem of reconstructing the clean voice from noisy signal has been challenging researchers over the past three decades. Todate, single channel techniques for noise reduction are widely used due to their simplicity and ease of implementation but offer little perspective for further progress. A variety of theoretical and relatively effective techniques have attempted to capitalize on specific characteristics or constraints with varying degrees of success. Recent literature addresses these approaches [4], such as spectral subtraction [5], model based SE [6], noise masking [7] etc.

On the other side, ICA is a relatively new technique proposed to solve the problem of blind source separation [8] but also recently applied to speech enhancement [9, 10]. ICA is a statistical technique for revealing hidden factors underlying sets of signals. In this model, data variables are assumed to be mixtures of some unknown latent variables with unknown mixing. The latent variables are assumed nongaussian and jointly independent. Comparing with some commonly used transformations, such as discrete Fourier transform, discrete cosine transform, and wavelet transform, ICA is a data-driven transformation adapted to the structure of clean speech data. In ICA domain, the speech signal is processed uniformly in both amplitude and phase. SE algorithms using ICA technique generally envolve a thresholding operation, therefore the technique has the potential of limited musical noise.

Here, we develop a Bayesian single channel SE estimator in the ICA domain. We derive the *maximum a posteriori* estimator in the general case of possibly correlated ICA domain noise components. In contrast, previous work [9, 10] unrealistically assumes uncorrelated noise components in the ICA domain. Next we elaborate the algorithm for Bayesian single channel SE in the ICA domain. Section 3 demonstrates the noise reduction capability of the proposed algorithm in a selection of real-world noisy data. Finally, Section 4 concludes the work and highlights future possible extensions.

2. BAYESIAN SPEECH ENHANCEMENT APPROACH

2.1 ICA Model for Speech Enhancement

Consider a time domain, additive-noise corrupted speech signal received at the microphone, $x(m) = s(m) + n(m)$, where m is the discrete time index, $s(m)$ is the clean speech signal and $n(m)$ is an additive noise.

The first step in ICA-based algorithms is to segment the received signals $x(m)$ with a time-domain window and form segments as columns of a matrix:

$$\mathbf{X} = \mathbf{S} + \mathbf{N}, \tag{1}$$

where matrices $\mathbf{X}$, $\mathbf{S}$ and $\mathbf{N}$ have size $M \times K$, M is the speech frame size (in samples) and K is the number of frames.

Since speech signals are characterized by higher order statistics [11], without loss of generality, we may assume that clean speech signal is the linear mixture of some independent components. ICA transforms a set of observed segments $\mathbf{s} = [s_1, s_2, \cdots, s_M]^T$, each representing a column of the matrix $\mathbf{S}$ introduced above, into a new representation $\boldsymbol{\varsigma} = [\varsigma_1, \varsigma_2, \cdots, \varsigma_M]^T$

$$\boldsymbol{\varsigma} = \mathbf{W} \cdot \mathbf{s} \tag{2}$$

where the components $\varsigma_i, 1 \le i \le M$ of $\boldsymbol{\varsigma}$ are jointly statistically independent, and $\mathbf{W}$ is an $M \times M$ invertible matrix, generally called demixing matrix. Finding the demixing matrix is the subject of the next section. By applying $\mathbf{W}$ from the left side to a column of each matrix in (1), we have:

$$\boldsymbol{\gamma} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{s} + \mathbf{W}\mathbf{n} = \boldsymbol{\varsigma} + \boldsymbol{\nu} \tag{3}$$

where $\mathbf{x}$ and $\mathbf{n}$ are columns of matrices $\mathbf{X}$ and $\mathbf{N}$, respectively, $\boldsymbol{\gamma}$ and $\boldsymbol{\nu}$ are the corrupted speech and noise segments in the ICA domain corresponding to $\mathbf{x}$ and $\mathbf{n}$. All the above variables, $\mathbf{x}$, $\mathbf{n}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\nu}$, are $M \times 1$ vectors.

Let $\hat{\boldsymbol{\varsigma}}$ denote the estimate of $\boldsymbol{\varsigma}$ in ICA domain when noise is present. By applying the inverse transformation to $\hat{\boldsymbol{\varsigma}}$, we obtain the enhanced speech segment vector, $\mathbf{z}$, as

$$\mathbf{z} = \mathbf{W}^{-1} \cdot \hat{\boldsymbol{\varsigma}} \tag{4}$$

Our task is therefore to estimate $\boldsymbol{\varsigma}$ given $\boldsymbol{\gamma}$. A maximum *a posteriori* algorithm is developed next.

2.2 MAP Estimator in ICA Domain

If noise $\mathbf{n}$ was Gaussian, $\boldsymbol{\nu} = \mathbf{W}\mathbf{n}$ is also Gaussian. On the other hand, if noise $\mathbf{n}$ is not Gaussian, $\boldsymbol{\nu}$ has a distribution closer to a Gaussian than the distribution of each component of $\mathbf{n}$, according to the Central Limit Theorem. Therefore, we assume that $\boldsymbol{\nu}$ has a Gaussian distribution.

The posterior p.d.f. of $\boldsymbol{\varsigma}$ can be expressed via the Bayes rule:

$$p(\boldsymbol{\varsigma}|\boldsymbol{\gamma}) = \frac{p(\boldsymbol{\gamma}|\boldsymbol{\varsigma}) \cdot p(\boldsymbol{\varsigma})}{p(\boldsymbol{\gamma})} \tag{5}$$

In this study, we assume that each component of the ICA transformed speech data $\boldsymbol{\varsigma}$ has Laplacian distribution, that is, $p(\varsigma_i) = \frac{1}{2\lambda_i} \exp\left(-\frac{|\varsigma_i|}{\lambda_i}\right)$, where $\lambda_i, 1 \le i \le M$ are positive constants [2]. Taking into account that the components of $\boldsymbol{\varsigma}$ are independent, we obtain the prior p.d.f. of $\boldsymbol{\varsigma}$ in equation 5

$$p(\boldsymbol{\varsigma}) = \prod_{i=1}^{M} \frac{1}{2\lambda_i} \exp\left(-\frac{|\varsigma_i|}{\lambda_i}\right) \tag{6}$$

Furthermore, since $\boldsymbol{\nu}$ has Gaussian distribution,

$$p(\boldsymbol{\gamma}|\boldsymbol{\varsigma}) = (2\pi)^{-\frac{M}{2}} \cdot det^{-\frac{1}{2}}(\mathbf{R}_{\boldsymbol{\nu}}) \cdot \exp\left[-\frac{1}{2}(\boldsymbol{\gamma}-\boldsymbol{\varsigma})^T \mathbf{R}_{\boldsymbol{\nu}}^{-1}(\boldsymbol{\gamma}-\boldsymbol{\varsigma})\right], \tag{7}$$

where the notation $det(\mathbf{A})$ is used for the determinant of a matrix $\mathbf{A}$, $\mathbf{R}_{\boldsymbol{\nu}}$ is the covariance matrix of noise $\boldsymbol{\nu}$ in ICA domain:

$$\mathbf{R}_{\boldsymbol{\nu}} = \mathbf{W}\mathbf{R}_{\mathbf{n}}\mathbf{W}^T, \tag{8}$$

where $\mathbf{R}_{\mathbf{n}}$ is the covariance matrix of the time-domain noise $\mathbf{n}$.

For a uniform cost function, the MAP estimate of $\boldsymbol{\varsigma}$ given $\boldsymbol{\gamma}$ is the value of $\boldsymbol{\varsigma}$ that maximizes $p(\boldsymbol{\varsigma}|\boldsymbol{\gamma})$:

$$\hat{\boldsymbol{\varsigma}} = \underset{\boldsymbol{\varsigma}}{argmax}[p(\boldsymbol{\varsigma}|\boldsymbol{\gamma})] = \underset{\boldsymbol{\varsigma}}{argmax}[p(\boldsymbol{\gamma}|\boldsymbol{\varsigma}) \cdot p(\boldsymbol{\varsigma})]. \tag{9}$$

By using the probability density functions (6) and (7), we have

$$p(\boldsymbol{\gamma}|\boldsymbol{\varsigma}) \cdot p(\boldsymbol{\varsigma}) = C \cdot \exp\left[-\frac{1}{2}(\boldsymbol{\gamma}-\boldsymbol{\varsigma})^T \mathbf{R}_{\boldsymbol{\nu}}^{-1}(\boldsymbol{\gamma}-\boldsymbol{\varsigma}) - \sum_{i=1}^{M} \frac{|\varsigma_i|}{\lambda_i}\right], \tag{10}$$

where $C = \left(\prod_{i=1}^{M} \frac{1}{2\lambda_i}\right) \cdot (2\pi)^{-\frac{M}{2}} \cdot det^{-\frac{1}{2}}(\mathbf{R}_{\boldsymbol{\nu}})$ is a constant.

The maximization problem in (9) is equivalent to the following minimization problem:

$$\hat{\boldsymbol{\varsigma}} = \underset{\boldsymbol{\varsigma}}{argmin}\left[\frac{1}{2}(\boldsymbol{\gamma}-\boldsymbol{\varsigma})^T \mathbf{R}_{\boldsymbol{\nu}}^{-1}(\boldsymbol{\gamma}-\boldsymbol{\varsigma}) + \sum_{i=1}^{M} \frac{|\varsigma_i|}{\lambda_i}\right] \tag{11}$$

The derivative of (11) with respect to $\boldsymbol{\varsigma}$ leads to

$$\mathbf{0} = \mathbf{R}_{\boldsymbol{\nu}}^{-1}(\boldsymbol{\gamma}-\boldsymbol{\varsigma}) - \underset{i=1,\cdots M}{diag}\left(\frac{1}{\lambda_i}\right) \cdot sign(\boldsymbol{\varsigma}) \tag{12}$$

where $sign(\boldsymbol{\varsigma})$ is a $M \times 1$ vector and $diag(\frac{1}{\lambda_i})$ is an $M \times M$ diagonal matrix with the ith diagonal element $\frac{1}{\lambda_i}$.

2.3 Approximate Estimator

There is no closed-form solution for (12). When M is larger than 20, exhaustive numerical search requires 2^M sign combinations, which is impractical in computational complexity. Iterative search approaches, such as Jacobi or Gauss-Seidel types of iteration (see chapter 10.1 in [12]), are not practical due to both computational complexity and convergence issues. Daubechies *et al* proposed an iterative thresholding algorithm for the minimization problem (11) [13]. In this paper, we follow a different approach by applying a computationally efficient approximation to the MAP estimate of $\boldsymbol{\varsigma}$.

When the input signal-to-noise ratio is moderate to high, the noise $\boldsymbol{\nu}$ has little effect on $\boldsymbol{\gamma}$, therefore,

$$|\varsigma_i| = |\gamma_i - \nu_i| \approx |\gamma_i|, \quad 1 \le i \le M. \tag{13}$$

where γ_i and ν_i are the ith element of $\boldsymbol{\gamma}$ and $\boldsymbol{\nu}$, respectively.

Substituting (13) into $sign(\boldsymbol{\varsigma}) = \underset{1,\cdots,M}{diag}\left(\frac{1}{|\varsigma_i|}\right)\boldsymbol{\varsigma}$ and the result into (12), we have

$$\mathbf{0} = \mathbf{R}_{\boldsymbol{\nu}}^{-1}(\boldsymbol{\gamma}-\boldsymbol{\varsigma}) - \underset{1,\cdots,M}{diag}\left(\frac{1}{\lambda_i}\right) \cdot \underset{1,\cdots,M}{diag}\left(\frac{1}{|\gamma_i|}\right) \cdot \boldsymbol{\varsigma}. \tag{14}$$

Now it is very easy to find the maximum *a posteriori* estimate of $\boldsymbol{\varsigma}$ as the root of (14):

$$\hat{\boldsymbol{\varsigma}} = \left[\mathbf{R}_{\boldsymbol{\nu}} \cdot \underset{i=1,\cdots,M}{diag}\left(\frac{1}{\lambda_i \cdot |\gamma_i|}\right) + \mathbf{I}\right]^{-1} \cdot \boldsymbol{\gamma}. \tag{15}$$

where $\mathbf{I}$ is an $M \times M$ identity matrix.

Once we obtain the estimate of $\boldsymbol{\varsigma}$, substituting it into equation (4) gives the time-domain enhanced speech signals. Finally, the enhanced waveform is obtained by reshaping the enhanced speech signals from matrix to vector form. What remains to be obtained are the demixing matrix $\mathbf{W}$ for ICA transform, parameters $\lambda_i, i = 1, \cdots, M$ required in the probability density function of the ICA transformed speech data $\boldsymbol{\varsigma}$, and noise covariance matrix $\mathbf{R}_{\mathbf{n}}$ required in obtaining $\mathbf{R}_{\boldsymbol{\nu}}$. The first two will be learned from a large ensemble of clean speech training frames, whereas $\mathbf{R}_{\mathbf{n}}$ can be estimated from noisy data in the absence of speech by using a voice activity detector (VAD). Figure 1 presents the block diagram of the resulting single channel SE approach.

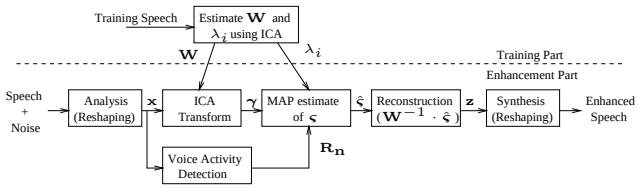

Figure 1: Block diagram of the ICA-based Bayesian single channel speech enhancement system.

2.4 Learning W and Estimating λ_i

Several practical methods for estimating the ICA basis are presently used [14]. In our study, a robust and efficient principle based on maxima of nongaussianity is employed to estimate the ICA model. Nongaussianity is measured by negentropy, a robust and in some sense optimal estimator the nongaussianity. We apply a new nonquadratic function to approximately measure the nongaussianity, defined by

$$G(\xi) = -(|\xi| + 1)e^{-|\xi|} + 1 \tag{16}$$

This is more robust than those defined in [14] and yields less dependece between the ICA components.

After obtaining the demixing matrix $\mathbf{W}$, we estimate parameters λ_i required in equation (6) from speech training data. They are approximations of the densities of independent components transformed from the test speech data. Several experiments demonstrate that these approximations model well the ICA transformed speech [15].

Parameters λ_i are estimated using the maximum likelihood (ML) technique. For K observations of training speech frames, the likelihood function of the i-th independent component is

$$p(\xi_i|\lambda_i) = \prod_{k=1}^{K} \frac{1}{2\lambda_i} \exp\left\{-\frac{|\xi_i(k)|}{\lambda_i}\right\}, \qquad (17)$$

where k is the index of observation, $\xi_i = \mathbf{w}_i^T \mathbf{s}^{(train)}$ is one of the independent components in vector $\boldsymbol{\xi} = \mathbf{W}\mathbf{s}^{(train)}$ with demixing matrix $\mathbf{W}$ and reshaped training speech matrix $\mathbf{s}^{(train)}$.

Taking the derivative of the natural logarithm of (17) with respect to λ_i leads to

$$\frac{\partial \ln p(\xi_i|\lambda_i)}{\partial \lambda_i} = \frac{1}{\lambda_i^2}\left(\sum_{k=1}^{K} |\xi_i(k)| - K\lambda_i\right). \qquad (18)$$

Setting the result of (18) to zero forms the likelihood equation. The ML estimate of λ_i is then:

$$\lambda_i = \frac{1}{K}\sum_{k=1}^{K} |\xi_i(k)|. \qquad (19)$$

3. EXPERIMENTAL RESULTS

We analyze the noise reduction capability of the proposed algorithm on real-world noisy data. The scenario of interest is that of speech enhancement with respect to given speech "keywords," (e.g. navigation commands in a car) independent of the actual speaker considered for testing.

We use eight different noise types: car, cafeteria, office, industrial, radio, street, tv and vacuum cleaner. Each noise type is superimposed with speech and scaled so that the global input SNR of the corrupted waveform ranges from -5 dB to 20 dB. For both training and testing, speech data from the TIMIT database or mixture of speech and noise data sampled at 16 kHz is windowed with a Hamming window in frames of size 50 samples (3.1msec) with 50% overlap between successive segments.

For learning the ICA model we use speech sentences from four different speakers. The demixing matrices $\mathbf{W}$ for each speaker are learned from two sentences with total length of about six seconds, by means of the FastICA algorithm [16], augmented with the nongaussianity function given by equation (16). The stopping criterion value of the iterative FastICA code is 0.0001, while the initial point for $\mathbf{W}$ is identity.

For each noise type, eight experiments are carried out (see Table 1) where clean speech representing the keywords is used to train for W and noisy speech from a different speaker is used to test the denoising capability of the system. 'F1', 'F2', 'M1' and 'M2' respresent two female and male speakers, respectively. 'SA1' and 'SA2' represent two different TIMIT sentences with the desired keywords. Therefore, in all experiments the training data is uncorrelated to the testing data.

The noise correlation $\mathbf{R_n}$ required in (8) for computing $\mathbf{R}_{\boldsymbol{\nu}}$ in (15) is obtained from the covariance of reshaped noise matrix $\mathbf{N}$, which is obtained using an AMR VAD component [17].

Test Cases	1	2	3	4
Train Data	F1-SA1 + F1-SA2			
Test Data	F2-SA1	F2-SA2	M1-SA1	M1-SA1
Test Cases	5	6	7	8
Train Data	M1-SA1 + M1-SA2			
Test Data	F1-SA1	F1-SA1	M2-SA1	M2-SA1

Table 1: Train and test data used in experiments.

Figures 2-3 present three objective speech quality criteria (global SNR, segmental SNR, and Itakura distance [4]) used to evaluate the performance of the proposed algorithm. Each row in the two figures represents results for one criterion. Figure 2 plots the measures for the enhanced speech. Figure 3 plots the difference between the enhanced speech and the input measures.

The first row presents the average global SNR. The algorithm yields high improvement in car, street and vacuum cleaner noise environments. For example, at -5 dB input global SNR, the SNR improvement is about 11.4 dB, 8 dB and 4.4 dB for these three noise types. The enhancement is about 1 dB on speech corrupted by cafeteria, industrial and radio noises. The algorithm provides little improvement in office and tv environments.

The second row illustrates segmental SNR results (see [4], equation 9.7, with SNR thresholds of -20 dB and +20 dB for the lower and upper bounds). The plot confirms our previous observation. Since segmental SNR is better correlated with speech intelligibility than the global SNR, we order the performance of the algorithm for different noise types based on this criterion. From highest improvement to lowest improvement, the order is car, street, vacuum cleaner, cafeteria, radio, industrial, tv, and office.

The third row shows the Itakura distance measure ([4], equation 5.191). This is a distortion measure, with a low value being better. The algorithm shows a reduction in distortion, naturally smaller for higher input SNR. The method reduces distortion by more than 50% in most environments.

Overall, the algorithm is promising in reducing real-world noise although its capability varies for different noises. Among the noise types studied, the algorithm provides the best overall performance for reducing car, street and home noises. With respect to computational load, one of the most time-consuming tasks of the algorithm is the learning of the demixing matrix $\mathbf{W}$, which can be computed offline only once.

4. CONCLUSIONS

This paper addresses the use of the speech sparseness assumption in the design of algorithms for speech enhacement. We propose a Bayesian single channel SE algorithm in the ICA domain. Speech frames are decoupled into independent components by a demixing matrix obtained by applying ICA to a large ensemble of clean speech training frames. The transformation facilitates the application of a maximum *a posteriori* criterion, whose solution can rely on sparseness assumptions. While recent literature on ICA SE relies on the unrealistic assumption of uncorelatedness of noise components in the ICA domain, we address the general case. Simulation results show that the proposed enhancement approach is able to reduce several types of real-world noises significantly with respect to objective quality measure criteria.

Present work is focused on a comparison with state-of-the-art mono noise reduction approaches, and ICA SE when training and testing are done on speech from same person with no limits on the vocabulary. For future work we suggest the optimization of parameters (criterion 10) using alternate

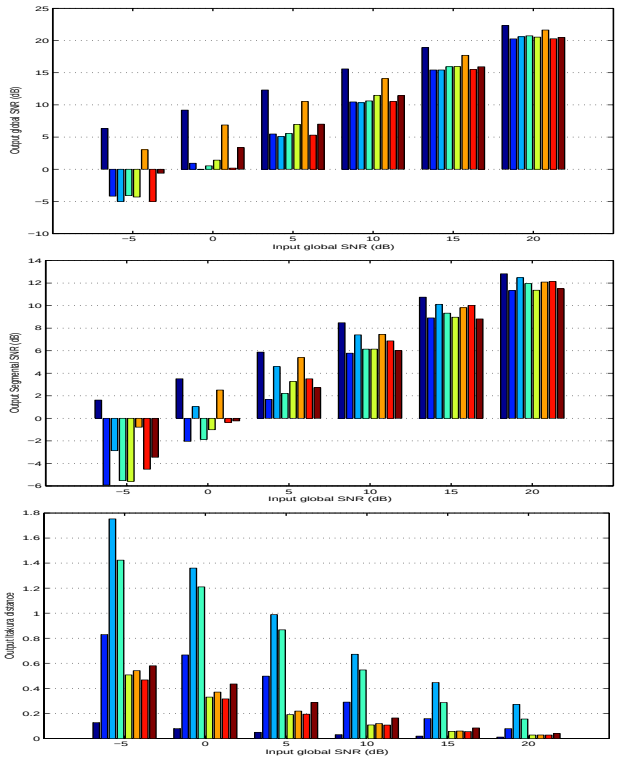

Figure 2: Objective quality measures: (1) Avg. SNR; (2) SegSNR; (3) Itakura; Grey bars represent various types of noises. X axis: -5, 0, 5, 10, 15, 20 dB input SNR.

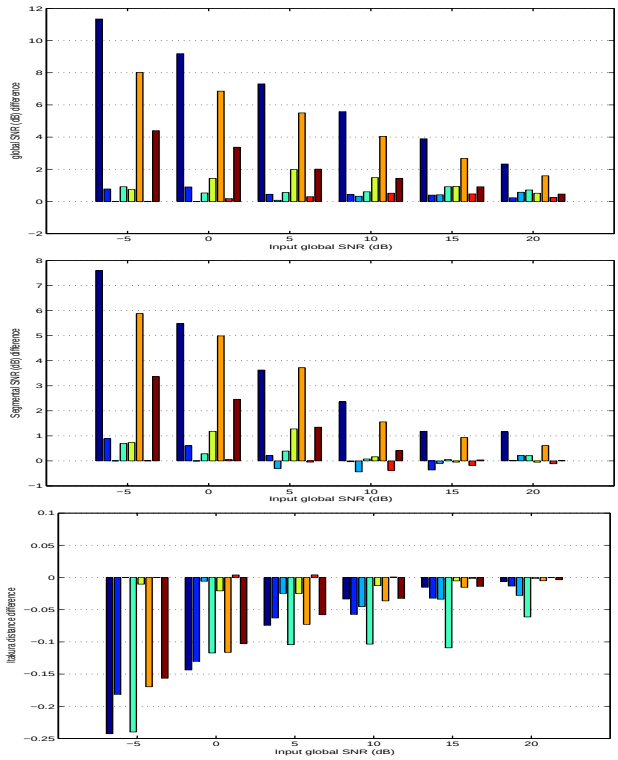

Figure 3: Objective quality measure of enhanced speech minus objective measure of noisy speech: (1) Avg. SNR; (2) SegSNR; (3) Itakura.

algorithms such as the one in [13], and online implementations where $\mathbf{R}_{\boldsymbol{\nu}}$ is updated online.

REFERENCES

[1] Michael Lewicki and Terrence Sejnowski, "Learning overcomplete representations," *Neural Computation*, , no. 12, pp. 337–365, 2000.

[2] J. Lee, H. Jung, T. Lee, and S. Lee, "Speech coding and noise reduction using ica-based speech features," in *International Workshop on Independent Component Analysis*, June 2000, pp. 417–422.

[3] J. Rosca and A. Kofmehl, "Cepstral-like ica representations for text-independent speaker recognition," in *Proc. 4th Int. Conf. on ICA and BSS (ICA2003), Nara, Japan*, April 2003.

[4] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*, IEEE Press, New York, NY, 2000.

[5] K. Wu and P. Chen, "Efficient speech enhancement using spectral subtraction for car hands-free application," in *International Conference on Consumer Electronics*, 2001, vol. 2, pp. 220–221.

[6] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "Hmm-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 5, pp. 445–455, Sept. 1998.

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

[8] Pierre Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[9] I. Potamitis, N. Fakotakis, and G. Kokkinakis, "Speech enhancement using the sparse code shrinkage technique," in *IEEE ICASSP*, May 2001, pp. 621–624.

[10] J. Lee, H. Jung, T. Lee, and S. Lee, "Speech feature extraction using independent component analysis," in *IEEE ICASSP*, June 2000, vol. 3, pp. 1631–1634.

[11] A.J. Bell and T.J. Sejnowski, "Learning the higher order structure of a natural sound," *Network: Computation in Neural Systems*, vol. 7, 1996.

[12] G.H. Golub and C.F.van Loan, *Matrix Computations*, The John Hopkins University Press, 1989.

[13] I. Daubechies, M. Defrise, and C. DeMol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," June 2003, vol. arXiv:math.FA/0307152 at http://arXiv.org/.

[14] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis*, Wiley, New York, NY, 2001.

[15] L. Hong, J. Rosca, and R. Balan, "Independent component analysis based single channel speech enhancement," in *Proceedings of ISSPIT 2003, Darmstadt, Germany*, December 14-17 2003.

[16] $***$, *The FastICA package for MATLAB*, Neural Networks Research Center, Helsinki University of Technology, http://www.cis.hut.fi/projects/ica/ fastica/.

[17] 26.094 TS, *Voice Activity Detector (VAD) for Adaptive Multi Rate (AMR) speech traffic channels*, 3GPP, June 1999.