

INDEPENDENT COMPONENT ANALYSIS BASED SINGLE CHANNEL SPEECH ENHANCEMENT USING WIENER FILTER

Liang Hong, Justinian Rosca, Radu Balan

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

ABSTRACT

Hands-free mobile phones and voice navigation systems are used more and more for safety and convenience when one drives. However, the car environment is noisy. Oftentimes the noise degrades substantially the intelligibility of speech. In this paper, we propose a single channel algorithm that reduce the car noise significantly. The approach employs the Wiener filter and the independent component analysis (ICA) technique. First, ICA is applied to a large ensemble of clean speech training frames to reveal their underlying statistically independent basis. The distribution of the ICA transformed data is also estimated in the training part. It is required for computing the covariance matrix of the ICA transformed speech data used in the Wiener filter. Then a Wiener filter is applied to estimate the clean speech from the received noisy speech. The Wiener filter minimizes the mean-square error between the estimated signal and the clean speech signal in ICA domain. Finally, an inverse transformation from ICA domain back to time domain reconstructs the enhanced signal. Extensive experiments show considerable noise reduction capability of the proposed algorithm. The evaluation is performed with respect to four objective quality measure criteria.

1. INTRODUCTION

Hands-free mobile phones and voice navigation systems are used more and more for safety and convenience when one drives. However, the car environment is noisy due to the presence of interfering sounds such as car engine and road noise, music, and other voices. Oftentimes the noise degrades substantially the intelligibility of speech and severely affect the ability of the driver to understand what the speaker is saying, or diminish the ability of the voice navigation system to process driver's commands.

Over the past three decades, a variety of speech enhancement techniques have been used to suppress such noises and improve the perceptual quality and intelligibility of speech. Boll proposed spectral subtraction [1]. Virag utilized the masking properties of the human auditory system for noise reduction [2]. Knecht *et al* tackled the problem by artificial neural networks [3]. Dahl *et al* introduce a digital self-calibrating microphone array to suppress the car noise [4]. Based on the number of microphones, the first two algorithms can be categorized into single channel systems and the other two are multiple channel systems. The multiple channel systems have greater potential for noise reduction by using spatial information. However, a large number of microphones are needed to achieve an acceptable performance. Generally, this is not practical in terms of spatial placement and the total cost of the whole system.

Taking into account algorithm simplicity and ease of implementation, we propose a single channel technique that applies the

Wiener filtering in independent component analysis (ICA) domain for car noise reduction. The Wiener filter is an optimal filter that minimizes the mean square error between the desired signal and the estimated signal. ICA is a data-driven transformation adapted to the structure of clean speech data. The properties of the two techniques will yield higher noise suppression capability and lower distortion by combining them. The algorithm can be divided into training phase and noise removing phase. The ICA model and the prior knowledge of speech required in Wiener filtering can be computed off-line only once during a training phase. No subsequent speech signal dependent parameter estimation is performed in the noise removing phase, therefore reducing the computation complexity of the algorithm.

The organization of the paper is the following: Section 2 elaborates in detail the single channel speech enhancement algorithm with Wiener filtering in the ICA domain. Then Section 3 demonstrates the noise reduction capability of the proposed algorithm through computer simulation. Finally, a conclusion is reached in Section 4.

2. SPEECH ENHANCEMENT ALGORITHMS

Let the received time domain additive noise corrupted speech signal be $x(m) = s(m) + n(m)$, where m is the discrete time index, $s(m)$ is a clean speech signal and $n(m)$ is the additive noise. To use the ICA technique, we first segment the received signal $x(m)$ with time-domain window and form the segments as the columns of matrices, that is

$$\mathbf{X} = \mathbf{S} + \mathbf{N}. \quad (1)$$

where the matrices \mathbf{X} , \mathbf{S} and \mathbf{N} are of size $M \times K$, M is the speech frame size in samples and K is the number of frames. The rest of the received signal that is shorter than a segment size is truncated during the reshaping.

The speech has specific higher order statistical characteristics [5]. 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 $\mathbf{s} = [s_1, s_2, \dots, s_M]^T$, that forms a column of matrix \mathbf{S} , into a new representation $\boldsymbol{\varsigma} = [\varsigma_1, \varsigma_2, \dots, \varsigma_M]^T$, where the components ς_i , $1 \leq i \leq M$ of $\boldsymbol{\varsigma}$ are jointly statistically independent, that is

$$\boldsymbol{\varsigma} = \mathbf{W} \cdot \mathbf{s}. \quad (2)$$

with \mathbf{W} a $M \times M$ invertible matrix called unmixing matrix. The approach to find the unmixing matrix will be introduced in the next section.

By applying \mathbf{W} from left side to a column of each matrix in (1), we have:

$$\boldsymbol{\gamma} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{s} + \mathbf{W}\mathbf{n} = \boldsymbol{\varsigma} + \boldsymbol{\nu}. \quad (3)$$

where \mathbf{x} and \mathbf{n} are one of the columns of matrix \mathbf{X} and \mathbf{N} , respectively, γ and ν are the corrupted speech segment and noise segment in ICA domain corresponding to \mathbf{x} and \mathbf{n} . All the above variables, \mathbf{x} , \mathbf{n} , γ and ν , are $M \times 1$ vectors.

Let $\hat{\zeta}$ denote the estimated ζ in ICA domain based on \mathbf{x} . By applying the inverse transformation to $\hat{\zeta}$, we obtain the enhanced speech segment vector, \mathbf{z} , as

$$\mathbf{z} = \mathbf{W}^{-1} \cdot \hat{\zeta}. \quad (4)$$

Our task therefore is to estimate ζ given γ .

Wiener filter is the linear statistically optimum discrete-time filter that minimizes the mean-square error between the estimated signal $\hat{\zeta}$ and the clean speech signal ζ . The estimated signal $\hat{\zeta}$ is obtained by filtering the received speech in ICA domain, γ , with a linear filter F ,

$$\hat{\zeta} = F \cdot \gamma, \quad (5)$$

where F is a $M \times M$ matrix to be determined.

Because the car noise is slow time-varying, we assume the filter input is a wide-sense stationary stochastic process with zero mean. The mean-square error is

$$J = \mathbf{E}\{\|\zeta - \hat{\zeta}\|^2\} = \mathbf{E}\{\|\zeta - F \cdot \gamma\|^2\}, \quad (6)$$

where \mathbf{E} denotes the statistical expectation operator. The problem therefore is to find F that minimize the mean-square error J . By assuming that the signal ζ and the noise ν in equation (3) are decorrelated and both have zero mean, the equation (6) turns into

$$\begin{aligned} J &= \mathbf{E}\{\|\zeta - F \cdot \gamma\|^2\} \\ &= \text{trace} \left(\mathbf{E}\{\zeta\zeta^T\} - F \cdot \mathbf{E}\{\gamma\zeta^T\} - \mathbf{E}\{\zeta\gamma^T\} \cdot F^T \right. \\ &\quad \left. + F \cdot \mathbf{E}\{\gamma\gamma^T\} \cdot F^T \right) \\ &= \text{trace} \left(\mathbf{R}_\zeta - F \cdot \mathbf{R}_\zeta - \mathbf{R}_\zeta \cdot F^T \right. \\ &\quad \left. + F \cdot (\mathbf{R}_\zeta + \mathbf{R}_\nu) \cdot F^T \right), \end{aligned} \quad (7)$$

where T denotes transpose, $\text{trace}(A)$ denotes the trace of matrix A , \mathbf{R}_ζ is the covariance matrix of the ICA transformed speech data, ζ and \mathbf{R}_ν is the covariance matrix of the noise, ν .

Taking the derivative of (7) with respect to F and setting the result to zero yields

$$F \cdot (\mathbf{R}_\zeta + \mathbf{R}_\nu) - \mathbf{R}_\zeta = \mathbf{0}, \quad (8)$$

where $\mathbf{0}$ is a $M \times M$ zero matrix. The optimum filter F that minimizes the mean-square error J is then:

$$F = \mathbf{R}_\zeta \cdot (\mathbf{R}_\zeta + \mathbf{R}_\nu)^{-1}. \quad (9)$$

The noise covariance matrix \mathbf{R}_ν can be computed easily from the covariance matrix of the time-domain noise \mathbf{n} , \mathbf{R}_n : $\mathbf{R}_\nu = \mathbf{W}\mathbf{R}_n\mathbf{W}^T$. On the other hand, taking into account that the ICA transformed speech signal components, ζ_i , $1 \leq i \leq M$, are statistically independent, the covariance matrix of a set of ICA transformed speech data \mathbf{R}_ζ is

$$\mathbf{R}_\zeta = \text{diag} (R_{\zeta_i}), \quad (10)$$

where $\text{diag} (R_{\zeta_i})$ is a $M \times M$ diagonal matrix with the i th diagonal element $R_{\zeta_i} = E\{|\zeta_i|^2\}$.

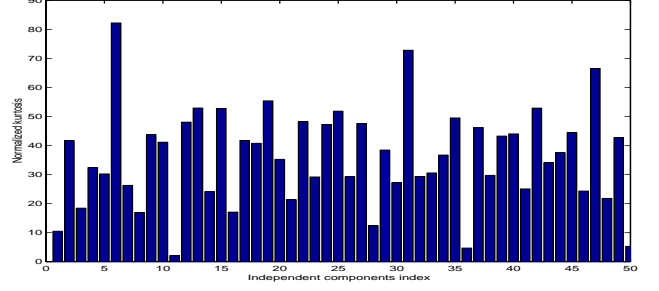


Fig. 1. Normalized kurtosis of the ICA transformed speech.

R_{ζ_i} can be estimated easily from the model of the ICA transformed speech signal. To find the characteristic of the ICA transformed signals, we studied their normalized kurtosis, which is defined as

$$\text{kurtosis}(\zeta_i) = \frac{E\{\zeta_i^4\}}{[E\{\zeta_i^2\}]^2} - 3, \quad (11)$$

Fig. 1 shows an example of the normalized kurtosis of the ICA transformed speech.

It is obvious that the kurtosis for all ICA transformed speech data are significantly larger than zero that is the value of the normalized kurtosis of a random variable with Gaussian distribution. So all independent components have super-Gaussian distributions. Laplacian distribution is a commonly used super-Gaussian distribution, therefore we assume that the ICA transformed speech data has Laplacian distribution, i.e.,

$$p(\zeta_i) = \frac{1}{2\lambda_i} \exp\left(-\frac{|\zeta_i|}{\lambda_i}\right). \quad (12)$$

From the distribution of the ICA transformed speech data ζ_i , we have

$$R_{\zeta_i} = \int_{-\infty}^{\infty} \zeta_i^2 p(\zeta_i) d\zeta_i = 2\lambda_i. \quad (13)$$

Substituting (13) into (10), then (10) into (9) and finally (9) into (5) gives

$$\hat{\zeta} = \text{diag} (2\lambda_i) \cdot \left[\text{diag} (2\lambda_i) + \mathbf{R}_\nu \right]^{-1} \cdot \gamma. \quad (14)$$

Once we obtain the estimate of ζ , substituting it into equation (4) gives the enhanced speech signals in time domain. Finally, reshaping the enhanced speech signal from matrix form to vector form reconstructs the enhanced waveform. Fig. 2 shows the block diagram of the ICA-based single channel speech enhancement. What remains to be estimated are the unmixing matrix \mathbf{W} of the ICA transform, parameters λ_i , $i = 1, \dots, M$, required in the probability density function of the ICA transformed speech data ζ , and noise covariance matrix \mathbf{R}_n required in obtaining \mathbf{R}_ν .

3. SIMULATION

In this section, the noise reduction capability of the proposed algorithm for car noises is studied. We first describe the approach to learn the ICA model and estimate the unknown parameters λ_i in training phase. Then the performance of the proposed ICA-based single channel speech enhancement algorithm using Wiener filtering is investigated.

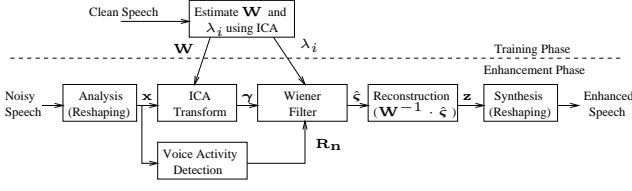


Fig. 2. Block diagram of the ICA-based single channel speech enhancement system with Wiener filtering.

3.1. Learning \mathbf{W} and Estimating λ_i

Several practical methods for estimating the ICA models have been proposed and applied [6]. In our study, a robust and efficient principle based on maxima of nongaussianity is employed to estimate the ICA model. The nongaussianity is measured by negentropy since it is robust and in some sense optimally estimates the nongaussianity. A novel nonquadratic function that is used to approximately measure the nongaussianity is applied. It is defined by $G(\xi) = -(|\xi| + 1)e^{-|\xi|} + 1$. It is more robust than those defined in [6] and yields less dependence between the ICA transformed components. To learn the ICA model, speech sentences from four different speakers of TIMIT database are used. The sampling frequency is 16KHz. The unmixing matrices \mathbf{W} for each speaker are learned from two sentences of that particular speaker with a total length of about six seconds. The FastICA code [7] is used to implement the algorithm of estimating the unmixing matrix \mathbf{W} based on the maxima of nongaussianity. The speech is analyzed (reshaped) into frames composed of 50 samples, i.e., 3.1ms time interval at 16KHz sampling rates, with 98% overlap between successive frames. Hamming window is applied to each frame in all cases. The stop criterion of the iterative FastICA code is 0.0001. The initial guess for \mathbf{W} is the identity matrix.

After obtaining the unmixing matrix \mathbf{W} , we can estimate the parameters λ_i required in (12), from the training speech data. They approximate the parameters λ_i of ICA domain testing speech data.

In this study, maximum likelihood technique is employed to estimate λ_i . For K observations of training speech frames, the likelihood function of the i -th independent component is

$$p(\xi_i | \lambda_i) = \prod_{k=1}^K \frac{1}{2\lambda_i} \exp \left\{ -\frac{|\xi_i(k)|}{\lambda_i} \right\}, \quad (15)$$

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 unmixing matrix \mathbf{W} and reshaped training speech frame $\mathbf{s}^{(train)}$.

Taking the derivative of the natural logarithm of (15) with respect to λ_i leads to

$$\frac{\partial \ln p(\xi_i | \lambda_i)}{\partial \lambda_i} = \frac{1}{\lambda_i^2} \left(\sum_{k=1}^K |\xi_i(k)| - K\lambda_i \right). \quad (16)$$

Setting the result of (16) to zero forms the likelihood equation. The maximum likelihood estimate of λ_i is the solution of the likelihood equation, that is

$$\lambda_i = \frac{1}{K} \sum_{k=1}^K |\xi_i(k)|. \quad (17)$$

3.2. Speech Enhancement in Car Noise Environments

In this subsection, we investigate the performance of the proposed algorithm on car noise reduction. The noise is added to speech sentences and scaled so that the input global SNR of the corrupted waveform ranges from -5 dB to 20 dB. To be more specific, let $x(m) = s(m) + \beta n_c(m)$ be the received signal, where β is a scale, and $\hat{s}(m)$ is the estimated signal. The global SNRs are defined as

$$\begin{aligned} input - gSNR_{dB} &= 10 \log_{10} \frac{E\{|s(m)|^2\}}{\beta^2 E\{|n_c(m)|^2\}}, \\ output - gSNR_{dB} &= 10 \log_{10} \frac{E\{|s(m)|^2\}}{E\{|\hat{s}(m) - s(m)|^2\}}, \end{aligned} \quad (18)$$

where E is the sample average.

Eight experiments are carried out as in Table 1. In this table, 'F1', 'F2', 'M1' and 'M2' represent the first female speaker, the second female speaker, the first male speaker and the second male speaker, respectively. 'SA1' and 'SA2' represent two different sentences. Therefore, in all the experiments, the training speech is uncorrelated to the testing speech.

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

Table 1. Training and testing data used in experiments.

In testing phase, the speech is also analyzed(reshaped) to frames composed of 50 samples with 98% overlap between successive frames. Hamming window is applied to each frame in all experiments. The noise correlation \mathbf{R}_n required for computing \mathbf{R}_ν in (14) is obtained from the covariance of reshaped noise matrix \mathbf{N} .

Fig. 3 illustrates an example of the waveforms of the original clean speech, the received noisy speech and the enhanced speech. The upper plot represents the original clean speech, the middle plot is for the received noisy speech and the lower plot shows the enhanced speech. The input global SNR is -5 dB. Clearly, the noise has been effectively reduced. Comparing the enhanced speech with the original clean speech and the received noisy speech by listening to them, we found that most of the noise is removed with no noticeable distortion.

To give a quantitative analysis of the performance of the proposed algorithm, four objective speech quality criteria, global SNR, segmental SNR, Itakura distance and weighted-spectral slope measure (WSSM) [8], are computed. Fig. 4 presents the objective evaluation results when Wiener filtering is applied in ICA domain. The solid line represents the evaluation results for the enhanced speech. The dashed line shows the quality measure for the input noisy speech.

The upper-left plot presents the results of the average global SNR. The proposed algorithm provides high improvement. At -5 dB input global SNR, the SNR improvement is about 18.6 dB. At 5 dB input global SNR, the SNR improvement is about 12.6 dB, and at 15 dB input global SNR, the SNR improvement is about 5 dB.

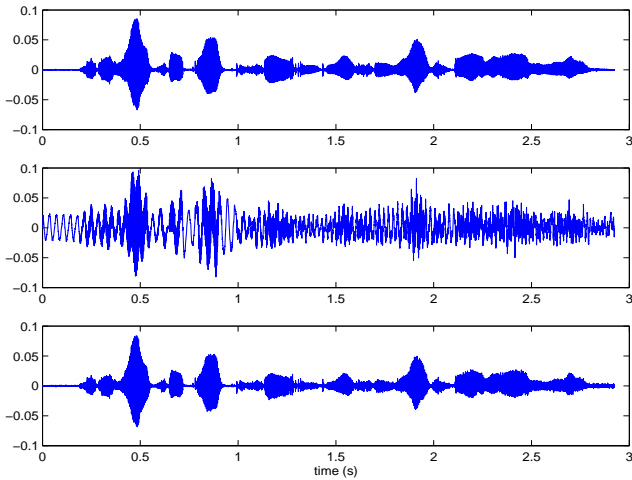


Fig. 3. Waveforms of the original clean speech, the received noisy speech and the enhanced speech when the input global SNR is -5 dB. Upper plot: original clean speech, Middle plot: received noisy speech, Lower plot: enhanced speech. (analysis frame size: 50 samples, analysis overlap rate: 98 %.)

The upper-right plot illustrates the results of the segmental SNR defined in [8], equation (9.7). The SNR thresholds are -20 dB for the lower bound and 20 dB for the upper bound. It is better related with speech quality evaluation than the global SNR measure. At -5 dB input global SNR, the segmental SNR improvement is about 13.3 dB. At 5 dB input global SNR, the segmental SNR improvement is about 8 dB, and at 15 dB input global SNR, there is still about 2 dB segmental SNR improvement.

The lower-left shows the results of the Itakura distance defined in [8], equation (5.191). The enhancement procedure improves the same way in distortion reduction for all input global SNR. The distortion reduction is more than 50%. When the input global SNR is less than 10 dB, the distortion suppression is higher than 75%.

The lower-right presents the results of the WSSM defined in [8], equation (9.14). WSSM measure illustrates a clear increase of intelligibility by the proposed algorithm. When the input global SNR is less than 5 dB, the improvement is higher than 70%. There are about 62%, 45% and 27% improvement when the input global SNR is 10 dB, 15 dB and 20 dB, respectively.

The proposed enhancement approach was applied successfully to reduce car noise. As regard to the computational load, one of the most time-consuming tasks of the algorithm is learning of the unmixing matrix \mathbf{W} , which is done off-line. The on-line computations are fast.

4. CONCLUSIONS

In this paper, we propose a single channel speech enhancement algorithm based on the ICA technique and Wiener filtering. The speech frames are transformed into their independent basis functions by a unmixing matrix, which is learned from a large ensemble of clean speech frames. The Wiener filter is applied to estimate the clean speech from the received noisy speech in ICA domain. The ICA model and the prior knowledge of speech required in Wiener filtering is computed off-line only once during the training part. No subsequent speech signal estimation is performed in the noise

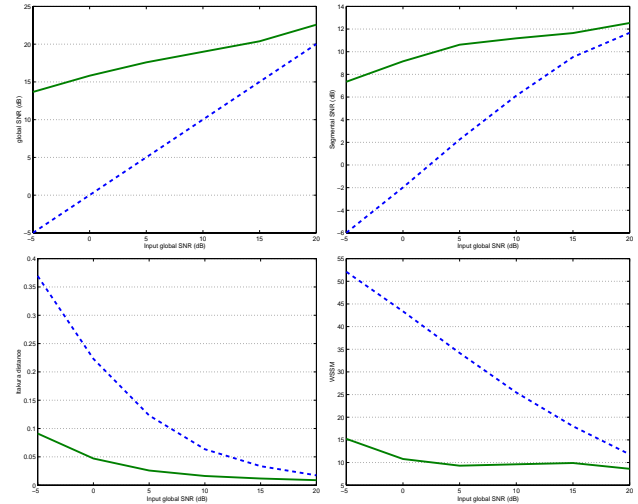


Fig. 4. Objective quality measure results as a function of input global SNR. Upper-left plot: Average global SNR, Upper-right plot: Segmental SNR, Lower-left plot: Itakura distance, Lower-right plot: WSSM. Solid line represents the measure of the enhanced speech signal. Dashed line represents the measure of the input noisy speech. (analysis frame size: 50 samples, analysis overlap rate: 98 %.)

removing phase. Simulations show the proposed enhancement approach is able to reduce car noise significantly with respect to four objective quality measure criteria. The computational load is moderate and can be adjusted according to the application.

5. REFERENCES

- [1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.
- [2] 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.
- [3] W. G. Knecht, M. E. Schenkel, and G. S. Moschytz, "Neural network filters for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 6, pp. 433–438, Nov. 1995.
- [4] M. Dahl and I. Claesson, "Acoustic noise and echo cancelling with microphone array," *IEEE Trans. Vehicular Technology*, vol. 48, no. 5, pp. 1518–1526, Sep. 1999.
- [5] 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.
- [6] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis*, Wiley, New York, NY, 2001.
- [7] * * *, *The FastICA package for MATLAB*, Neural Networks Research Center, Helsinki University of Technology, <http://www.cis.hut.fi/projects/ica/fastica/>.
- [8] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*, IEEE Press, New York, NY, 2000.