Noise Cancellation | a Posteriori Probability, Maximum Criteria | Independent Component Analysis

Technology Reports

High-speed Noise Cancellation with Microphone Array

We propose the use of a microphone array based on independent component analysis as a method for high-speed elimination of noise in speech input to mobile terminals. The effectiveness of the proposed method is confirmed by evaluation experiments that reproduce an actual mobile environment.

Zhipeng Zhang and Minoru Etoh

1. Introduction

Noise cancellation is an important technology for improving call quality for mobile terminals and for speech user interfaces such as speech recognition and speech translation. Techniques for enhancing speech in signals that include background noise are being researched widely. The single microphone-based spectral subtraction [1] has been widely known as a technique for suppressing background noise. The technique differentiates between intervals with speech and intervals without speech. The input signal for the intervals without speech is recognized as background noise and use it to create a noise spectrum. The noise spectrum is then subtracted from the input signal for the intervals with speech and noise are mixed to suppress the noise. Accordingly, good performance is achieved in the noise cancellation when the noise signal is stationary and the technique is easily implemented, so this technique is currently in wide use. When the background noise is a non-stationary signal, however, such as in a noisy restaurant or with vehicles coming and going in busy traffic, good noise cancellation performance cannot be obtained. For this reason, methods that use multiple microphones (microphone arrays) are being investigated. Since a microphone array uses spatial information such as the phase differences among the signals that arrive from the sound source to each microphone to suppress noise, it has better noise suppression performance than a single microphone that assumes a stationary noise signal. There are a beam forming method [2] and a Blind Source Separation (BSS) method [3] that uses Independent Component Analysis (ICA) in the microphone array methods. The comparison of these two methods are shown in **Table 1**.

Because of limited installation space and computing power, it is currently very difficult to mount many microphones on a mobile terminal. Nevertheless, practical use of a small microphone array with a limited number of microphones is now possible. The beam forming method has a long history, and PDA terminals that combine adaptable beam forming and nonlinear signal processing for hands-free communication are already on the market. However, beam forming in principle assumes that the desired sound source is in a different position from other sound sources, so when the posi-

Table 1 Comparison of methods based on a microphone array

	Beam forming method	ICA-based BSS method
Advantages	Already proven in commercial use	Can automatically follow a moving sound source
Disadvantages	Poor performance in automatic tracking of a moving sound source	Estimation of the separation matrix takes computation Hardware and microphone cost are barriers to commercialization

tion of the sound source changes or the noise and target sound sources are in the same direction, the noise canceling performance may be reduced.

As opposed to the beam forming method, which separates sound sources that are in different spatial locations, the ICA-based BSS method uses signal independence to separate sound sources that have independent statistical properties, and so does not require location information in principle. That is to say, even if the noise and the target sound source are in the same direction, only the target sound can be extracted, and has an advantage of a wider range of application. On the other hand, however, the ICA requires successive learning of the statistical properties (maximization of a non-Gaussian distribution[*1] to be exact). The solution to this optimization problem requires nonlinear iterative computing, which makes the ICA unsuitable for real-time processing. Although a module that achieves greatly improved separation performance in a real-time actual environment by implementing high-speed computation has been developed, the implementation cost is high because it requires dedicated hardware and a directional microphone. This becomes an obstacle to commercialization as a product.

In this article, we propose a new ICA-based BSS method for obtaining a high quality speech signal with a small number of microphones. The proposed method uses the fact that the parameters of the transfer function[*2] from the user's mouth to the microphone of the mobile terminal settle within a prescribed range, and adopts the Maximum a Posteriori Probability (MAP) of these parameters. This is a way to estimate the parameters, and the parameters are estimated so as to maximize the a posteriori probability of the parameters based on the speech data. Furthermore, since the ICA of this method converges faster than the conventional ICA, speech of higher quality can be extracted.

Below we describe evaluation experiments on ICA-based noise cancellation and the ICA that makes use of the transfer function.

2. Noise Cancellation Using ICA

In the previous chapter, ICA-based BSS was described as a method for separating out the target signal. The method estimates multiple linear mixed signals without using any knowledge about the original signal or the mixing process. There are two types of BSS method: time domain ICA and frequency domain ICA. The proposed method uses the frequency domain ICA for simplicity in handling the transfer function. Furthermore, the environment is assumed to have two sound sources, which are speech (target sound source) and noise (interference sound source) to simplify the target system. This makes it possible to use only two microphones, thus reducing both the computational complexity and the implementation cost.

2.1 Model for Mixed Signals (Measured Signal) in an Actual Environment

The model for mixed signal separation in a two-microphone ICA system for mobile terminal is shown in **Figure 1**, where s represents the sound source signal. The term s_1 is the user speech from the target sound source and s_2 is noise from the interference sound

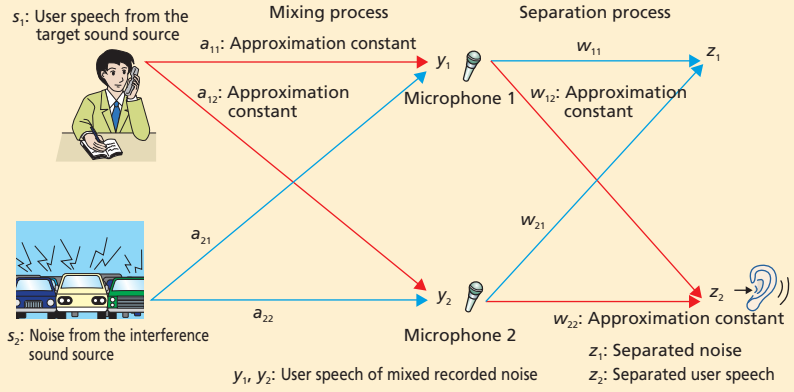

Figure 1 Mixed signal separation model for a two-microphone ICA system for mobile terminals

***1 Non-Gaussian distribution**: The property of a probability variable that does not show a Gaussian probability distribution.

***2 Transfer function**: Ratio of the Laplace transform of the output signal and the Laplace transform of the input signal in a transmission system.

source. The signals y_1 and y_2 detected at the two microphones are recorded via the transmission path from the sound source signal. Denoting the transfer function from the signal source to the microphone as A, the relation between the signal source and transfer function and the measured data is linear, $y=As$. However, A is a mixed array, and each element of A represents the transfer situation from the respective target and interference sound sources to the two microphones. Here, the interference sound source and the target sound source are assumed to be statistically independent.

2. 2 Separation Signal Model

The BSS method uses the signal independence to reconstruct the original signal from the mixed signal y. The separation matrix W is obtained and the separation signal z is obtained from equation (1).

$$z = Wy \tag{1}$$

If the entire transfer function from the signal source to the microphone is known, then z can be reconstructed to the sound source signal by calculating W with $W=A^{-1}$, but the details of transfer function A are not known in advance. Also, the transfer function changes if the sound source moves, some means of changing the separation matrix W to track the moving sound source is required. Therefore, on the basis of the assumed independence with respect to the signal source, we adopted the BSS method that reconstruct the original signal source, by estimating the separation matrix W for the independence to be maximum.

2.3 Estimation of the Separation Matrix

Here we explain a method based on the maximum likelihood criterion[*3], which is often used for estimating the separation matrix [4][5]. If $p(y)$ is the probability distribution function for measured signal y, then the likelihood with W as a variable is expressed as $p\{y(t)/W\}$. Actually, the log of the likelihood is used more often for convenience in computation. Estimation with the log maximum likelihood is expressed by the following equation (2).

$$\hat{W} = \arg\max \sum_{t=1}^{T} \log p\{y(t)/W\} \tag{2}$$

Generally, $\hat{W}$ cannot be solved analytically by using this equation, so an iterative computation using an equation such as equation (3) must be used.

$$W_{i+1} = W_i + \eta \Delta W \tag{3}$$

From the gradient method[*4],

$$\Delta W = \frac{\partial \log p\{y(t)/W\}}{\partial W} = (I - E[\phi(y) y^T])W \tag{4}$$

Here, I is the unit matrix, $E[X]$ is the expected value of X, and $\phi(y)$ is the differential of the probability distribution function of y. Finally, the updating equation for W is as follows.

$$W_{i+1} = W_i + \eta \{I - E[\phi(y) y^T]\} W_i \tag{5}$$

This W is used as the basis for independent component separation in the frequency domain, and then the frequency separation signal is reconverted into the time domain to obtain the independent signals in the time domain.

3. ICA Based on a Transfer Function

Conventional estimation method based on the maximum likelihood criterion is one of the general non-linear optimization methods that involve multiple local optima and require iteration. Unless some means is contrived to overcome it, the result may depend on the initial value and the number of iterations needed may be large and indeterminate. These issues are particularly troublesome in the case of non-stationary noise, for which tracking is a strikingly difficult problem. To solve that problem, we propose using the a priori knowledge that the parameters of the transfer function for the space between the user's mouth and the microphone are confined to a certain range to achieve an optimization method that is both fast and stable.

The flow of the proposed method is

***3 Maximum likelihood criterion**: A criterion in which the probability of obtaining the observed data is maximum, assuming a particular probability model.

***4 Gradient method**: One of the methods for numerical optimization using the slope of a function (differential of a vector).

shown in **Figure 2**. A major feature of the proposed method is that the computation is done in two stages, which are initialization and iterative estimation. In the first stage, the initial values of the transfer function parameters are obtained from the relation between the user's speaking position and the microphone position. The obtained initial values are then used to estimate the separation matrix. In the mobile phone environment, the position of the user's mouth relative to the microphone can be considered as constrained to a certain range. Therefore, we use the parameters of the transfer function for between the user's mouth and the microphone. For the background noise, on the other hand, the parameters are unknown, but they can be estimated by using data from the noise intervals at the beginning of the speaking. If the parameter distribution is even partially known in advance, parameter estimation can be formalized by MAP estimation rather than maximum likelihood estimation, allowing accurate and stable estimation of the separation matrix with fewer iterations. Because the position of the user's mouth relative to the microphone is confined to a certain range, we can assume that a part of the mixed matrix (a_{11} and a_{12}) to be nearly constant, and the part that corresponds to the separation matrix (w_{12} and w_{22}) can also be assumed as nearly constant.

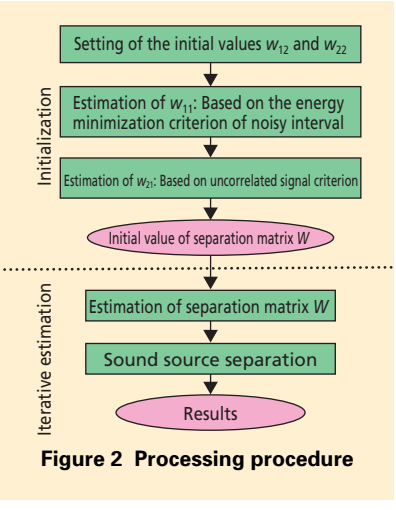

Figure 2 Processing procedure

3.1 Estimation of the Initial Values

We first measure the frequency response between the user's mouth and the microphone. Those measurements are used in the following procedure to estimate the initial values of the separation matrix elements that relate to the transmission of the target sound source.

1) Set the Initial Value of w_{11}

Detect intervals during which the user is not speaking. Use the measured data from the speechless intervals (y_1 and y_2) and equation (6) to obtain w_{11} such that of energy of output z_1 is minimized.

$$w_{11} = \arg\min R[z_1, \bar{z}_1] \tag{6}$$

Here, $R[x, y]$ expresses the correlation of x and y [5].

Estimate w_{11} from equations (5) and (6) in the following way.

$$w_{11} = a_1(R_{22} - R_{21})/(R_{11} - R_{12}) \tag{7}$$

Here,

$$R_{ij} = R[y_i, y_j].$$

2) Set the Initial Value of w_{21}

Based on the uncorrelated signal criterion between z_1 and z_2.

$$R[z_1, z_2] = 0 \tag{8}$$

With this criterion, obtain the initial value of w_{21}.

$$W_{21} = -\frac{W_{22}W_{11}E(y_1y_2) + W_{12}W_{22}E(y_2y_2)}{W_{11}E(y_1y_1) + W_{12}E(y_1y_2)} \tag{9}$$

3.2 Estimation of the Separation Matrix

Generally, if there is a priori knowledge about the parameters, use of the MAP criterion for estimation is considered effective. The a posteriori probability of the separation matrix, $p(W/y)$, is expressed as the product of the a priori probability of W, $p(W)$, and the likelihood, $p(y/W)$.

$$p(W/y) = p(W)p(y/W) \tag{10}$$

As we can see from the above equation, if, without prior knowledge of W, $p(W)$ is assumed to be a uniform distribution, then the MAP criterion and the maximum likelihood criterion become the same. Given a priori probability $p(W)$ about W, a more accurate estimation is possible. The equation for estimation of W on the basis of the MAP

criterion is as follows.

$$\hat{W}=\arg\max\sum_{t=1}^{T}\log\left[p(W)p\{y(t)/W\}\right] \quad (11)$$

Here, assuming that a priori probability $p(W)$ concerning W is a normal distribution, the density function $p(W)$ is as shown below. The expected value μ is taken as the initial value of the W obtained in Section 3.1. The variance σ^2 represents the change in the prior estimated value of the separation matrix.

$$p(W)=\frac{1}{\sigma\sqrt{2\pi}}\exp\{\frac{-(W-\mu)^2}{2\sigma^2}\} \quad (12)$$

When estimating W with the gradient method based on the MAP criterion,

$$\Delta W=\frac{\partial\log[\,p(W)\,p\{y(t)/W\}\,]}{\partial W}=\frac{\partial\log p(W)}{\partial W}+\frac{\partial\log[\,p\{y(t)/W\}\,]}{\partial W} \quad (13)$$

The second term in equation (13) is the same as the maximum likelihood estimation, and the first term is as follows.

$$\frac{\partial\log p(W)}{\partial W}=(W-\mu)/\sigma^2 \quad (14)$$

Thus,

$$\Delta W=\eta\{I-\phi(y)y^T\}W+(W-\mu)/\sigma^2 \quad (15)$$

The update formula is

$$W_{i+1}=W_i+\eta\{I-\phi(y)y^T\}W_i+(W_i-\mu)/\sigma^2 \quad (16)$$

The separation matrix estimated as above is used to perform the separation and extract the target signal.

4. Evaluation Experiments

4.1 Evaluation Data

We evaluated the proposed method by recognition of digits in connected speech using 30 digits uttered continuously by one female speaker. The sampling rate was 16 kHz. We used the mixed matrix

$$A=\begin{bmatrix}2, & 3\\ 4, & 1\end{bmatrix}$$

to create noisy speech data by mixing airport noise and noise-free speech (in the frequency domain). The experiments described below assume that part of the separation matrix (w_{12}: 3.0, w_{22}: 2.0) is already known.

4.2 Overview of the Speech Recognition Experiments

For the speech recognition, we used the Hidden Markov Model (HMM)[*5] speech recognition software [6] openly published by the University of Cambridge. The software uses a 12-dimensional feature vector that comprises the Mel-Frequency Cepstrum Coefficient (MFCC)[*6] frequency characteristics and the normalized power[*7]. The HMM parameters include the countable states and the probability distribution function for the output of each state. In the speech recognition, the output probability function for each state is represented by a mixture of multiple normal Gaussian distributions. In these experiments, the HMM parameter for the number of states is five, and the number of normal Gaussian distribution mixtures is four for each state.

4.3 Evaluation Results

Samples of the speech signal extracted by the BSS method based on the conventional and proposed ICA are shown in **Figure 3**. The proposed method suppresses the noise component in the speech signal more than does the conventional method. To confirm the effectiveness in practical use, we performed evaluation experiments in which the method was used as speech recognition preprocessing. The target sound source was extracted by sound source separation and then evaluated by speech recognition. The proposed method and the conventional method evaluation results (accuracy, %) are shown in **Figure 4**, where the horizontal axis is the number of iterations required for separation matrix estimation. With a single estimation, the proposed method performed with about the same results as did the conventional method with multiple rounds of estimation results. We confirmed that the proposed method improved the recognition rate compared to the conventional method by from 79% to 84%.

5. Conclusion

We have described noise cancellation technology that uses a microphone

***5 HMM:** A statistical method for modeling indeterminate time series data.
***6 MFCC:** A series of speech feature coefficients modeled on human auditory perception.
***7 Normalized power**: The normalized value of a speech signal power in log domain.

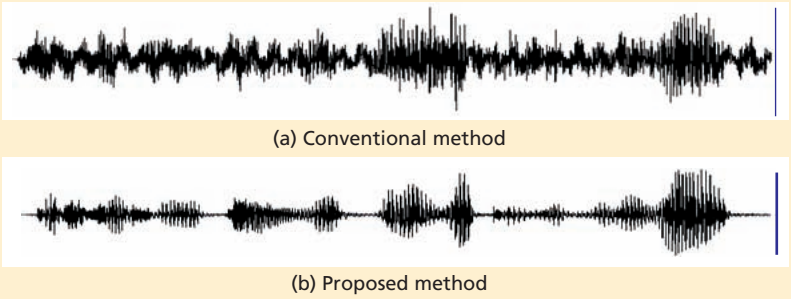

Figure 3 Speech signal sample: BSS method based on the conventional and proposed ICA

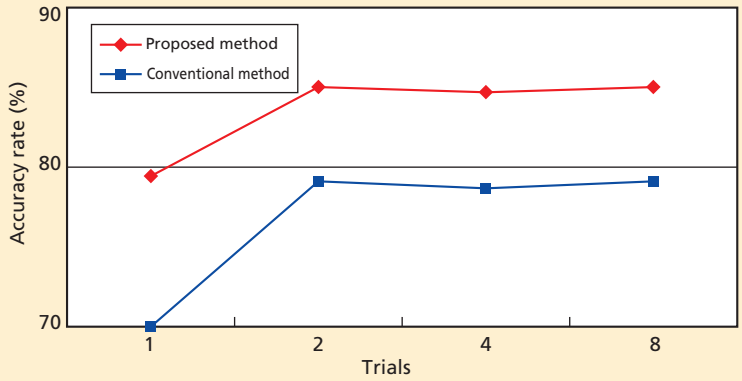

Figure 4 Evaluation results

array. This technology uses a two-microphone array and introduces an optimization method that uses sound source location information obtained from an actual mobile phone environment to highly general ICA statistical signal processing. The experiments, which reproduced a use scenario in an actual mobile phone environment, confirmed that measuring and using the parameters of the transfer function between the user's mouth and the microphone resulted in better speech recognition performance with less computational complexity compared to the conventional method.

We expect this microphone array noise cancellation technology to broaden the scope of future speech communication and serve as a basic technology for speech recognition and translation services.

REFERENCES

[1] S.F. Boll: "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on ASSP, No. 2, pp. 113-120, 1979.

[2] Y. Kaneda: "Audio Systems and Digital Processing," 1995.

[3] T. Nishikawa, S. Araki, S. Makino and H. Saruwatari: "Optimization of Band Divisions in Blind Source Separation using Band-division ICA," 2001 Spring Meeting of the Acoustical Society of Japan, 2001.

[4] T-W. Lee, et al.: "Independent Component Analysis Using an Extended Infomax Algorithm for Mixed Sub-Gaussian and Super-Gaussian Sources," Neural Computation, Vol. 11, pp. 417-441, 1999.

[5] A. Bell and T. Sejnowski: "An Information Maximization Approach to Blind Separation and Blind Deconvolution," Neural Computation, Vol. 7, pp. 1129-1159, 1995.

[6] M.J.F. Gales and P.C. Woodland: "Recent advances in large vocabulary continuous speech recognition: An HTK perspective," ICASSP, 2006.