ICA-based Technique in Air and Bone-Conductive Microphones for Speech Enhancement

Zhipeng Zhang, Kei Kikuiri, Nobuhiko Naka and Tomoyuki Ohya

Multimedia Laboratories, NTT DoCoMo
3-5 Hikari-no-oka, Yokosuka, Kanagawa, 239-8536 Japan
zzp@mml.yrp.nttdocomo.co.jp

Abstract

How to obtain clean speech signal in noisy environments is a crucial issue for improving the appeal of mobile phones. This paper proposes to supplement the existing normal air-conductive microphone with a bone-conductive microphone for noise reduction. We propose to apply the ICA (Independent Component Analysis)-based technique to the air and bone-conductive microphone combination for speech enhancement. The speech signal output by the bone-conductive microphone has the advantage of very high SNR, which well supports the generation of a clean speech signal in combination with a normal microphone. We evaluate this method by a Japanese digital recognition system. The results confirm that the proposed method can allow a mobile phone to obtain a clean speech signal even if the background noise is relatively high.

1. Introduction

Speech capturing devices are usually disturbed by diffused noise, especially in mobile phone applications where devices are frequently used in noisy environments. In highly adverse conditions, the ambient noise has to be reduced. How to obtain a clean speech signal in noisy environments is a crucial issue if mobile phone communication is to be more widely adopted. A lot of research has been carried out in this area. Spectral Subtraction (SS) is a typical technique [1]. The main disadvantage of SS is that it fails to handle time varying noise well. To achieve even better performance, SPIRIT offers the two-microphone NC (noise canceling) solution [2]: one microphone close to and the other far from the speaker. The first one picks up both the desired signal and the background noise, while the other microphone picks up just the noise. By subtracting the second signal from the first, the speech signal is cleaned due to cancellation of the noise signal. Compared to one-microphone SS method, SPIRIT provides better sound quality and suppresses the annoying noise. However, it is difficult to catch a pure noise-only signal given that the user's voice can travel so widely; this problem restricts the degree of quality enhancement possible.

This paper proposes using a bone-conductive microphone to supplement the existing normal air-conductive microphone of a mobile phone for noise reduction. The bone-conductive microphone outputs a speech signal that has higher SNR than that provided by normal microphones. Speech enhancement by combining air and bone-conductive microphones appears to be a very promising approach.

Liu et al. proposed a speech enhancement method using air and bone-conductive microphones [3]. In their paper, they used the direct filtering method for standard and throat microphones in a noisy environment, and very good results were reported. There are many differences between their work and ours. In their paper [3], they assumed the noise signals in the air microphone and bone channels were zero-mean Gaussian random variables. To estimate these parameters, their method needs non-speech frames that are usually detected by a speech/non-speech detector module. System performance largely depends on the accuracy of voice activity detection (VAD) and it fails when the number of detected non-speech frames is insufficient. Another problem is that the noise in real environments is not always a zero-mean Gaussian random variable; a speech enhancement method that can cope with more general environments is needed.

This paper proposes an ICA (Independent Component Analysis)-based technique [4] to create an effective air and bone-conductive microphone arrangement for speech enhancement. We first explain the signals acquired by air and bone-conductive microphones in mobile phones, and then propose the ICA technique for processing the speech signals. In the next section, we report some experiments and evaluation by the perceptual evaluation and speech recognition. The paper concludes with a general discussion and issues related to future research.

2. Signal acquisition by air and bone-conductive microphones in mobile device

2.1 Air and bone-conductive microphones for mobile phone

The ideal placement of air and bone-conductive microphones in mobile phone use is shown in Fig. 1. The bone-conductive microphone is fixed near the top of mobile phone. When

Fig. 1: bone-conductive microphone fixed in mobile phone.

speaking, the bone-conductive signal can be recorded when the mobile phone, held by the subject's fingers, is pressed to his face (Fig. 2); the normal air microphone picks up the speech signal near mouth. This use style strongly supports integration with mobile terminals.

2.2 Properties of speech signals captured by air and bone-conductive microphones

Compared to the regular microphone, the bone-conductive microphone is insensitive to ambient noise, but the high frequency portion of the speech signal is insufficient. Therefore we can not directly use bone-conductive microphone for mobile communication even though they have high SNR; this is in contrast to the air-conductive microphone which yields a full frequency signal with low SNR. By combining the two speech signals, we are able to significantly suppress the background noise and obtain a signal that covers the full frequency range and has high SNR.

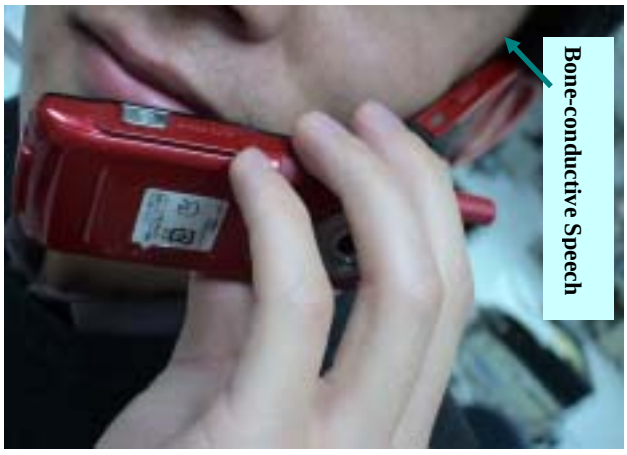

Figure2. Signal acquisition by air and bone-conductive microphones in mobile phone

3. ICA-based technique

3.1 Concept of ICA

ICA is a statistical method that decomposes multivariate data into a linear sum of non-orthogonal basis vectors. ICA has been widely used for multi-channel signal processing [4,5,6]. Assuming that the original signals are independent we can apply an ICA algorithm to blindly recover the unknown sources.

Assume that there is an *M* dimensional zero-mean vector *S(t)* such that the components of *S(t)* =[s_1(t), s_M(t)] are mutually independent. The signals are transmitted through a medium so that an array of sensors picks up a set of signals, *X(t)* =[x_1(t), x_M(t)], each of which has been mixed, delayed and filtered as follows:

$$Xi(t) = \sum_{j=1}^{N} \sum_{i=1}^{p-1} a_{ij} S_j (t - k - D_{ij}) \text{.....(1)}$$

where *Dij* are entries in the delay matrix and there is a *P*-point filter, *aij*, between the *i*th source and the *j*th sensor. It can be simplified as follows:

$$X = AS \text{ .....(2)}$$

where *A* is the mixing matrix. The problem is to recover original signal *S*, given only sensor outputs *X*(*t*). The unmixing matrix W can be formulated as follows:

$$Y = WS \text{ .....(3)}$$

where *Y* is the estimated source signal. If the *W= inv(A)* the estimated signal will be same as the source signal.
The basic assumption of ICA is that the source components are, at each time instant, mutually independent and that each component is white, ie: there are no dependencies between time points. This assumption usually holds for speech signal in the real world [7]. The ICA have bees successfully used in microphone array for speech enhancement [5 7]. However, in ICA-based multi-channel speech signal processing, it is difficult to extract the desired source signal in highly adverse conditions. By applying ICA to a bone-conductive speech signal that has higher SNR, it is easy to extract the desired source signal.

3.2 Learning rule

Learning is performed by maximizing the likelihood of *Y*. The general learning rule is:

$$\Delta W \propto [I - (\frac{\partial \log(p(y))}{\partial y}) y^T] W \text{ .....(4)}$$

where *I* is the identity matrix and *P(y)* is the probability function of *y*. This is the learning rule using the natural gradient extension as described by [8]. We may keep the form of the equation for the full filter system by moving into the frequency domain representation where the elements of the matrices are filters.

A problem with frequency-domain processing is permutation. We need to align the permutation in each frequency bin so that

the separated signal in the time domain contains frequency components from the same source signal. We adopted the method introduced by Asano [9] to solve the permutation problem. After ICA was performed, the uncorrupted speech signal is obtained as well as the noise signal.

4. Experiment

4.1 Task and Data

We evaluate our proposed method on speech recognition. The task of the system is the recognition of connected Japanese digits, each having 2-8 digits, such as “3429” and “246858”. In a preliminary experiment, we used a bone-conductive throat microphone instead of one fixed near the top of phone. The tested speech uttered by one speaker was simultaneously captured by air and bone-conductive microphones. 30 utterances from one speaker were recorded at 16kHz with 16bit resolution in a clean condition (room). The speech signal was corrupted by adding noisy data (“Exhibition hall”) that were recorded in the same room using air and bone-conductive microphones. The noisy data were made at three SNR conditions: 0,5,10 dB. The noise data were non-stationary and the SNR were low.

4.2 Feature vector and HMM

The speech signals were converted into a 25-dimensional acoustic vector consisting of 12-dimensional cepstral-mean-normalized MFCCs and their first derivatives, as well as normalized log energy coefficients. The HMM used in our experiments was a 5 state left-to-right HMM, each state has 4 mixtures.

4.3 Spectrogram of extracted speech data by ICA

ICA was performed to extract clean speech. The length of FFT is 512. The number of iteration in ICA is 1 as to reduce the computational time. Figures 3 and 4 show samples of the noisy signals captured by the air and bone-conductive microphones. We can observe that the high frequency portion of the bone-conductive speech is not ample. Figure 5 shows the signals output by ICA. It is clear that this signal not only has a higher SNR but also is ample at high frequency portions.

4.4 Recognition result

Recognition experiments were performed and Table 1 shows the comparison result (accuracy %) for the three arrangements: air microphone only, bone-conductive microphone only, and ICA output. The proposed method achieved a 4.8% and 7.8% improvement in word accuracy compared to air and bone-conductive microphone arrangements. This confirms the effectiveness of the proposed method.

	0dB	5dB	10dB
Air-conductive Mic	39.6	58.4	82.6
Bone-conductive Mic	43.5	62.7	85.4
ICA	49.3	66.7	88.1

Table1: Comparison of results achieved with air microphone only, bone-conductive microphone only, and ICA.

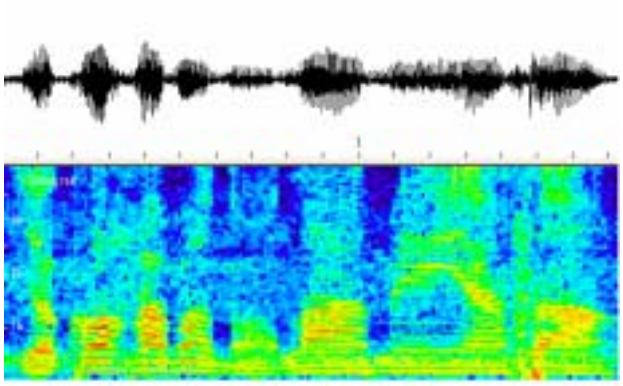

Figure3: Spectrogram of speech data by air-mic.

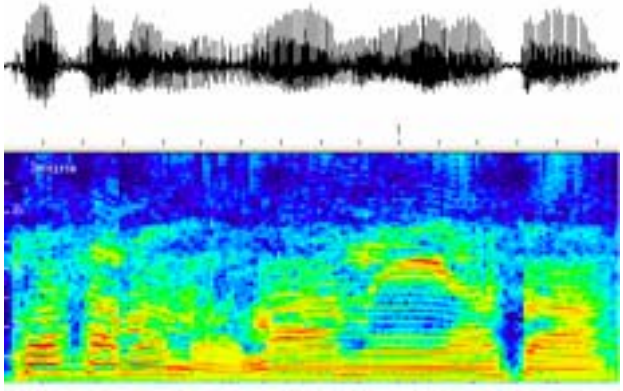

Figure4: Spectrogram of speech data by bone-mic.

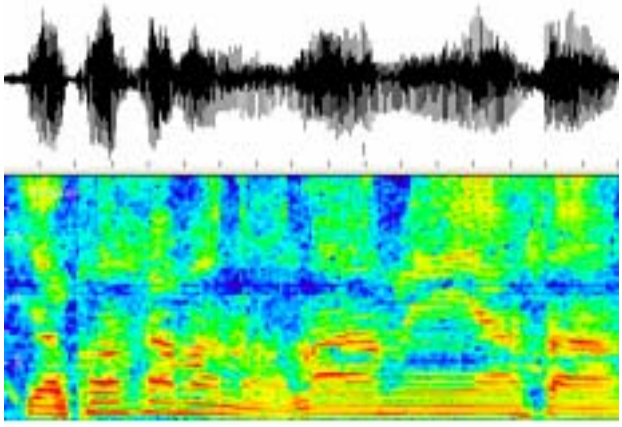

Figure5: Spectrogram of speech data by ICA.

5. Conclusion

This paper reported a new method of using a bone-conductive microphone to supplement an existing normal air-conductive microphone for noise reduction in mobile communication devices. We proposed a ICA-based technique to process the outputs of the air and bone-conductive microphone combination and so enhance speech quality. The proposed method has several advantages: it does not need voice activity detection (VAD), it can handle the noise of real environments, and the physical arrangement of the microphones is practical.

By observing the spectrogram of speech signal obtained by ICA, we found that the proposed method improves speech quality. It not only has a higher SNR but also is ample at high frequency portions. The proposed method has also been evaluated by a speech recognition test. Recognition results show that the proposed method reduces the recognition error compared to the air microphone only and the bone-conductive microphone only arrangements, respectively. Future research includes increasing the variation of test data, investigating the effects of speaker variation, and adding a compensation technique that improves the quality of bone-conductive speech.

Acknowledgements

This research has been conducted in cooperation with Furui Lab at Tokyo Institute of Technology. The authors wish to express their thanks to Prof. Furui and Dr.Iwano for their discussion.

References

[1] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Transactions on ASSP*, vol. 27, No. 2, pp. 113-120. (1979)

[2] A. Guerin, R. Le Bouquin, and G. Faucon, "A Two-Sensor Noise Reduction System: Applications for Hands-free Car Kit," in *EURASIP JASP*, pp.1125-1134. (2003)

[3] Z Liu, Z Zhang, A. Acero, J. Droppo, X. Huang, “Direct Filtering for Air- and Bone-Conductive Microphones”, *IEEE MMSP*,2004

[4]Torkkola, K. (1996). Blind separation of convolved sources based on information maximization. In *IEEE Workshop on Neural Networks for Signal Processing*, 423–432, Kyoto, Japan.

[5]Ehlers, F. and Schuster, H. (1997). Blind separation of convolutive mixtures and an application in automatic speech recognition in noisy environment. *IEEE Transactions on Signal processing*, 45(10):2608–2609.

[6]Lee, T.-W., Bell, A., and Lambert, R. (1997). Blind separation of convolved and delayed sources. In *Advances in Neural Information Processing Systems* 9,758–764. MIT Press.

[7] S. Makino, S. Araki, R. Mukai, and H. Sawada, "ICA-based audio source separation, " in Proc. International Workshop on Microphone Array Systems - Theory and Practice, (2003)

[8] Bell, A. and Sejnowski, T.. “An Information Maximization Approach to Blind Separation and Blind Deconvolution”. *Neural Computation*, 7:1129–1159. (1995)

[9]F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, “A combined approach of array processing and independent component analysis for blind separation of acoustic signals,” *ICASSP*, pp.2729–2732.2001