APPLYING BLIND SIGNAL SEPARATION TO THE RECOGNITION OF OVERLAPPED SPEECH

Tomohiko TANIGUCHI, Shoji KAJITA, Kazuya TAKEDA and Fumitada ITAKURA
Dept. of Information Electronics, Graduate School of Engineering
Nagoya University, Nagoya, 464-01, Japan
Tel:+81 52 789 3629, FAX: +81 52 789 3172, E-mail:takeda@nuee.nagoya-u.ac.jp

ABSTRACT

Blind signal separation method based on minimizing mutual information is applied to deal with multi-speaker problem in speech recognition. Recognition experiments performed under different acoustic environments, in a soundproof room and a reverberant room, clarify that 1) the method can improve recognition accuracy by about 20% where SNR condition is 0 dB, 2) the method is more effective when many speakers' speech exist than the simple overlapped situation, and that 3) the method does not work well under reverberant conditions.

1. INTRODUCTION

Since the multi-speaker situation is quite usual in speech communication, recognizing overlapped speech is one of the most typical and important problems of speech recognition in real world. As known as "cocktail party effect", humans can focus on the particular speaker's speech under the existence of interfering speech. Among various interpretations of cocktail party effect such as speaker localization [1,2] and tracking a formant structure[3], blind separation methods seem to provide the most general framework.

A blind signal separation method based on minimization of mutual information [4] is potentially suited for recognizing speech under multi-speaker conditions, since the separation method does not make any assumptions concerning spatial or spectral characteristics of sound sources except for; 1) the source signals are super-Gaussians, 2) no time difference in the mixing process, 3) the source signals are statistically independent and, 4) the time differences of source signals are identical across the mixed signals.

In this paper, the effectiveness of the separation is discussed in terms of acoustic preprocessing for multi-speaker speech recognition, in the following order. In Section 2, the blind separation method will be briefly described. Then, in Section 3, the experimental conditions for evaluating the method will be described. After discussions on the results of experiments in Section 4, we will conclude this paper in Section 5.

2. METHOD

The signal separation method can be summarized as follows [4].

If the statistically independent source signals $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_\mathrm{N}$ are mixed by an N×N mixing matrix $\mathbf{A}$, the mixed signals $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\mathrm{N}$ are obtained as follows:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1}$$

where $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\mathrm{N})^t$ and $\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_\mathrm{N})^t$. Note that there is no time difference among terms of a source signal, $\mathbf{s}_\mathrm{i}[\mathbf{n}]$, in contributing the mixed signal, $\mathbf{x}_\mathrm{i}[\mathbf{n}]$.

Based upon this modeling, the blind separation of signals can be formalized as a problem of finding mixing matrix $\mathbf{A}$, or more directly, inverse matrix, $\mathbf{A}^{-1}$. From the assumption that the source signals are independent, the inverse matrix $\mathbf{A}^{-1}$ is expected to be estimated as a matrix of minimizing mutual information between the signals obtained by multiplying an unknown matrix, $\mathbf{W}$ to the mixed signal, i.e.

$$\hat{\mathbf{A}^{-1}} = \underset{\mathrm{W}}{\operatorname{argmin}} I\left(\mathbf{W}\mathbf{x} + \mathbf{w_0}\right). \tag{2}$$

Bell et al. found that, as far as $\mathbf{s}$ follows a super-Gaussian distribution, the minimization of the mutual information, $I\left(\mathbf{W}\mathbf{x}\right)$, can be achieved by maximizing joint entropy after nonlinearly squashing $\mathbf{W}\mathbf{x}$ so as to bound its variance. Defining $\mathbf{y}$ as the squashed version of $\mathbf{W}\mathbf{x}$, therefore, the above minimizing procedure can be accomplished by

$$\hat{\mathbf{A}^{-1}} = \underset{\mathrm{W}}{\operatorname{argmax}} H\left(\mathbf{y}\right), \tag{3}$$

$$\mathbf{y} = g\left(\mathbf{W}\mathbf{x} + \mathbf{w}_0\right), \tag{4}$$

where $\mathbf{w}_0$ is a bias vector and $g(x)$ is a non-linear squashing, e.g. sigmoid, function.

Joint entropy can be calculated as the expected value of the log likelihood of joint distribution by

$$H(\mathbf{y}) = \mathbf{E}[-\ln \mathbf{f}_\mathbf{y}(\mathbf{y})], \tag{5}$$

and the joint distribution of $\mathbf{y}$ can be determined in terms of joint distribution of $\mathbf{x}$, and Jacobian $|J|$ of the mapping from $\mathbf{x}$ to $\mathbf{y}$

$$\begin{aligned} H(\mathbf{y}) &= E[-\ln f_{\mathbf{y}}(\mathbf{y})] \\ &= E[-\ln \frac{f_{\mathbf{x}}(\mathbf{x})}{|J|}] \\ &= E[\ln|J|] - E[\ln f_{\mathbf{x}}(\mathbf{x})]. \end{aligned} \tag{6}$$

Since the joint distribution of $\mathbf{x}$ is independent from $\mathbf{W}$, it is enough to find $\mathbf{W}$ which maximizes the first term, $E[\ln|J|]$ to maximize joint entropy of $\mathbf{y}$. The gradient descent method can, then, be applied to search for the $\mathbf{W}$ by differentiating logarithm of Jacobian with respect to each element of $\mathbf{W}$ matrix:

$$\mathbf{W}^{(i)} = \mathbf{W}^{(i-1)} + \lambda \mathbf{\Delta W} \tag{7}$$

$$\Delta \mathbf{W} = E\left[\frac{\partial}{\partial \mathbf{W}} \ln|J|\right]. \tag{8}$$

Where λ is a learning factor and i is the number of iteration. The iteration can be regarded being converged when the $\lambda \Delta \mathbf{W}$ becomes less than predetermined threshold value. Finally, with the estimated matrix $\hat{\mathbf{W}}$ the source signal can be recovered by inverting the mixing

$$\mathbf{s} = \hat{\mathbf{W}}\mathbf{x}. \tag{9}$$

3. EXPERIMENTAL SETUP

3.1 Recording Environment

The overlapped speech are recorded both in a soundproof room and a quiet but reverberant (T_{60} = 0.70 s) laboratory room. The back ground noise levels are 28 (dBA), and 36 (dBA) respectively in those rooms. The arrangement of loudspeakers and microphones for recording multi-speaker speech, are illustrated in Figure 1, for the soundproof room condition. The same loudspeakers and microphones are arranged in the same way in a larger (3m × 5m) laboratory room for the recording under reverberant condition. The test and the interfering speech are generated from loudspeakers 1 and 2 respectively. Two directional microphones are located at the same point so as to form a 90° angle. As shown in Figure 2, the characteristics of the transfer function from a loudspeaker to microphones are almost the same except for their gains. Therefore, the resultant mixed signals satisfy the mixing condition of linear addition, i.e. $x_1 = a_{11}s_1 + a_{12}s_2$. Furthermore, under that condition, the difference of arriving time from the same speaker is negligible between two microphones.

In the other words, $\tau_{11} = \tau_{21}$ and $\tau_{12} = \tau_{22}$ hold for the mixing process of the below form.

$$\begin{cases} x_1[n] = a_{11} \cdot s_1[n - \tau_{11}] + a_{12} \cdot s_2[n - \tau_{12}] \\ x_2[n] = a_{21} \cdot s_1[n - \tau_{21}] + a_{22} \cdot s_2[n - \tau_{22}] \end{cases} \tag{10}$$

Table 1: Analysis Conditions for Recognition experiments.

Common Conditions	
Frame Length	20 ms
Frame Shift	10 ms
DTW	
Sampling Freq.	8 kHz
Feature Vector	16 CH. SGDS
Analysis Freq.	4 - 17 Bark
CSR	
Sampling Freq.	16 kHz
Feature Vector	12 MFCC + ΔMFCC + ΔΔMFCC ΔPOW + ΔΔPOW
Vocabulary	381
Grammar	no grammar

Sources of speech and interfering signals are stored in digital 16 bit form, 16 kHz sampling. The mixed signals are recorded in 8 kHz and 16 kHz for DTW and CSR experiments, respectively.

As for interfering speech, human speech-like noise (HSLN), which is a kind of bubble noise generated by superimposing independent speech signals [5], is utilized. By changing the number of superpositions, we can simulate various multi-speaker conditions, e.g. when HSLN of one superposition is used for interference, the overlap of two speakers is simulated, whereas when the number of superpositions is large, then cock-tail party environment is simulated.

3.2 Recognition Systems

Both DTW isolated word discrimination and HMM continuous speech recognition experiments are performed in speaker dependent manner. The isolated word discrimination task is to discriminate the utterances of a phonetically similar Japanese city name pair. For the experiment, SGDS [6] is used as a spectral measure and 68 utterances of five male speakers are used. The rest of conditions are summarized in Table 1 together with CSR case.

For continuous speech recognition (CSR) experiment, HTK speech recognizer is used with the conditions listed in Table 1. For the CSR experiment each of ten sentences of four male and four female speakers are used as test data. The monophone HMM model is trained by 140 phonetically balanced sentences.

4.RESULTS

4.1 DTW Word Discrimination

Figure 3 shows the recognition accuracy across global SNR of the test speech against the interfering signal of HSLN of 256 superpositions. The baseline performance, using the test signal recorded by microphone 1 with no interfering signal, is shown as the top vertical line (CLEAN). From the directional property of

the microphone 10 to 20% better accuracy is obtained by microphone 1 (MIC 1) than microphone 2 (MIC 2.) However, the accuracy is 5 to 25 % lower than that of clean signal. (Note: the accuracy usually never falls down 50% in the word discrimination task.)

After separation (SEP 1), the recognition accuracy is improved by about 2, 5, 7 and 10 % under SNR of 12, 6, 0 and -6 dB respectively, or the error rate is reduced by two thirds or a half of the pre-separated data. On the other hand, after separation, the accuracy of interfering signal (SEP 2) is greatly reduced. This is also the evidence of the effectiveness of signal separation. From these results, it is clarified that the separation method works well for pre-processing of speech recognition systems.

4.2 HMM Continuous Speech Recognition

The same tendency with isolated word discrimination were observed in HMM continuous speech recognition experiment, which is illustrated in Figure 4, as the word correctness (% Corr.: disregarding insertion errors). The baseline recognition accuracy, however, is quite low, especially when the SNR is lower than 0 dB. By applying the separation, an improvement in accuracy of more than 20 % is obtained in 0 and 6 dB conditions.

The conspicuous difference from the DTW results is that there is saturation in improving performance at 12 dB of SNR, where the signals before separation out perform the separated speech. From these results, it can be concluded that the blind separation improves the performance more effectively when the baseline accuracy is low.

Figure 5 shows the results across the type of interfering speech, i.e. the number of superpositions of HSLN. The SNR is fixed to be 0 dB throughout the experiments. The results include simple overlapping of single speaker's speech of a male (1M) and a female (1F). As shown from the figure, in general, as the number of interfering speech increases, the overall recognition accuracy decreases. However, the improvement obtained by the separation increases as the number of superpositions increases.

4.3 Reverberant Room Results

From the CSR results of laboratory room recording, in Figure 6, it can be shown that the separation method does not work well in the reverberant condition as in the soundproof room. In our previous simulation experiment [7], the most critical condition that governs the separation performance was the time difference in the mixing process (described by equation (10)). In the reverberant room, where various acoustic echo paths exist, this condition does not hold due to the mixing of delayed signal.

5.CONCLUSION

In this paper we evaluated the Bell's blind separation method in order to utilize for preprocessing of recognizing overlapped speech. In both DTW word discrimination and HMM continuous word recognition, the separation method can improve recognition accuracy by more than 10 % where SNR of the signal is below 10 dB. From these results, it can be concluded that the blind separation method is effective to separate overlapped speech.

However, two problems are found; 1) saturation of performance improvement in rather clean conditions and 2) insufficient results under highly reverberant conditions. The latter is a consequence of the theoretical assumption of the method and the most important problem to be solved in future works.

ACKNOWLEDGEMENT

The authors are grateful to Dr. Heni Yehia for his suggestions and discussions on this work.

REFERENCE

[1] J.L.Flanagan, A.C.Surendran and E.E.Jan: "Spatially Selective Sound Capture for Speech and Audio Processing", Speech Communication 13, pp.207-222, 1993

[2] T.Yamada, S.Nakamura and K.Shikano: "Robust Speech Recognition with Speaker Localization by a Microphone Array", Proc. ICSLP-96, Vol.3, pp.1317-1320, 1996.

[3] Thomas W. Parsons:"Separation of speech from interfering speech by means of harmonic selection", J.Acoust.Soc.Am., vol.60, no.4, pp.911-918, 1976.

[4] Anthony J. Bell and Terrence J. Sejnowski: "An information-maximization approach to blind separation and blind deconvolution", Neural Computation, 7, pp.1129-1159, 1995.

[5] D.Kobayashi et al.: "Extracting Speech Features from Human Speech-like Noise.", Proc ICSLP96, vol.1, pp.418-421, 1996

[6] F.Itakura and T.Umezaki: "Distance measure for speech recognition based on the smoothed group delay spectrum", Proc. ICASSP-87, vol.3, pp.1257-1260, 1987.

[7] T.Taniguchi, H.Yehia, S.Kajita, K.Takeda and F.Itakura: "On the Problems in Applying Bell's Blind Separation to Real Environments", ASJ-ASA Third Joint Meeting, Dec. 1996, 1pSP4 pp.2602 (in JASA Special Issue) pp.1257-1260, (in ASJ Proceeding)

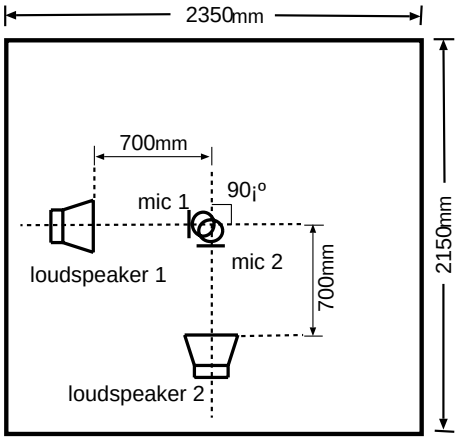

Figure 1: Arrangement of speakers and microphones for recoding overlapped speech. The outer frame corresponds to the size of the soundproof room.

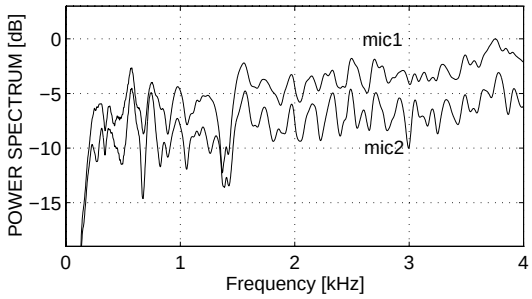

Figure 2: Transfer characteristics from loud speaker 1, to microphone 1 and 2. Since the spectral responses of the two channels are almost identical, the mixed signal can be represented by linear addition, i.e. $x_1 = a_{11}s_1 + a_{12}s_2$.

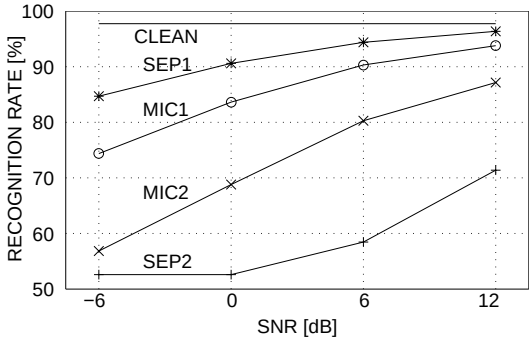

Figure 3: Recognition accuracy, correct rate in DTW based word discrimination, of interfered and separated speech under various SNR conditions. MIC1 is located in front of the loudspeaker 1, which is for the test speech. SEP1 is the separated signal corresponding to the test speech, whereas SEP2 corresponds to that of interfering speech.

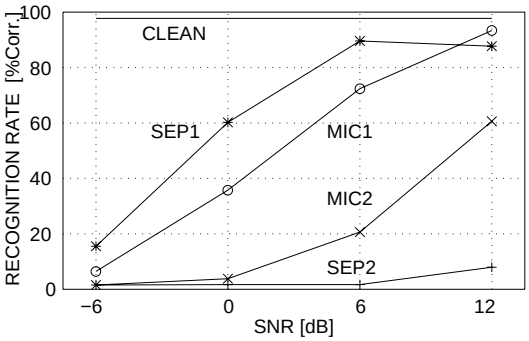

Figure 4: Word correct rate (% Corr) of HMM based continuous speech recognition results under various SNR conditions. The arrangement of microphones and loudspeakers are same as the DTW word discrimination experiment.

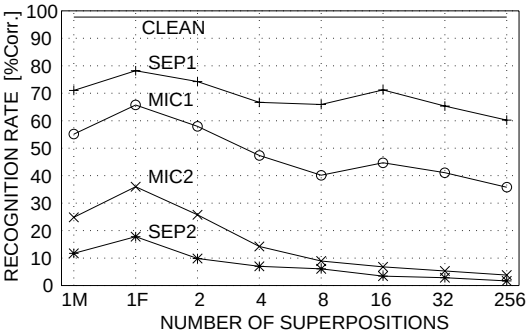

Figure 5: % Corr. of HMM based continuous speech recognition results plotted across the number of superpositions of interfering human-speech-like noise (HSLN). 1M and 1F indicate the case when an utterance of a male and female speaker is used as interfering speech, i.e. simple overlapping speech situation.

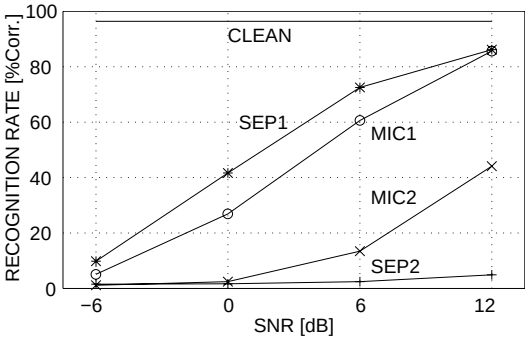

Figure 6: Word correct rate (% Corr) of HMM based continuous speech recognition results applied to the reverberant laboratory room conditions with interfering signals of various SNR.