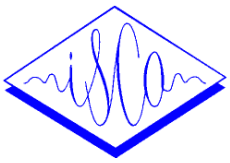

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

MULTICHANNEL SIGNAL SEPARATION FOR COCKTAIL PARTY SPEECH RECOGNITION: A DYNAMIC RECURRENT NETWORK

Seungjin CHOI [§], *Heonseok HONG* [§], *Hervé GLOTIN* [†], *Frédéric BERTHOMMIER* [†]

[§] Department of Electrical Engineering, Chungbuk National University, KOREA
{schoi,hshong}@lisa.chungbuk.ac.kr

[†] ICP, INPG, Grenoble CEDEX, FRANCE
{glotin, bertho}@icp.inpg.fr

ABSTRACT

This paper addresses the method of multichannel signal separation with its application to cocktail party speech recognition. First, we present a fundamental principle for multichannel signal separation which describes what spatial independence criterion results in. Second, for practical implementation of the signal separation filter, we consider a dynamic recurrent network and develop a new simple learning algorithm. The performance of the proposed method is evaluated in terms of word recognition error rate (WER). Experimental results show that our proposed method dramatically improves the word recognition performance in the case of two simultaneous speeches.

1. INTRODUCTION

Current automatic speech recognition systems can understand clean speeches well in relatively noiseless laboratory environments but its performance is severely degraded in the presence of loud noise or interfering speeches. Unlike automatic speech recognition systems, humans are able to well recognize mixtures of speeches produced by two simultaneous speakers. This ability is known as *binaural cocktail party effect*.

One promising approach to tackle the cocktail party speech recognition task might be multichannel signal separation (MSS) method which aims at estimating multiple unknown sources from the sensor signals observed at an array of microphones. In general, the propagation characteristics from acoustic sources to an array of microphones includes multipath effect and reverberation. Thus it is reasonable to model the propagation characteristics as an multivariate FIR filter. Then the task of multichannel signal separation is to recover sources from their convolutive mixtures. This task can be done in either time-domain [9, 2, 1, 3] or in frequency-domain [5]. Although frequency-domain approach enjoys the computational power of FFT, it requires the block processing that causes system latency. Moreover different permutation occurs in each frequency bin, which might results in severe degradation of performance without special case [7].

In this paper we stress out two issues (i.e., theory and implementation) in the time-domain approach to convolutive source separation. First we discuss some fundamental theory for convolutive source separation. Then we consider a dynamic recurrent network and develop a new learning algorithm for multichannel signal separation. The high performance of the method is confirmed by cocktail party speech recognition experiment.

2. MIXING MODEL

Let $\boldsymbol{x}(t) \in \mathbb{R}^n$ be a vector whose elements $\{x_i(t)\}$ are the signals measured at an array of microphones. Denote the source vector by $\boldsymbol{s}(t) = [s_1(t) \cdots s_n(t)]^T$. Then the ith sensor signal $x_i(t)$ is given by

$$x_i(t) = \sum_{j=1}^{n} \sum_{p} h_{ij,p} s_j(t-p), \tag{1}$$

where $\{h_{ij,p}\}$ is the room impulse response between the jth source and the ith microphone.

Define $H_{ij}(q^{-1}) = \sum_p h_{ij,p} q^{-p}$ where q^{-1} is the time-shift operator, i.e., $q^{-1}s_j(t) = s_j(t-1)$. Then we can rewrite (1) as

$$x_i(t) = \sum_{j=1}^{n} H_{ij}(q^{-1}) s_j(t). \tag{2}$$

In compact form, we have

$$\boldsymbol{x}(t) = \boldsymbol{H}(q^{-1})\boldsymbol{s}(t), \tag{3}$$

where $\boldsymbol{H}(q^{-1})$ is the polynomial matrix whose (i,j)-element is $H_{ij}(q^{-1})$.

3. SEPARATION PRINCIPLE

Throughout this paper the following terminology is used. Let us consider two stochastic sequences, $s_1(t)$ and $s_2(t)$. The collection of each sequence over the time is denoted by $\mathcal{S}_1$ and $\mathcal{S}_2$, i.e.,

$$\begin{aligned} \mathcal{S}_1 &= \{s_1(t)\ \forall t\} \\ \mathcal{S}_2 &= \{s_2(t)\ \forall t\}. \end{aligned} \tag{4}$$

Two stochastic sequences $s_1(t)$ and $s_2(t)$ are said to be *spatially independent* if any two finite subsets of $\mathcal{S}_1$ and $\mathcal{S}_2$ are statistically independent. A stochastic sequence $s_1(t)$ is

said to be *temporally independent* if any two finite disjoint subsets of $\mathcal{S}_1$ are statistically independent.

In this paper, we assume that source signals $\{s_i(t)\}$ are spatially independent. In addition, we assume that there exists an inverse of mixing filter $\boldsymbol{H}(q^{-1})$. Existence of the inverse or identifiability will not be addressed in this paper.

In the case of simple instantaneous mixing, the data model in (1) is simplified as

$$\boldsymbol{x}(t) = \boldsymbol{H}_0 \boldsymbol{s}(t). \tag{5}$$

In such a case, it is well known that independent component analysis (ICA) performs blind source separation. It can be easily justified by the Skitovich-Darmois theorem that is summarized below. The estimate of source vector, $\hat{\boldsymbol{s}}(t)$ has the form

$$\hat{\boldsymbol{s}}(t) = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{s}(t), \tag{6}$$

for some permutation matrix $\boldsymbol{P}$ and some nonsingular diagonal matrix $\boldsymbol{\Lambda}$.

Theorem 1 (Skitovich-Darmois) *Let $\{s_1, s_2, \ldots, s_n\}$ be a set of independent random variables. Consider two random variables x and y which are linear combinations of $\{s_i\}$,*

$$\begin{aligned} x &= a_1 s_1 + \cdots + a_n s_n, \\ y &= b_1 s_1 + \cdots + b_n s_n, \end{aligned} \tag{7}$$

where $\{a_i\}$ and $\{b_i\}$ are real constants. If x and y are statistically independent, then each variable s_i for which $a_i b_i \neq 0$ is Gaussian.

Liu and Luo [6] extended the Skitovich-Darmois theorem to the case of convolutive mixtures of i.i.d. sources. The key result in [6] is summarized in the following lemma and theorem which will be useful to develop the separation principle in the case of colored sources.

Lemma 1 (Liu-Luo) *Let $x_1(t)$ and $x_2(t)$ be two spatially independent stochastic sequences. Let $P(q^{-1})$ and $Q(q^{-1})$ be polynomials in q^{-1}. Then any pair of random variables from two stochastic sequences $y_1(t)$ and $y_2(t)$ defined by*

$$\begin{aligned} y_1(t) &= P(q^{-1})x_1(t), \\ y_2(t) &= Q(q^{-1})x_2(t), \end{aligned} \tag{8}$$

are independent.

Theorem 2 (Liu-Luo) *Let $\{e_1(t), e_2(t), \ldots, e_n(t)\}$ be a set of stochastic sequences which are temporally and spatially independent. Consider two random sequences $y_1(t)$ and $y_2(t)$ defined by*

$$y_1(t) = P_1(q^{-1})e_1(t) + \cdots + P_n(q^{-1})e_n(t) \tag{9}$$

$$y_2(t) = Q_1(q^{-1})e_1(t) + \cdots + Q_n(q^{-1})e_n(t), \tag{10}$$

where $\{P_i(q^{-1})\}$ and $\{Q_i(q^{-1})\}$ are polynomials in q^{-1}. If $y_1(t)$ and $y_2(t)$ are two spatially independent stochastic sequences, then each $e_i(t)$ for which $P_i(q^{-1})Q_i(q^{-1}) \neq 0$ is Gaussian.

Now we consider the case where sources are spatially independent but temporally colored, since speech signal has non-vanishing temporal correlation. Let us write the output of the separation filter, $\boldsymbol{y}(t)$ in terms of the global system $\boldsymbol{G}(q^{-1})$ (which combines the effect of mixing and demixing) and source vector $\boldsymbol{s}(t)$,

$$\boldsymbol{y}(t) = \boldsymbol{G}(q^{-1})\boldsymbol{s}(t). \tag{11}$$

Or

$$\begin{aligned} y_1(t) &= G_{11}(q^{-1})s_1(t) + \cdots + G_{1n}(q^{-1})s_n(t), \\ &\vdots \quad\quad \vdots \\ y_n(t) &= G_{n1}(q^{-1})s_1(t) + \cdots + G_{nn}(q^{-1})s_n(t), \end{aligned}$$

where $G_{ij}(q^{-1})$ is the (i,j)-element of $\boldsymbol{G}(q^{-1})$.

Theorem 3 (Fundamental Theorem) *Suppose that source signals $\{s_i(t)\}$ are regular processes with their innovations $\{e_i(t)\}$ being non-Gaussian. The innovations $\{e_i(t)\}$ are assumed to be spatially and temporally independent. If $y_i(t)$ and $y_j(t)$ are spatially independent for $i \neq j$, then $\boldsymbol{y}(t)$ has the following form*

$$\boldsymbol{y}(t) = \boldsymbol{P}\boldsymbol{C}(q^{-1})\boldsymbol{s}(t), \tag{12}$$

where

$$\boldsymbol{C}(q^{-1}) = diag\left\{C_1(q^{-1}), \cdots, C_n(q^{-1})\right\}, \tag{13}$$

This fundamental theorem is the direct consequence of Theorem 2. Let us assume source signals are regular processes. Then their innovations $\{e_i(t)\}$ exist. Hence each source $s_i(t)$ has the form

$$s_i(t) = A_i^{-1}(q^{-1})e_i(t). \tag{14}$$

In fact this is the auto-regressive (AR) model.

Combining (11) and (14), we can write the output $\boldsymbol{y}(t)$ as

$$\boldsymbol{y}(t) = \boldsymbol{G}(q^{-1})\left[A^{-1}(q^{-1}) \odot \boldsymbol{e}(t)\right], \tag{15}$$

where $\odot$ represents the Hadamard product (element-wise product) and $A^{-1}(q^{-1})$ is defined by

$$A^{-1}(q^{-1}) = \left[A_1^{-1}(q^{-1}) \cdots A_n^{-1}(q^{-1})\right]^T. \tag{16}$$

Let us consider the ith and jth elements of $\boldsymbol{y}(t)$,

$$\begin{aligned} y_i(t) &= \widetilde{G}_{i1}(q^{-1})e_1(t) + \cdots + \widetilde{G}_{in}(q^{-1})e_n(t), \\ y_j(t) &= \widetilde{G}_{j1}(q^{-1})e_1(t) + \cdots + \widetilde{G}_{jn}(q^{-1})e_n(t), \end{aligned} \tag{17}$$

where $\widetilde{G}_{ij}(q^{-1}) = G_{ij}(q^{-1})A_j^{-1}(q^{-1})$.

Note that the innovations $\{e_i(t)\}$ are spatially and temporally independent non-Gaussian stochastic sequences. If $y_i(t)$ and $y_j(t)$ are spatially independent, then it follows from Theorem 2 that $\widetilde{G}_{ik}(q^{-1})\widetilde{G}_{jk}(q^{-1}) = 0$ for $k = 1, \ldots, n$. This should be true for all $i \neq j$. Thus $y_i(t) = G_{ik}(q^{-1})s_k(t)$ (k is the index which explains the permutation), provided that $\boldsymbol{G}(q^{-1})\boldsymbol{A}^{-1}(q^{-1})$ is nonsingular for every z. Hence we are able to estimate the acoustic sources, each of which corresponds to the filtered version of acoustic source by minimizing the statistical dependence between $y_i(t-\tau_i)$ and $y_j(t-\tau_j)$ for all t, τ_i, τ_j, and $i \neq j$. Without further prior information, the suppression of echo is not possible blindly, however, cross-talking can be eliminated.

4. DEMIXING FILTER

Here we describe an implementation (demixing filter and associated learning algorithm) of the system which is able to eliminate cross-talking in the presence of simultaneous speeches. The theory that we explained in previous section states that the minimization of spatial dependence among the microphone signals results in the elimination of cross-talking. To this end we consider a dynamic recurrent network whose ith output $y_i(t)$ is described by

$$y_i(t) = x_i(t) + \sum_{p=0}^{L} \sum_{j \neq i} w_{ij,p} y_j(t-p). \quad (18)$$

In compact form, the output vector of the demixing network, $\boldsymbol{y}(t)$ is

$$\begin{aligned} \boldsymbol{y}(t) &= \boldsymbol{x}(t) + \sum_{p=0}^{L} \boldsymbol{W}_p \boldsymbol{y}(t-p) \\ &= [\boldsymbol{I} - \boldsymbol{W}_0]^{-1} \left\{ \boldsymbol{x}(t) + \sum_{p=1}^{L} \boldsymbol{W}_p \boldsymbol{y}(t-p) \right\}, \quad (19) \end{aligned}$$

where all diagonal elements of synaptic weight matrices $\{\boldsymbol{W}_p\}$ are zeros. In other words the network does not allow self-feedback connections. This recurrent network was already considered in [9, 2, 1]. In [2, 1], the learning algorithm for updating $\{\boldsymbol{W}_p\}$ has the form

$$\boldsymbol{W}_p(t+1) = \boldsymbol{W}_p(t) - \eta_t \varphi(\boldsymbol{y}(t)) \varphi^T(\boldsymbol{y}(t-p)), \quad (20)$$

where $\eta_t > 0$ is a learning rate. Note that only off-diagonal elements of $\{\boldsymbol{W}_p\}$ are adapted here. One can see that when the learning algorithm (20) achieves the convergence, the correlation between $\varphi_i(y_i(t))$ and $y_j(t-p)$ vanishes. Thus spatial dependence between $y_i(t)$ and $y_j(t)$ is minimized.

The nonlinear function $\varphi(\boldsymbol{y}) = [\varphi_1(y_1), \ldots, \varphi_n(y_n)]^T$ is a element-wise function. Popular choices of $\varphi_i(y_i(t))$ are signum function, $\varphi_i(y_i(t)) = \mathrm{sgn}(y_i(t))$ and hyperbolic tangent function, $\varphi_i(y_i(t)) = \tanh(y_i(t))$. This learning algorithm can be viewed as an extension of Jutten-Herault algorithm [4]. Although the algorithm (20) is successful in some simulations, it is rather slow and shows poor performance.

Now we derive our new learning algorithm for updating $\{\boldsymbol{W}_p\}$. To this send, we consider the loss function that is motivated from the minimization of mutual information. The loss function L that we consider here is

$$L = -\sum_{i=1}^{n} \log p_i(y_i), \quad (21)$$

where $\{p_i(\cdot)\}$ denote the probability density functions of sources or hypothesize densities for sources.

Let us define

$$\varphi_i(y_i) = -\frac{d \log p_i(y_i)}{dy_i}. \quad (22)$$

With this definition, the infinitesimal increment of the loss function is

$$\begin{aligned} dL &= \sum_{i=1}^{n} \varphi_i(y_i(t)) dy_i(t) \\ &= \varphi^T(\boldsymbol{y}(t)) d\boldsymbol{y}(t), \quad (23) \end{aligned}$$

where

$$d\boldsymbol{y}(t) = (\boldsymbol{I} - \boldsymbol{W}_0)^{-1} \left\{ d\boldsymbol{W}_p \boldsymbol{y}(t-p) + \sum_{k=1}^{L} \boldsymbol{W}_k d\boldsymbol{y}(t-k) \right\}.$$

We assume that the small change of $\boldsymbol{y}(t)$ is affected only by the small variation of the synaptic weight matrices $\{\boldsymbol{W}_p\}$ at time instant t. Here we the instantaneous gradient in our mind, so the second term in the bracket $\sum_{k=1}^{L} \boldsymbol{W}_k d\boldsymbol{y}(t-k)$ is neglected. As will be shown in experimental results, this approximation does not cause trouble. Then the approximated $d\boldsymbol{y}(t)$ is

$$d\boldsymbol{y}(t) = (\boldsymbol{I} - \boldsymbol{W}_0)^{-1} d\boldsymbol{W}_p \boldsymbol{y}(t-p). \quad (24)$$

The gradient descent method leads to the learning algorithm that has the form

$$\begin{aligned} \Delta \boldsymbol{W}_p(t) &= \boldsymbol{W}_p(t+1) - \boldsymbol{W}_p(t) \\ &= -\eta_t \frac{dL}{d\boldsymbol{W}_p} \\ &= -\eta_t [\boldsymbol{I} - \boldsymbol{W}_0(t)]^{-T} \varphi(\boldsymbol{y}(t)) \boldsymbol{y}^T(t-p) \quad (25) \end{aligned}$$

We can further simplify the algorithm (25) by adopting the pseudo-gradient,

$$\begin{aligned} \Delta \boldsymbol{W}_p(t) &= -\eta_t [\boldsymbol{I} - \boldsymbol{W}_0(t)] [\boldsymbol{I} - \boldsymbol{W}_0(t)]^T \frac{dL}{d\boldsymbol{W}_p} \\ &= -\eta_t [\boldsymbol{I} - \boldsymbol{W}_0(t)] \varphi(\boldsymbol{y}(t)) \boldsymbol{y}^T(t-p). \quad (26) \end{aligned}$$

Since $(\boldsymbol{I} - \boldsymbol{W}_0)(\boldsymbol{I} - \boldsymbol{W}_0)^T$ is positive definite, the gradient direction is not changed.

Remarks:

- The algorithm (20) can be viewed as an approximated version of (25) by eliminating the term $(\boldsymbol{I} - \boldsymbol{W}_0)^{-T}$.
- In our computer simulations, we found that the algorithms (25) and (26) converged to a solution much faster than the algorithm (20).
- In our experiment, the algorithms (25) and (26) showed similar performance and convergence speed. However, the algorithm (26) has less computational complexity compared to (25) because it does not require the matrix inversion.

5. EXPERIMENTAL RESULTS

The performance of the proposed method here was evaluated using the stereo database ST-NB95 [8]. The ST-NB95 was built from Numbers95 (NB95) in order (1) to spatialize the signal of NB95 in azimuth (2) to introduce a minimal distortion of the original signal and (3) to mix the signals of NB95 with a relative level controlled well. The configuration of microphones and loudspeakers is shown in Figure 1.

In the recurrent network for multichannel signal separation, we used the length of delay, $L = 250$ and the learning rate $\eta_t = 10^{-7}$.

We used single state HMM/ANN context independent phone models. Multilayer perceptron (MLP) was trained by 3590 utterances of the monophonic NB95 database. There

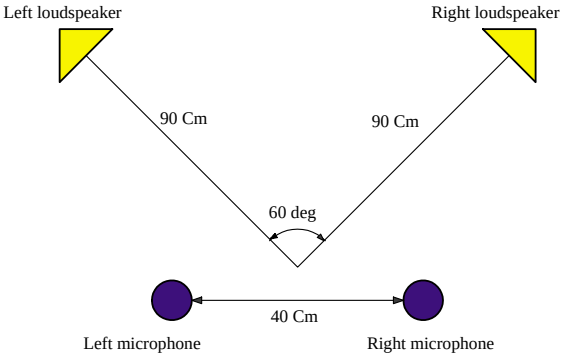

Figure 1: The configuration of loudspeakers and microphones.

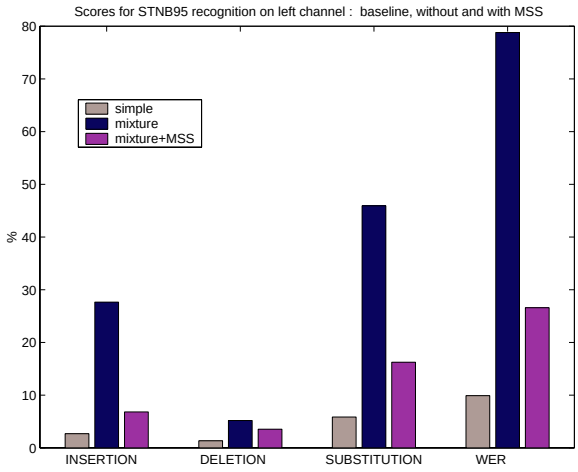

Figure 2: JRASTA WER results by insertion, deletion, substitution, and WER: (1) STNB95 on speaker (simple); STmixed two speakers (mixtures); STmixed+MSS (mixture+MSS).

was no adaptation to generate local probabilities for HMMs. Because the NB95 database consists of telephone signals, their low frequency components were filtered out. Then signals' frequencies are between [115, 3769] Hz. The JRASTA-PLP was used for the common feature's preprocessing. The MLP had 11 lpc order analysis, 12 cepstral coefficient and energy. We extracted delta and delta delta of all previous parameters. This set of parameters was taken for 9 successive frames of 25 ms shifted of 12.5 ms. Then MLP had a total of 351 input units (9 * 3 * (12+1)), 1750 hidden units, and 27 output posteriors' phoneme class. Test set had 613 sentences for each source.

Baseline score on clean original test set is 8 $\pm$ 0.7 % WER. Despite the noise introduced during the stereo recording (mostly due to the microphone transfer function), the baseline scores for simple source are quite similar: 9.9 % WER for left source; 8 % for right source. This might be due to the good performance of JRASTA that aims to remove additive and convolutive noise. With left microphone and simultaneous speech (cocktail party simulation) we had : 78.8 %WER with the left reference. The proposed method gave for left 26.6 and right source 26.2 %WER, so in mean 26.4 $\pm$ 1.2 % WER. This result is shown in Figure 2.

6. CONCLUSIONS

We have presented a fundamental theorem for multichannel signal separation which described what spatial dependence minimization criterion results in when sources are spatially independent but temporally correlated. For practical implementation, a dynamic recurrent network was considered and a new simple learning algorithm was developed. The method was applied to the task of cocktail party speech recognition. Experimental results showed that the method improves the word recognition performance dramatically in the case of two simultaneous speeches.

7. ACKNOWLEDGMENTS

IDIAP-CH has supported some of the recognition tasks. This work was supported in part by Korean Ministry of Science and Technology under Brain Science and Engineering Research Program and by the EEC project LTR RESPITE (Task 2.1).

8. REFERENCES

[1] N. Charkani and Y. Deville. Self-adaptive separation of convolutively mixed signals with a recursive structure - part I: Stability analysis and optimization of asymptotic behaviour. *Signal Processing*, 73(3):255–266, 1999.

[2] S. Choi and A. Cichocki. Adaptive blind separation of speech signals: Cocktail party problem. In *Proc. Int. Conf. Speech Processing*, pages 617–622, 1997.

[3] S. Choi, Y. Lyu, F. Berthommier, H. Glotin, and A. Cichocki. Blind separation of delayed and superimposed acoustic sources: Learning algorithms and experimetal study. In *Proc. Int. Conf. Speech Processing*, pages 109–113, 1999.

[4] C. Jutten and J. Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[5] R. H. Lambert. *Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures.* PhD thesis, University of Southern California, May 1996.

[6] R. Liu and H. Luo. Direct blind separation of independent non-gaussian signals with dynamic channels. In *Proc. Int. Workshop on Cellular Neural Networks and their Applications*, pages 34–37, London, UK, 1998.

[7] L. C. Parra and C. Spence. Convolutive blind source separation of non-stationary sources. *IEEE Trans. Speech and Audio Processing*, to appear.

[8] E. Tessier, F. Berthommier, H. Glotin, and S. Choi. A model of source segregation using the localization cue for robust cocktail-party speech recognition. In *Proc. Int. Conf. Speech Processing*, pages 97–102, 1999.

[9] K. Torkkola. Blind separation of convolved sources based on information maximization. In *Proc. IEEE Workshop Neural Networks for Signal Processing*, pages 423–432, 1996.