Adaptive Blind Separation of Speech Signals: Cocktail Party Problem

Seungjin CHOI and Andrzej CICHOCKI
Lab for Artificial Brain Systems
Frontier Research Program, RIKEN
2-1 Hirosawa, Wako-shi
Saitama 351-01, Japan
{schoi, cia}@lion.riken.go.jp

Abstract

In this paper, we present an on-line adaptive scheme for blind separation of speech signals from their convolutive mixtures. This problem is often referred as cocktail party problem. When multiple speakers speak simultaneously in tele-conferencing studio, we need to separate out each speaker from their mixtures. If mixtures are assumed as instantaneous mixtures, then it becomes standard blind separation problem [19, 13, 5, 1, 6, 10, 11]. In practical situations, however, the observed signals from microphones are usually convolutive mixtures due to the propagating source signals through the dynamic medium and parasitic effects like multiple echoes and reverberation. Thus, the received signals are weighted sum of mixed and delayed components. In other words, the received signals at each microphone are the convolutive mixtures of speech signals. We present two slightly different neural networks, i.e., a dynamic recurrent network and a feedforward-feedback network, which can separate the convolutive mixtures of speech signals without any knowledge of the propagation media. Both theoretical and computer simulation results are provided.

1 Introduction

The problem of blind source separation is to recover the statistically independent source signals from their instantaneous mixtures. Since Jutten and Herault [19] proposed a neural computational approach to source separation (using a linear feedback neural network) this problem has drawn many attractions from a variety of applications such as digital communication, image processing, array signal processing, sensory processing, and biomedical applications. Jutten and Herault approach has been further developed by Cichocki et al. [13] by employing self-connections for normalization. Bell and Sejnowski [5] proposed an information maximization approach, which gave a bright insight. This information-theoretic approach to source separation was further clarified by Amari et al. [1]. Amari et al. [1] have introduced a natural gradient [3] and provides a robust source separation algorithm. Equivariant source separation algorithm [6] has been proposed by Cardoso and Laheld. The source separation criteria based on only finite order of cross-cumulants has been proposed by Choi and Lin [10], Choi and Cichocki [11]. It has been recently shown [7] that information maximization approach coincides with maximum likelihood in source separation.

However, source separation problem has been restricted to model a channel as a linear memoryless system. In practical applications, propagation media result in multiple echoes and reverberation. Thus the received signals from an array of sensors are convolutive mixtures of source signals. In this paper, we address time-domain approach to separate convolutive mixtures of speech signals. First, we formulate the problem of speech signal separation of their convolutive mixtures. Second, from an information theoretic approach, we derive adaptive algorithms which can eliminate the spatial dependence by using a dynamic recurrent neural network and a hybrid network combining a linear memoryless feedforward network and a dynamic recurrent network.

2 Problem Formulation

The problem of speech signal separation can be formulated as follows. An $m$ dimensional vector of re-
received signals $x(t)$ is assumed to be generated from an $n$ dimensional vector of independent sources $s(t)$ (each speaker's contribution) using the multi-variate linear time invariant filter, i.e.,

$$
x(t) = \sum_{i=0}^{M} H_i s(t - i) + n(t),
$$

$$
= H(z) s(t) + n(t),
$$

(1)

where $H(z) = \sum_{i=0}^{M} h_i z^{-i}$ ($H(z)$ is an $m \times n$ polynomial matrix and $z^{-i}$ is delay operator such that $z^{-i} s(t) = s(t - i)$) is the channel transfer function and $n(t)$ is an additive white Gaussian noise. It is assumed that source signals $s(t)$ are spatially independent. If each component of source signals $s(t)$ are temporally uncorrelated, i.e., $E[s_k(t) s_j(t - \tau)] = 0$, $\forall \tau \neq 0$ besides the spatial independence, we can recover the source signals $s(t)$ up to their scaling, ordering, and delay ambiguity [15, 14, 2, 18, 12]. It has been shown [12] that spatio-temporal decorrelation can decorrelate the channels up to the instantaneous mixtures which can be solved by source separation techniques.

In contrast to multichannel blind deconvolution of LTI system excited by spatially independent and temporally uncorrelated source signals, the separation of speech signals from their convolutive mixtures are more complicated problem. Due to the statistical dependence of speech signals in temporal domain, it is meaningless to recover the exact waveform of speech signals. What we can do is to extract the contribution of each speech signal from their mixtures. In other words, each recovered signal is still unknown filtering version of one of original speech signals. Thus the task is to build a second filter $W(z)$ such that the output $y(t) = W(z) x(t) = P \Delta D(z) s(t)$, where $P$ is a permutation matrix, $\Delta$ is a nonsingular diagonal matrix, and $D(z) = \text{diag}(D_1(z), \ldots, D_n(z))$. ($D_i(z)$ is a polynomial in terms of a delay operator $z^{-i}$). As past work, Platt and Faggini [26], Nguyen Thi and Jutten [20] have extended Jutten's network [19] to the case of convolutive mixture. Cicchetti et al [14] have extended their original network. The separation of signals based on output deconvolution has been developed by [16, 29, 21]; Bell and Sejnowski's informax principle [3] has been extended to speech signal separation [27] and to echo cancellation [28] by Torkkola. Frequency-domain approach was proposed by Lambert and Nikias [22]. Infomax in frequency domain has been developed by Lee et al. [23, 24].

3 Neural Network Models and the Algorithms

From an information-theoretic approach, we describe a local algorithm for speech signal separation using a dynamic recurrent network (see Figure 1) and a hybrid network (see Figure 2).

3.1 A Dynamic Recurrent Neural Network

We assume that the number of source signals, $n$, are known. The same number of sensors are used, i.e., $m = n$. Consider a dynamic recurrent neural network shown in Figure 1, described by

$$
y_k(t) = x_k(t) + \sum_{i=0}^{L} \sum_{j=1}^{n} w_{ijk}(t) y_j(t - k),
$$

(2)

where $y_k(t)$ is the output of the network, $w_{ijk}(t)$ is the synaptic weight between $y_i(t)$ and $y_j(t - k)$, and $x_k(t)$ is the input to the network. Or in matrix form

$$
y(t) = x(t) + \sum_{k=0}^{L} W_k(t) y(t - k),
$$

(3)

where $W_k(t) \in \mathbb{R}^{n \times n}$ for $k = 0, \ldots, L$ are connection matrices with all diagonal elements being zero (there are only cross-connections). Let $w_{ijk}(t)$ denote the $(i,j)$th element of the matrix $W_k(t)$. Redundancy reduction or minimization of mutual information among output of the network leads to the following loss function,

$$
l(W_z(t)) = E\{ - \sum_{i=1}^{n} \log p_i(y_k(t)) \},
$$

(4)

where $p_i(y_k(t))$ is the probability density function (PDF) of $y_k(t)$. In contrast to a feedforward network, it is not necessary to have normalization constraint in (4). This redundancy reduction principle coincides with maximum likelihood here. We now develop an algorithm for minimizing (4). Let us define

$$
f_k(y_k(k)) = - \frac{d \log p_k(y_k(k))}{dy_k(k)},
$$

(5)

By stochastic gradient descent approach, the updating equation for $w_{ijk}(t)$ is

$$
w_{ijk}(t + 1) = w_{ijk}(t) - \eta \frac{dl(W(z,t))}{dw_{ijk}(t)}
$$

$$
= w_{ijk}(t) - \eta f_k(y_k(t)) y_j(t - k),
$$

(6)
Note that when $f_d(y_i(t)) = y_i(t)$ (this is the case where $y_i(t)$ is Gaussian), (6) becomes the decorrelation algorithm which was proposed by [16]. Depending on the probability distribution of speech signals, we can choose a good nonlinear function $f_d(y_i(t))$. Recently Charni and Deville [9] have reported that the PDF of speech signals are close to Laplacian distribution and the choice, $f_d(y_i(t)) = \text{sign}(y_i(t))$ outperformed linear learning $f_d(y_i(t)) = y_i(t)$ or $f_d(y_i(t)) = \tanh(\gamma y_i(t))$ [5, 27].

\begin{equation}
    y_i(t) = x_i(t) + \sum_{j \neq i} w_{ij0} x_j(t) + \sum_{k=1}^{L} \sum_{j \neq i} w_{ijk}(t) y_j(t-k),
\end{equation}

where $y_i(t)$ is the output of the $i$th unit, $w_{ij0}(t)$ is the synaptic weight between $y_i(t)$ and $x_j(t)$, $w_{ijk}(t)$ is the synaptic weight between $y_i(t)$ and $y_j(t-k)$, and $x_j(t)$ is the input to the network. Or in matrix form

\begin{equation}
    y(t) = W_0(t)x(t) + \sum_{k=1}^{L} W_k(t)y(t-k),
\end{equation}

where $W_0(t)$ is a synaptic weight matrix consisting of connection strength between $y(t)$ and $x(t)$ with its diagonal elements being unity; $W_k(t) \in \mathbb{R}^{m \times n}$ for $k = 1, \ldots, L$ are connection matrices of feedback part of the network with all diagonal elements being zero. Let $w_{ijk}(t)$ denote the $(i,j)$th element of the matrix $W_k(t)$. In similar fashion, we can derive the following learning algorithm for $w_{ijk}(t)$ to minimize the loss function given in (4) by stochastic gradient descent approach, described by

\begin{equation}
    w_{ijk}(t+1) = w_{ijk}(t) - \eta_i f_i(y_i(t)) x_j(t-k),
\end{equation}

\begin{equation}
    w_{ijk}(t+1) = w_{ijk}(t) - \eta_i f_i(y_i(t)) y_j(t-k),
\end{equation}

for $i \neq j$ and $k \neq 0$.

\begin{figure}[h]
    \centering
    \includegraphics[width=\textwidth]{figure1.png}
    \caption{A dynamic recurrent network for blind separation of speech signals.}
\end{figure}

\begin{figure}[h]
    \centering
    \includegraphics[width=\textwidth]{figure2.png}
    \caption{A hybrid network for blind separation of speech signals.}
\end{figure}

\section{Computer Simulations}

We have tested the performance of two proposed networks with associated learning algorithms (6) and (9), (10). Two different digitized speech signals $s(t)$ as shown in Figure 3 were used in this simulation. We have tested proposed algorithms with three different $f_d(y_i(t))$: 1) $f_d(y_i(t)) = y_i(t)$; 2) $f_d(y_i(t)) = \text{tanh}(3y_i(t))$; 3) $f_d(y_i(t)) = \text{sign}(y_i(t))$. The received signals (see Figure 4) collected from two different microphones, i.e., $x(t)$ were generated by

\begin{equation}
    x(t) = H(z)s(t),
\end{equation}

where

\begin{align*}
    H_{11}(z) &= 0.818 + 0.937z^{-1} + 0.846z^{-2} - 0.824z^{-3} \\
    &- 0.327z^{-4} + 0.894z^{-5} + 0.881z^{-10} - 0.149z^{-20}, \\
    H_{12}(z) &= -0.906 + 2.57z^{-1} + 0.741z^{-2} - 0.463z^{-3} \\
    &- 0.774z^{-4} - 2.59z^{-5} + 0.664z^{-10} - 0.214z^{-20}, \\
    H_{21}(z) &= -0.333 - 0.831z^{-1} + 0.671z^{-2} - 0.756z^{-3} \\
    &- 2.64z^{-4} - 4.58z^{-5} - 1.4z^{-10} + 0.998z^{-20}, \\
    H_{22}(z) &= 0.342 - 0.7z^{-1} - 0.933z^{-2} - 0.468z^{-3} \\
    &+ 0.701z^{-4} - 2.24z^{-5} - 0.994z^{-10} - 0.672z^{-20}.
\end{align*}

As a performance measure, we have used a signal to noise ratio improvement [9], defined as

\begin{equation}
    SNRI_t = 10 \log_{10} \frac{E(x_k(t) - s_k(t))^2}{E(y_k(t) - s_k(t))^2},
\end{equation}

where

\begin{align*}
    H_{11}(z) &= 0.818 + 0.937z^{-1} + 0.846z^{-2} - 0.824z^{-3} \\
    &- 0.327z^{-4} + 0.894z^{-5} + 0.881z^{-10} - 0.149z^{-20}, \\
    H_{12}(z) &= -0.906 + 2.57z^{-1} + 0.741z^{-2} - 0.463z^{-3} \\
    &- 0.774z^{-4} - 2.59z^{-5} + 0.664z^{-10} - 0.214z^{-20}, \\
    H_{21}(z) &= -0.333 - 0.831z^{-1} + 0.671z^{-2} - 0.756z^{-3} \\
    &- 2.64z^{-4} - 4.58z^{-5} - 1.4z^{-10} + 0.998z^{-20}, \\
    H_{22}(z) &= 0.342 - 0.7z^{-1} - 0.933z^{-2} - 0.468z^{-3} \\
    &+ 0.701z^{-4} - 2.24z^{-5} - 0.994z^{-10} - 0.672z^{-20}.
\end{align*}

As a performance measure, we have used a signal to noise ratio improvement [9], defined as

\begin{equation}
    SNRI_t = 10 \log_{10} \frac{E(x_k(t) - s_k(t))^2}{E(y_k(t) - s_k(t))^2},
\end{equation}
The recovered signals using a dynamic recurrent network with three different $f_d(y_k(t))$ are shown in Figure 5, 6, and 7. The simulation results using a hybrid network are shown in Figure 8, 9, and 10. It can be observed that $f_d(y_k(t)) = \text{sign}(y_k(t))$ and $f_d(y_k(t)) = \tanh(3y_k(t))$ performs better than $f_d(y_k(t)) = y_k(t)$.

5 Conclusion

We have proposed neural network architectures and associated learning algorithms for blind separation of convolutive and linearly mixed speech signals. The learning algorithms which are derived from information-theoretic approach are relatively simple local rules which can be easily implemented in VLSI technology. Preliminary computer simulation experiments confirm the validity of the developed approach. The method allows us to separate the speech signals from their convolutive mixtures corrupted by environmental noise or even another speech signals. The proposed model of separation of convolutive mixtures can be used as a preprocessing module for speech recognition systems. We focus our current research works on more sophisticated and realistic models like ARMA filters and the case with the number of microphones larger than the number of source (speech) signals.

![Figure 3. Two digitized original speech signals](image1.png)

![Figure 4. Two mixtures of speech signals](image2.png)

![Figure 5. Two recovered signals using the dynamic recurrent network in Figure 1, $f_d(y_k(t)) = y_k(t)$, SNR$R_1 = 2.88$dB and SNR$R_2 = 3.45$dB](image3.png)

![Figure 6. Two recovered signals using the dynamic recurrent network in Figure 1, $f_d(y_k(t)) = \tanh(3y_k(t))$, SNR$R_1 = 3.49$dB and SNR$R_2 = 6.24$dB.](image4.png)

References


Figure 7. Two recovered signals using the dynamic recurrent network in Figure 1, \( f_i(y_k(t)) = \text{sign}(y_i(t)) \), \( \text{SNRI}_1 = 3.60\text{dB} \) and \( \text{SNRI}_2 = 6.32\text{dB} \).

Figure 8. Two recovered signals using the hybrid network in Figure 2, \( f_i(y_k(t)) = y_i(t) \), \( \text{SNRI}_1 = 2.91\text{dB} \) and \( \text{SNRI}_2 = 3.43\text{dB} \).

Figure 9. Two recovered signals using the hybrid network in Figure 2, \( f_i(y_k(t)) = \tanh(3y_i(t)) \), \( \text{SNRI}_1 = 3.64\text{dB} \) and \( \text{SNRI}_2 = 5.54\text{dB} \).

Figure 10. Two recovered signals using the hybrid network in Figure 2, \( f_i(y_k(t)) = \text{sign}(y_i(t)) \), \( \text{SNRI}_1 = 3.62\text{dB} \) and \( \text{SNRI}_2 = 5.57\text{dB} \).


