Speech Communication 92 (2017) 132–141

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

A permutation algorithm based on dynamic time warping in speech frequency-domain blind source separation

Zhao Lv[a,b], Bei-bei Zhang[a,b], Xiao-pei Wu[a,b,*], Chao Zhang[a], Bang-yan Zhou[a,b]

[a] The Key Laboratory of Intelligent Computing & Signal Processing, Anhui University, Hefei 230039, China
[b] Co-Innovation Center for Information Supply & Assurance Technology, Anhui University, Hefei 230601, China

ARTICLE INFO

Article history:
Received 4 February 2016
Revised 20 June 2017
Accepted 22 June 2017
Available online 1 July 2017

Keywords:
Dynamic time warping (DTW)
Blind source separation (BSS)
Independent component analysis (ICA)
Frequency-domain (FD)
Time-domain (TD)
Permutation

ABSTRACT

Frequency-Domain Blind Source Separation (FD-BSS) is an efficient way to analyze convolutive mixed speech. To improve the quality of the separated speech, a permutation algorithm based on *Dynamic Time Warping* (DTW) is proposed in this paper. Because signals in adjacent frequency bins have high similarity, DTW technology is used to compare them and generate adjustment matrices to solve the permutation ambiguity. Our approach is evaluated through simulated and practical experiments. Using *Signal to Distortion Ratio* (SDR), *Signal to Interference Ratio* (SIR), *Signal to Artifacts Ratio* (SAR), and *Perceptual Estimation of the Speech Quality* (PESQ) for measurements. To examine the quality of the separated speech in a practical acoustic environment, we adopt the accuracy ratio of *Automatic Speech Recognition* (ASR). In the experiments, we compare our approach with other classical permutation criteria such as K-L divergence distance, envelope correlation and higher-order statistics. The experimental results show that the proposed algorithm performs permutation alignment more accurately and improves the acoustic quality of separation.

1. Introduction

Blind Source Separation (BSS) is an effective approach for estimating the original source signals using only the information in the mixed signals observed by each sensor, which consist of mixtures of the original signals (Xie et al., 2012; Vincent et al., 2006). This technique can acquire independent source signals without any prior knowledge of either the source signals or the mixing matrix (Thi and Jutten, 1995), and it is applicable to both speech recognition and high-quality telecommunication systems. BSS is also an important method for auditory scene analysis in convolutive environments. The fundamental principle of BSS is shown in Fig. 1 (Haykin, 2000).

As shown in Fig. 1, the observation vector $\mathbf{x} = [x_1(t), x_2(t), \cdots, x_N(t)]^T$ can be modeled as

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

where “instantaneous mixing matrix” $\mathbf{A}$ is invertible and $\mathbf{s} = [s_1(t), s_2(t), \cdots, s_N(t)]^T$ denotes an independent source. This mixing is termed “instantaneous” because the observation vectors at the current time depend on the sources at the same, but no earlier, time point. The goal of BSS is to find a “separating matrix”, $\mathbf{W}$, such that

$$\hat{\mathbf{s}} = \mathbf{W} \cdot \mathbf{x} \tag{2}$$

is an optimal estimation of the source signal $\mathbf{s}$.

Several algorithms have been proposed to achieve BSS (Haykin, 2000, Bell. and Sejnowski., 1995; Haykin, 2000, Deville. and Duarte., 2015). Among them, *Independent Component Analysis* (ICA) (Comon, 1994) plays an important role in speech BSS. Research has proven that the ICA method can separate “clean” speech and noise effectively in instantaneous mixing conditions. However, the sound may become distorted by influences such as time-delay and reverberation imposed in the real acoustic environment when propagated in a medium. This means that signals are spatially filtered, i.e., they are convolutively mixed before arriving at the receiver (Chen et al., 2015; Jan et al., 2011). Convolutive mixing signals are more difficult to separate than instantaneous mixing signals because of the more complicated conditions under which they are generated.

Two methods based on ICA have been proposed. The first is to construct a *Time-Domain* (TD) deconvolutive filter and directly employ an instantaneous ICA algorithm to separate the mixed signals (Gao et al., 2013; Mahajan and Betrabet, 2015; Buchner et al., 2005). However, the high computation and parameter adjusting load required for the deconvolutive filter may present practical limitations owing to the excessive length of the *Finite Im-*

[*] Corresponding author at: Co-Innovation Center for Information Supply & Assurance Technology, Anhui University, Hefei 230601, China.
E-mail addresses: wxp2001@ahu.edu.cn, iiphci_ahu@163.com (X.-p. Wu).

http://dx.doi.org/10.1016/j.specom.2017.06.007
0167-6393/© 2017 Elsevier B.V. All rights reserved.

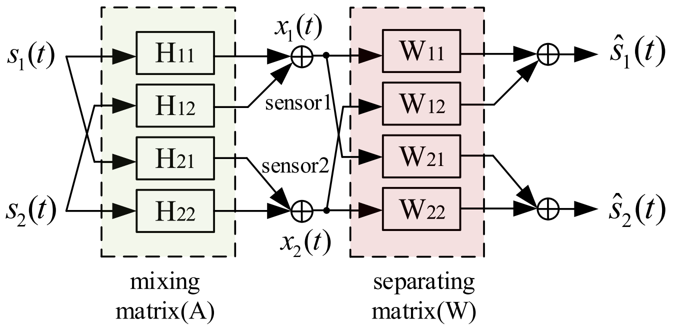

Fig. 1. Fundamental principle of BSS.

pulse Response (FIR) filter. The second method is to transform the mixed signals from the TD to the *Frequency-Domain* (FD) using a *Short Time Fourier Transform* (STFT) (Koldovsky and Tichavsky, 2011). In the FD, signals from the convolutive mixture in the TD are transformed into instantaneous mixtures; consequently, a complex-valued ICA method can be applied to separate the signals of the instantaneous mixture in each frequency bin (Wang et al., 2005; Koya et al., 2011; Fu et al., 2015; Hyvärinen, 2013).

Permutation ambiguity is one of the inherent indeterminacies in the ICA algorithm (Wang et al., 2004). For the instantaneous ICA model, the relationship between the source signal **s** and the estimated signals $\hat{\mathbf{s}}$ can be inferred from Eq. (1) and Eq. (2) as follows:

$$\hat{\mathbf{s}} = \mathbf{W}\cdot\mathbf{x} = \mathbf{W}\cdot\mathbf{A}\cdot\mathbf{s} = \mathbf{D}\cdot\mathbf{s}. \tag{3}$$

Considering that the input sources are independent, matrix **D** exists only with a non-zero element in each column and line, i.e.,

$$\mathbf{D} = \mathbf{P}\boldsymbol{\Lambda}, \tag{4}$$

where **P** and **Λ** are the transport matrix and the diagonal matrix, respectively. Matrix **P** will lead to output permutation indeterminacy, namely, the channel orders of the estimated output signals are not fixed. In the instantaneous ICA model, permutation ambiguity is present only in the output channel indeterminacy because a single separation operation is processed. In contrast, in a *Frequency Domain BSS* (FD-BSS) model, the ICA algorithm is executed in each frequency bin independently, and permutation ambiguity becomes a serious problem. The order of recovered signals in each frequency bin must be aligned so that the reconstructed signals in the TD will not be mixed with other sources. Various methods have been proposed to solve this problem. Smaragdis smoothed the separation matrices, but this method is ineffective when the mixing filter is too long (Smaragdis, 1998; Schobben and Sommen, 2002). Asano detected the correlation between envelopes of signals from adjacent frequency bins in (Asano et al., 2003), but the approach is insufficiently robust because only some envelope information can be extracted to match and, consequently, some detail information may be ignored. Ikram proposed a method to estimate the *Direction of Arrival* (DOA) of signals to overcome the permutation ambiguity in (Ikram and Morgan, 2002), however, this method requires the sources to be far from the microphones, and the signal propagation is a planar wave front. In that case, DOA can estimate the source's direction from the separation matrix, but for some frequency bins, especially at low frequencies, the direction of the sources cannot be estimated (Sawada et al., 2004).

In this paper, a new permutation algorithm based on *Dynamic Time Warping* (DTW) is presented and used to solve the permutation ambiguity. First, the feature vectors for the first frequency bin signals are extracted and saved as reference templates. When the next adjacent frequency bin signals are input, the proposed algorithm compares the similarities of the featured input signals with the reference templates and then outputs an adjustment matrix, which is determined by the minimum distortion between two adjacent frequency bin signals. In this way, the signals of all frequency bins are processed completely to obtain the reconstructed signals in the TD. To improve the algorithm's reliability, it uses *linear predictive coding cepstrum* (LPCC) as the feature parameters. The experimental results show that the proposed algorithm improves the permutation performance and the quality of separated speech.

The remainder of this paper is organized as follows: Section 2 introduces the FD-BSS algorithm based on ICA. Section 3 describes how to use DTW algorithm to solve the permutation ambiguity in detail. By first depicting the fact that signals in adjacent frequency bins have high similarity (meaning the DTW algorithm can be used to solve the problem of permutation ambiguity) and second, by introducing the fundamental principle of the DTW algorithm. Finally, it illustrates the process of the proposed algorithm using a case of the *Two-Input-Two-Output* (TITO) convolutive mixing model. Simulated and practical acoustic environment experiments are described in Section 4, and Section 5 presents conclusions.

2. Frequency-domain BSS based on the ICA algorithm

In a practical acoustic environment, the existence of reverberation and time-delay creates a convolutive mixing effect between the observed signal x, and the speech sources s. This process can be expressed as described in (Yousefian et al., 2015),

$$x_i(t) = \sum_{j=1}^{N} a_{ij}(t) * s_j(t) = \sum_{j=1}^{N}\sum_{k=0}^{P-1} a_{ij}(k)s_j(t-k) \quad i = 1, 2, \cdots, N, \tag{5}$$

where P is the order of the mixing filter, and a_{ij} denotes the impulse response from the j^{th} source to the i^{th} speech sensor. The FD-BSS algorithm acts to convert the TD signals, $x_i(t)$, into the FD signals, $X_i(f)$, using STFT:

$$X_i(f) = \sum_{j=1}^{N} A_{ij}(f)\cdot S_i(f) \qquad i = 1, 2, \cdots, N. \tag{6}$$

In Eq. (6), an instantaneous ICA algorithm can be used directly to separate the mixing speech signals into different frequency bins. To achieve this transformation, a windowing and STFT process is applied to the original mixing signals

$$X_i(f_l, \tau) = \sum_{t=0}^{L-1} x_i(t) win(t-\tau)\exp(-j2\pi f_l t),\ i = 1, 2, \cdots, N, \tag{7}$$

where $l = 0, 1, ..., L-1,\ f_l = (l/L)f_s$ represents the l_{th} frequency bin, $win(t)$ denotes a windowing function, τ means the position of the window function, and f_s is the sampling frequency. For simplicity, we illustrate the process using a case of a TITO convolutive mixture model.

An $L \times M$ complex value matrix stemming from the observation signals $x_i(t)$ will be obtained after the STFT operation:

$$STFT(x_i) = \begin{bmatrix} X_i(f_0,\tau_0) & X_i(f_0,\tau_1) & \cdots & X_i(f_0,\tau_{M-1}) \\ X_i(f_1,\tau_0) & X_i(f_1,\tau_1) & \cdots & X_i(f_1,\tau_{M-1}) \\ \vdots & \vdots & \vdots & \vdots \\ X_i(f_{L-1},\tau_0) & X_i(f_{L-1},\tau_1) & \cdots & X_i(f_{L-1},\tau_{M-1}) \end{bmatrix} \quad i = 1, 2, \tag{8}$$

where M denotes the frame movement time and x_i is the i^{th} channel of observation signals. Then, two observation channel vectors