Signal Processing 96 (2014) 261–265

Contents lists available at ScienceDirect

Signal Processing

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

Fast communication

Extension of DUET to single-channel mixing model and separability analysis

N. Tengtrairat, W.L. Woo *

School of Electrical and Electronic Engineering, Newcastle University, England, United Kingdom

ARTICLE INFO

Article history:
Received 27 April 2013
Received in revised form
21 August 2013
Accepted 29 August 2013
Available online 16 September 2013

Keywords:
Single-channel mixture
DUET
Autoregressive process
Maximum likelihood estimation

ABSTRACT

In this paper, the DUET binaural model is extended to the single-channel mixing model where only one microphone is available for recording. A novel "artificial stereo" mixing model is proposed to create a synthetic stereo signal by weighting and time-shifting the original single-channel mixture. Separability analysis of the proposed model has also been derived to verify that the artificial stereo mixture is separable. This work, therefore, relaxes the underdetermined ill-conditions associated with monaural source separation and path the way for binaural source separation approaches to solve monaural mixture.

1. Introduction

The Degenerate Unmixing Estimation Technique (DUET) [1] has been proposed as a separating method using binary time–frequency (TF) masks. A major advantage of DUET is that the estimates from two channels are combined inherently as part of the clustering process. The DUET algorithm has been demonstrated to recover the underlying sparse sources given two anechoic mixtures in the TF domain. However, the DUET algorithm has been practically handicapped to separate signals when only one recording channel is available which leads to the problem of single channel signal separation [2,3]. In this paper, a new method for formulating an artificial binaural mixture using a monaural mixture is presented. To the best knowledge of the authors, the proposed method is the first artificial stereo mixture based on the autoregressive (AR) process.

This paper is organized as follows: Section 2 proposes an artificial stereo mixing model. Section 3 presents the assumptions. Next, the artificial stereo mixing model in time–frequency representation is given in Section 4. The separability of the proposed model is analyzed in Section 5. Finally, Section 6 concludes the paper.

2. Artificial stereo mixing model

In this paper, we consider the case of a mixture of two sources in time domain which can be expressed as

$$x_1(t) = s_1(t) + s_2(t) \tag{1}$$

where $x_1(t)$ is the observed mixture, and $s_1(t)$ and $s_2(t)$ are the original source signals which are assumed to be modeled by the autoregressive (AR) process [4]

$$s_j(t) = -\sum_{m=1}^{M_j} a_{s_j}(m;t)\, s_j(t-m) + e_j(t) \tag{2}$$

where $a_{s_j}(m;t)$ denotes the m^{th} order AR coefficient of the j^{th} source at time t, M_j is the maximum AR order, and $e_j(t)$ is an independent identically distributed (i.i.d.) random signal with zero mean and variance $\sigma^2_{e_j}$. This AR process is particularly interesting in source separation; firstly, many audio signals satisfy this process [5] and secondly, it enables us to formulate a virtual mixture by weighting and time-shifting the single channel mixture $x_1(t)$ as

$$x_2(t) = \frac{x_1(t) + \eta x_1(t-\delta)}{1+|\eta|} \tag{3}$$

* Corresponding author. Tel.: +44 191 222682; fax: +44 191 2228180.
E-mail addresses: w.l.woo@ncl.ac.uk, lok.woo@ncl.ac.uk (W.L. Woo).

0165-1684/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2013.08.017

In Eq. (3), $\eta \in \Re$ is the weight parameter, and δ is the time-delay. The mixture in (1) and (3) is termed as "artificial stereo" because it has an artificial resemblance of a stereo signal except that it is given by one location which results in the same time-delay but different attenuation of the source signals. To show this, we can express (3) in terms of the source signals, AR coefficient and time-delay as

$$\begin{aligned} x_2(t) &= \frac{x_1(t) + \eta x_1(t-\delta)}{1+|\eta|} \\ &= \frac{s_1(t)+s_2(t)+\eta[s_1(t-\delta)+s_2(t-\delta)]}{1+|\eta|} \\ &= \frac{-\sum_{m=1}^{M_1} a_{s_1}(m;t)s_1(t-m) + e_1(t)}{1+|\eta|} + \frac{\eta s_1(t-\delta)}{1+|\eta|} \\ &\quad + \frac{-\sum_{m=1}^{M_2} a_{s_2}(m;t)s_2(t-m) + e_2(t)}{1+|\eta|} + \frac{\eta s_2(t-\delta)}{1+|\eta|} \\ &= \frac{(-a_{s_1}(\delta)+\eta)}{1+|\eta|}s_1(t-\delta) + \frac{(-a_{s_2}(\delta)+\eta)}{1+|\eta|}s_2(t-\delta) \\ &\quad + \frac{e_1(t) - \sum_{m=1}^{M_1} a_{s_1}(m;t)s_1(t-m)}{1+|\eta|} \\ &\quad + \frac{e_2(t) - \sum_{m=1}^{M_2} a_{s_2}(m;t)s_2(t-m)}{1+|\eta|}. \end{aligned} \quad \begin{matrix} \\ \\ \\ \\ \\ \\ m \neq \delta \\ \\ m \neq \delta \end{matrix} \tag{4}$$

Define

$$a_j(t;\delta,\eta) = \frac{-a_{s_j}()\delta;t+\eta}{1+|\eta|} \tag{5}$$

$$r_j(t;\delta,\eta) = \frac{e_j(t) - \sum_{m=1}^{M_j} a_{s_j}(m;t)s_j(t-m)}{1+|\eta|}, \quad m \neq \delta \tag{6}$$

where $a_j(t;\delta,\eta)$ and $r_j(t;\delta,\eta)$ represent the mixing attenuation [1] and residue of the j^{th} source, respectively. Using Eqs. (5) and (6), the overall proposed mixture model[1] can now be formulated in terms of the sources as

$$\begin{aligned} x_1(t) &= s_1(t)+s_2(t) x_2(t) = a_1(t;\delta,\eta)s_1(t-\delta) \\ &\quad + a_2(t;\delta,\eta)s_2(t-\delta) + r_1(t;\delta,\eta) + r_2(t;\delta,\eta). \end{aligned} \tag{7}$$

By comparing with the single-channel mixture $x_1(t)$, the artificial stereo mixture $x_2(t)$ contains extra information i.e. $a_j(t;\delta,\eta)$, δ, and $r_j(t;\delta,\eta)$. The terms $a_j(t;\delta,\eta)$ and $r_j(t;\delta,\eta)$ represent the signature of the j^{th} source. In this light, these parameters can be used for estimating the parameter of the j^{th} source.

3. Assumptions

For the purposed of analysis, the following assumptions are used:

Assumption 1. The sources satisfy the windowed-disjoint orthogonality (WDO) [6] condition where different signals are approximately orthogonal to each other.

$$S_i(\tau,\omega)S_j(\tau,\omega) \approx 0, \quad \forall i \neq j, \ \forall \tau, \omega, \tag{8}$$

where $S_j(\tau,\omega)$ is the Short-Time Fourier Transform (STFT) of $s_j(t)$ defined as

$$\begin{aligned} S_j(\tau,\omega) &= F^W[s_j(t)](\tau,\omega) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} W(t-\tau)s_j(t)e^{-i\omega t}dt \end{aligned} \tag{9}$$

and $W(t)$ is the window function. The STFT is performed on the signal frame-by-frame and thus, τ represents the window shift.

Assumption 2. The sources satisfy the local stationarity of the time–frequency representation. This refers to the approximation of $S_j(\tau,\omega) \approx S_j(\tau-\phi,\omega)$ where ϕ is the maximum time-delay (shift) associated with $F^W(\cdot)$ with an appropriate window function $W(\cdot)$. If ϕ is small compared with the length of $W(\cdot)$ then $W(-) \approx W(\cdot)$ [7]. Hence, the Fourier transform of a windowed function with shift ϕ yields approximately the same Fourier transform without ϕ for the proposed method, the artificial stereo mixture is shifted by δ and by invoking the local stationarity this leads to

$$\begin{aligned} s_j(t-\delta) &\overset{STFT}{\rightarrow} e^{-i\omega\delta}S_j(\tau-\delta,\omega) \\ &\approx e^{-i\omega\delta}S_j(\tau,\omega), \forall \delta, |\delta| \leq \phi \end{aligned} \tag{10}$$

Thus, the STFT of $s_j(t-\delta)$ where $|\delta| \leq \phi$ is approximately $e^{-i\omega\delta}S_j(\tau,\omega)$ according to the local stationarity.

4. Time–frequency representation

Based on the assumptions, the mixing model is transformed into the TF domain by using the STFT of $x_j(t)$, $j = 1, 2$ as

$$\begin{aligned} X_1(\tau,\omega) &= S_1(\tau,\omega)+S_2(\tau,\omega)X_2(\tau,\omega) \\ &= a_1(\tau)e^{-i\omega\delta}S_1(\tau-\delta,\omega) + a_2(\tau)e^{-i\omega\delta}S_2(\tau-\delta,\omega) \\ &\quad - \left(\sum_{\substack{m=1 \\ m\neq\delta}}^{M_1} \frac{a_{s_1}(m;\tau)}{1+|\eta|}e^{-i\omega m}S_1(\tau-m,\omega) + \sum_{\substack{m=1 \\ m\neq\delta}}^{M_2} \frac{a_{s_2}(m;\tau)}{1+|\eta|}e^{-i\omega m}S_2(\tau-m,\omega) \right), \quad \forall(\tau,\omega) \end{aligned} \tag{13}$$

for δ and $m \leq \phi$. In Eq. (13), we have used the fact that $e_j(t) \ll s_j(t)$, thus the TF of $r_j(t)$ in (6) simplifies to

$$R_j(\tau,\omega) = -\sum_{\substack{m=1 \\ m\neq\delta}}^{M_j} \frac{a_{s_j}(m;\tau)}{1+|\eta|}e^{-i\omega m}S_j(\tau-m,\omega) \tag{14}$$

[1] In DUET [1] $x_1(t) = \sum_{n=1}^{N} s_n(t)$, $x_2(t) = \sum_{n=1}^{N} c_n(t-\beta_n)$ where N is the number of sources, β_n is the arrival delay between the microphones, and c_n is a relative attenuation factor corresponding to the ratio of the attenuations of the paths between sources and microphones.