Chapter 5
A Multimodal Solution to Blind Source Separation of Moving Sources

Syed Mohsen Naqvi  
Loughborough University, UK

Yonggang Zhang  
Harbin Engineering University, China

Miao Yu  
Loughborough University, UK

Jonathon A. Chambers  
Loughborough University, UK

ABSTRACT

A novel multimodal solution is proposed to solve the problem of blind source separation (BSS) of moving sources. Since for moving sources the mixing filters are time varying, therefore, the unmixing filters should also be time varying and can be difficult to track in real time. In this solution the visual modality is utilized to facilitate the separation of moving sources. The movement of the sources is detected by a relatively simplistic 3-D tracker based on video cameras. The tracking process is based on particle filtering which provides robust tracking performance. Positions and velocities of the sources are obtained from the 3-D tracker and if the sources are moving, a beamforming algorithm is used to perform real time speech enhancement and provide separation of the sources. Experimental results show that by utilizing the visual modality, a good BSS performance for moving sources in a low reverberant environment can be achieved.

INTRODUCTION

Professor Colin Cherry in 1953 first asked the question: “How do we [humans] recognise what one person is saying when others are speaking at the same time?” (Cherry, 1953). This was the genesis of the so-called machine cocktail party problem, i.e. mimicking the ability of a human to separate sound sources within a machine, and
A Multimodal Solution to Blind Source Separation of Moving Sources

Attempts to solve it have evolved from the signal processing community in the form of convolutive blind source separation (CBSS), which is a topic of considerable active research due to its potential applications (Haykin, Eds., 2000). CBSS consists of estimating sources from observed audio mixtures with only limited information and the associated algorithms have been conventionally developed in either the time or frequency domains (Bregman, 1990; Cichocki & Amari, 1990; Yilmaz & Rickard, 2004; Wang et al., 2005; Parra & Spence, 2000; Bingham et al. 2000; Makino et al., 2005; Sanei et al., 2007; Naqvi et al., 2008, 2009). Frequency domain convolutive blind source separation (FDCBSS) has however been a more popular approach as the time-domain convolutive mixing is converted into a number of independent complex instantaneous mixing operations. The permutation problem inherent to FDCBSS presents itself when reconstructing the separated sources from the separated outputs of these instantaneous mixtures and grows geometrically with the number of instantaneous mixtures (Wang et al., 2005).

Most existing BSS algorithms assume that the sources are physically stationary and based on statistical information extracted from the received mixed audio data (Cichocki & Amari, 1990; Wang et al., 2005; Parra & Spence, 2000). However, in many real applications, the sources may be moving. In such applications, there will generally be insufficient data length available over which the sources are physically stationary, which limits the application of these algorithms. Only a few papers have been presented in this area (Mukai et al., 2003; Koutras et al., 2000; Naqvi et al., 2008; Prieto & Jinachitra 2005; Hild-II et al., 2002). In (Mukai et al., 2003), sources are separated by employing frequency domain ICA using a block-wise batch algorithm in the first stage, and the separated signals are refined by postprocessing in the second stage which constitutes crosstalk component estimation and spectral subtraction. In the case of (Koutras et al., 2000), they used a framewise on-line algorithm in the time domain. However, both these two algorithms potentially assume that in a short period the sources are physically stationary, or the change of the mixing filters is very slow, which are very strong constraints. In (Prieto & Jinachitra, 2005), BSS for time-variant mixing systems is performed by piecewise linear approximations. In (Hild-II et al., 2002), they used an online PCA algorithm to calculate the whitening matrix and another online algorithm to calculate the rotation matrix. However, both algorithms are designed only for instantaneous source separation, and can not separate convolutive mixed signals. Fundamentally, it is very difficult to separate convolutively mixed signals by utilizing the statistical information only extracted from audio signals, and this is not the manner in which humans solve the problem (Haykin, Eds., 2007) since they use both their ears and eyes.

In this chapter, a multimodal approach is therefore proposed by utilizing not only received linearly mixed audio signals, but also the video information obtained from cameras. A video system can capture the approximate positions and velocities of the speakers, from which we can identify the directions and motions, i.e., stationary or moving, of the speakers. A source is identified as moving if the velocity is larger than an upper bound value. In this case, a beamforming method is used to enhance the signal from one source direction and reduce the energy received from another source direction, so that source separation can be obtained. Although the beamforming approach can only reduce the signal from a certain direction and the reverberance of the interference still exists, which are also limitations in the BSS approach, it can obtain a good separation performance in a low reverberation environment. To obtain an online real time source separation, the beamforming approach used in this work only depends on the angle of arrival information to the sensor array, and no received audio data are required (Veen & Buckley, 1988).