Signal Processing 92 (2012) 1916–1927

Contents lists available at SciVerse ScienceDirect

Signal Processing

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

Use of bimodal coherence to resolve the permutation problem in convolutive BSS

Qingju Liu *, Wenwu Wang **, Philip Jackson ***

Centre for Vision, Speech and Signal Processing, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, GU2 7XH, United Kingdom

ARTICLE INFO

Article history:
Received 30 January 2011
Received in revised form
4 November 2011
Accepted 7 November 2011
Available online 17 November 2011

Keywords:
Convolutive blind source separation (BSS)
Audio–visual coherence
Gaussian mixture model (GMM)
Feature selection and fusion
Adapted expectation maximization (AEM)
Indeterminacy

ABSTRACT

Recent studies show that facial information contained in visual speech can be helpful for the performance enhancement of audio-only blind source separation (BSS) algorithms. Such information is exploited through the statistical characterization of the coherence between the audio and visual speech using, e.g., a Gaussian mixture model (GMM). In this paper, we present three contributions. With the synchronized features, we propose an adapted expectation maximization (AEM) algorithm to model the audio–visual coherence in the off-line training process. To improve the accuracy of this coherence model, we use a frame selection scheme to discard nonstationary features. Then with the coherence maximization technique, we develop a new sorting method to solve the permutation problem in the frequency domain. We test our algorithm on a multimodal speech database composed of different combinations of vowels and consonants. The experimental results show that our proposed algorithm outperforms traditional audio-only BSS, which confirms the benefit of using visual speech to assist in separation of the audio.

1. Introduction

Human speech perception is essentially bimodal as speech is perceived by the interactions of auditory and visual sensory processing [1,2]. Looking at the speaker's lips improves the intelligibility of human speech embedded in cocktail party noise due to the contribution of the complementary visual information [2]. There is a complex nonlinear relationship between the auditory and visual streams, usually referred to as the audio–visual coherence or correlation [3]. In feature space, the coherence can be coded by audio–visual atoms or dictionaries [4,5] with matching pursuit [6] techniques, or characterized statistically with models such as Gaussian mixture models (GMM) [7]. Exploiting these cross-modal interactions, the visual stream has proven a success in improving the robustness to noise in many fields of applications, including automatic speech recognition [8], speaker localization [4,9], speech enhancement or audio filtering [10,11], and blind source separation [3,5,12–16].

In traditional blind source separation (BSS) for auditory mixtures, typically only audio signals are considered. Under the framework of independent component analysis (ICA) [17], the BSS problems have been extensively studied and many classical algorithms have been proposed for the instantaneous mixing model such as the "J–H" algorithm [18], JADE [19], Infomax [20], SOBI [21] and FastICA [22] algorithms. For the more complex convolutive mixing model, one can apply either the time domain deconvolution algorithms [23–25] or the frequency domain separation algorithms [12–15,26–31], which often suffer from the permutation and scaling ambiguity problems.

* Corresponding author. Tel.: +44 1483 683413; fax: +44 1483 686031.
** Corresponding author. Tel.: +44 1483 686039.
*** Corresponding author. Tel.: +44 1483 686044.
E-mail addresses: Q.Liu@surrey.ac.uk (Q. Liu), W.Wang@surrey.ac.uk (W. Wang), P.Jackson@surrey.ac.uk (P. Jackson).

0165-1684/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2011.11.007

Considering the bimodal nature of human speech, we could potentially improve the separation of the source signals from their audio mixtures utilizing the audio–visual coherence obtained by the integration of visual speech. This is known as audio–visual or bimodal BSS [3,5,12,13,15,16], a recent development in multi-modal signal processing. Sodoyer et al. [3] addressed the separation problem for an instantaneous mixture of decorrelated sources, with no further assumptions on independence or non-Gaussianity. Wang et al. [13] implemented a similar idea by applying the Bayesian framework to the fused feature observations for both instantaneous and convolutive mixtures. Rivet et al. [12] proposed a new statistical tool utilizing the log-Rayleigh distribution for modeling the audio–visual coherence, and then used the coherence to address the permutation and scaling ambiguities in the spectral domain. Casanovas et al. [5] detected temporal audio–visual structures represented by atoms taken from redundant dictionaries, and extracted sources from a soundtrack. Naqvi et al. [16] utilized beamforming in the frequency domain for moving sources in the teleconference-like scenario, incorporating the geometrical model derived on the basis of the beamforming theory.

Despite being promising, these approaches are also limited in some situations. For example, the algorithm proposed in [3] was designed only for instantaneous mixtures. The method in [13] considered a convolutive model with a relatively small number of taps for the mixing filters. The approach in [12] modeled the audio–visual coherence in a high dimensional feature space, which often results in an over-fitting problem and therefore is sensitive to outliers. Cross-modal correlation was not exploited in the separation stage in [5], where visual information was used only for voice activity detection. In [16], the video provided the position information about the distance and azimuth angles between the moving speakers and the microphone array, however, source separation was still performed in the audio domain.

In this paper, we attempt to address some of these limitations. Motivated by the work in [12,13], we follow a similar two-stage framework which includes off-line training and online separation. In particular, we consider a convolutive mixing model and address the permutation problem associated with the frequency domain BSS (FD-BSS). In the off-line training stage, we build a model to statistically characterize the audio–visual coherence in the feature space. This coherence is built on the audio–visual features extracted from the target speech. Mel-frequency cepstral coefficients (MFCCs) are used as the audio features, and the lip width and height as visual features, which are synchronized with the audio features on a frame-by-frame basis before statistical training. In the separation stage, coherence maximization is applied for the alignment of the ICA-separated spectral components. Different from [12,13], however, we have proposed three new techniques to improve the training and separation processes. First, a frame selection scheme is proposed to remove the non-stationary features which consequently improves the robustness and accuracy of the estimation of the audio–visual coherence. Second, the classical expectation maximization (EM) algorithm is modified to take into account the different influences of the audio features, resulting in an adapted EM (AEM) algorithm, which further improves the estimation of the joint audio–visual probability. Third, a novel sorting scheme is proposed to address the permutation problem. A preliminary version of this work was presented in [15]. Different from [15], in this paper, we have developed a robust feature selection scheme for audio–visual modeling as mentioned above. In addition, we have further improved the audio feature representation as described in Section 3.1. Moreover, here we have performed systematic evaluations on real recordings, and compared the performance of the proposed method with the state-of-the-art methods.

The remainder of the paper is organized as follows. An overview of traditional frequency domain convolutive BSS and the framework of the proposed audio–visual BSS system are presented in Section 2. Then Section 3 introduces the feature extraction and fusion method for the modeling of the cross-model correlation, including a new frame selection approach and an adapted expectation maximization algorithm to improve the accuracy of this model. The proposed de-permutation algorithm exploiting the audio–visual coherence is presented in Section 4. The simulation results are analyzed and discussed in Section 5, followed by the conclusions.

2. BSS for convolutive mixtures

2.1. Convolutive model

BSS aims to recover sources from their mixtures without any or with little prior knowledge about the sources or the mixing process. Consider a cocktail party scenario, the observation at each sensor is the sum of K filtered source signals, which can be approximated by the convolutive model

$$x_p(n)=\sum_{k=1}^{K}\sum_{m=0}^{+\infty}h_{pk}(m)s_k(n-m)+\xi_p(n),$$

$$\mathbf{x}(n)=\mathbf{H}*\mathbf{s}(n)+\boldsymbol{\xi}(n), \tag{1}$$

where h_{pk} represents the room impulse response filter from source k to sensor p. We denote $\mathbf{x}(n)=[x_1(n),\ldots,x_P(n)]^T$ as the observation vector at the discrete time index n; $\mathbf{s}(n)=[s_1(n),\ldots,s_K(n)]^T$ the source vector and $\boldsymbol{\xi}(n)=[\xi_1(n),\ldots,\xi_P(n)]^T$ the additive noise vector, where T is vector transpose. $\mathbf{H}$ is the mixing matrix whose elements are filters h_{pk} and $*$ denotes convolution.

Convolutive BSS aims to find a set of separation filters $\{w_{kp}\}$ that satisfy

$$\hat{s}_k(n)=y_k(n)=\sum_{p=1}^{P}\sum_{m=0}^{+\infty}w_{kp}(m)x_p(n-m),$$

$$\hat{\mathbf{s}}(n)=\mathbf{y}(n)=\mathbf{W}*\mathbf{x}(n), \tag{2}$$

where $\mathbf{W}$ is the separation matrix whose entry w_{kp} is the impulse response filter from observation p to the estimate of source k ($\mathbf{y}(n)$ or $\hat{\mathbf{s}}(n)$ represents the estimated version of $\mathbf{s}(n)$). We consider a time-invariant system where both