Digital Signal Processing 36 (2015) 174–183

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Binaural source separation based on spatial cues and maximum likelihood model adaptation

Roohollah Abdipour [a], Ahmad Akbari [a,*], Mohsen Rahmani [a,b], Babak Nasersharif [a,c]

[a] Audio & Speech Processing Lab, School of Computer Engineering, Iran University of Science & Technology, Tehran, Iran
[b] Computer Engineering Department, Faculty of Engineering, Arak University, Arak, Iran
[c] Electrical & Computer Engineering Department, K.N. Toosi University of Technology, Tehran, Iran

ARTICLE INFO

Article history:
Available online 17 September 2014

Keywords:
Binaural source separation
Model adaptation
Maximum likelihood linear regression
Statistical signal processing
Speech enhancement

ABSTRACT

This paper describes a system for separating multiple moving sound sources from two-channel recordings based on spatial cues and a model adaptation technique. We employ a statistical model of observed interaural level and phase differences, where maximum likelihood estimation of model parameters is achieved through an expectation-maximization algorithm. This model is used to partition spectrogram points into several clusters (one cluster per source) and generate spectrogram masks accordingly for isolating individual sound sources. We follow a maximum likelihood linear regression (MLLR) approach for tracking source relocations and adapting model parameters accordingly. The proposed algorithm is able to separate more sources than input channels, i.e. in the underdetermined setting. In simulated anechoic and reverberant environments with two and three speakers, the proposed model-adaptation algorithm yields more than 10 dB gain in signal-to-noise-ratio-improvement for azimuthal source relocations of 15° or more. Moreover, this performance gain is achievable with only 0.6 seconds of input mixture received after relocation.

1. Introduction

Sound source separation is a well-known challenge with important applications. For example, consider a hearing-aid device that should separate utterances of a target speaker from competing sound sources. An effective source separation algorithm is rewarding in this context. As a result, various algorithms have been proposed in the literature.

Independent component analysis (ICA) is one of the well-known source separation approaches that rely on the availability of multi-channel observations [1–6]. Commonly, source signals are assumed to be statistically independent. This independence assumption makes it possible to use optimization methods based on higher order statistics (HOS) [7]. Alternatively, ICA methods based on the maximum likelihood principle [8] and second-order statistics (SOS) [9–11] can be applied. SOS-based solutions are especially useful in situations with uncorrelated and non-stationary sources. Traditional ICA methods are famous due to their ability to separate signals without any *apriori* knowledge about sound sources and environmental conditions (such as the configuration of microphones). However, traditional ICA methods fail in underdetermined conditions (i.e., when the number of sources exceeds the number of input channels). Another major limitation is that the mixing coefficients should be stationary for a period of time. But this constraint is not satisfied in real situations where sound sources can move.

Model-based source separation is another well-studied approach that incorporates *apriori* knowledge about sources. For example, code-book based methods [12,13] and hidden Markov model (HMM) based methods [14–16] have been widely used for speech enhancement. In these methods, some models are considered for noise and speech signals and the model parameters are estimated using a training set in advance. Nonnegative matrix factorization (NMF) [17] is another model-based source separation approach which was initially used in single-channel situations [18–24]. NMF-based methods decompose a nonnegative matrix of observations into a multiplication of two nonnegative matrices, a basis matrix containing a set of basis vectors for each source and a gain matrix containing the mixing coefficients. Source signals are usually obtained by calculating a wiener-like filter based on the basis and gain matrices (e.g., see [21,23–27]).

* Corresponding author.
E-mail addresses: r_abdipour@iust.ac.ir (R. Abdipour), akbari@iust.ac.ir (A. Akbari), m-rahmani@araku.ac.ir (M. Rahmani), bnasersharif@eetd.kntu.ac.ir (B. Nasersharif).

http://dx.doi.org/10.1016/j.dsp.2014.09.001
1051-2004/

The NMF-based solutions have shown promising results, especially for non-stationary signals. However, the spectrograms of real signals are strongly diverse and are often poorly modeled using a low-rank structure such as NMF. Furthermore, NMF is usually applied on whole excerpts of data and hence do not appear as very appropriate for real-time processing, but rather for off-line applications.

Other methods employ binaural cues, such as interaural time or phase difference (ITD or IPD) and interaural level difference (ILD), for separating sound sources [28–31]. Generally, for each spatially-fixed sound source, the corresponding subband-level ⟨IPD, ILD⟩ observations concentrate in a specific region in the IPD–ILD space. So, the IPD–ILD space of a multi-source environment can be modeled as a set of observation clusters (one cluster per source). The position of each cluster depends on the location of the source, and differs for spatially-disjoint sources. Based on this point, many source separation methods aim to find clusters of observations in IPD–ILD space, and assign each cluster to a source. For example, in [28] a two-dimensional histogram of ITD and ILD features is constructed and each peak of the histogram is assigned to a source. Then, a time-frequency mask is constructed accordingly to partition the input mixture into the original signals. For another example, [29] describes a supervised classification algorithm that learns the rules of separating the target source based on ITD and ILD features. This classifier is used to calculate a binary mask for source separation. The method described in [30] learns the probability distribution of a target speaker given its ⟨ITD, ILD⟩ observation pair. This method uses this probability distribution as a look-up table to determine the probability that each spectrum bin be dominated by the target source and calculate a mask accordingly.

Incorporating source models in conjunction with spatial models is also common and usually improves performance. For example, in [32] spatial cues are employed as prior knowledge to separate sound sources based on nonnegative tensor factorization. For another example, in [31], spatial models of sources are combined with *apriori* trained source models. These models give the likelihood of each source based on the current observation and a time-frequency mask is built accordingly for source separation. In [33] a library of source models is employed to incorporate prior knowledge about each source. It also considers models that represent the spatial and environmental conditions. The proposed framework is shown to be flexible enough to be applicable in different conditions.

Promising results have been reported for the localization-based methods. However, these methods are only useful in offline scenarios where sources do not relocate. That is due to the fact that they are based on localization models of spatially-fixed sources and need a relatively long segment of observations to estimate model parameters. In effect, these methods fail in real situations with moving sources where model parameters should be updated over time according to new source locations.

Our main idea is to exploit model-adaptation techniques to adjust model parameters according to new source locations. We employ a bivariate Gaussian model to represent observations of each source, where the model parameters are estimated using an expectation-maximization algorithm. We also employ the maximum likelihood linear regression (MLLR) technique to adjust model parameters after possible source relocations. We use this model to partition the observations into source-related clusters and build a separation mask accordingly.

It's worth mentioning that although incorporating source models besides the spatial models can improve the performance, this study is limited to the update of spatial models and source movement tracking in order to build a solution for online applications. Obviously, one can still utilize source models besides our up-to-date spatial models to improve performance.

The remainder of the paper is organized as follows. In Section 2, we recall a state-of-the-art statistical model for sound source separation, which is useful for spatially-fixed sources. Then, in Section 3, we propose a model adaptation algorithm to update the parameters of this model as sources relocate over time. The performance of this algorithm is evaluated in Section 4 for different conditions. Finally, the paper concludes in Section 5.

2. Background

A state-of-the-art approach for building the spatial model of concurrent sound sources is proposed in [31]. Therein, a Gaussian mixture model (GMM) is employed to represent the localization cues of spatially-fixed sources. Moreover, an expectation-maximization (EM) algorithm is derived to estimate model parameters. We use this model as our spatial model. We also use its parameter estimation algorithm to initialize our model. This model and its parameter estimation algorithm are detailed in this section. Then, in the next section, we propose a model adaptation algorithm to track source movements and update the spatial model accordingly.

Consider I spatially-fixed sources of signals $\{s_i(t), i = 1..I\}$. The binaural recordings $x^l(t)$ and $x^r(t)$, corresponding to mixtures arriving at the left and right ears, respectively, are modeled as:

$$x^l(t) = \sum_{i=1}^{I} s_i\left(t - \tau_i^l\right) * h_i^l(t) \tag{1}$$

$$x^r(t) = \sum_{i=1}^{I} s_i\left(t - \tau_i^r\right) * h_i^r(t) \tag{2}$$

In this model, $\tau_i^{l,r}$ are the delays of the direct path of source i to the left and right ears, and $h_i^{l,r}(t)$ show the effects of room and head-related impulse responses (RIR and HRIR), excluding the delay of arrival.

Supposing that the spectra of these mixtures are approximately disjoint (i.e., supposing that each time-frequency bin corresponds to one source), the element-wise ratio of time-frequency units of these mixtures can be expressed as:

$$R(\lambda, f) = \frac{X^l(\lambda, f)}{X^r(\lambda, f)} = \frac{H_i^l(\lambda, f)}{H_i^r(\lambda, f)} e^{-j2\pi f(\tau_i^l - \tau_i^r)} \tag{3}$$

where upper-case letters show the short-term Fourier transform (STFT) of their corresponding lower-case signals, and λ and f are the frame and frequency-bin indices, respectively. The interaural level and phase differences between the two ears are written as:

$$IPD(\lambda, f) = 2\pi f\left(\tau_i^l - \tau_i^r\right) = 2\pi f \tau_i, \quad \tau_i = \tau_i^l - \tau_i^r \tag{4}$$

$$ILD(\lambda, f) = \ln\left(\frac{H_i^l(\lambda, f)}{H_i^r(\lambda, f)}\right) \tag{5}$$

The IPD values are constrained to the interval $(-\pi, +\pi]$. For spatially-disjoint sources, their subband observation $\boldsymbol{o}(\lambda, f) = [IPD(\lambda, f), ILD(\lambda, f)]$ form distinct clusters in IPD–ILD space [28,31]. The cluster related to the source i can be modeled using a two-variable Gaussian distribution as:

$$p\left(\boldsymbol{o}(\lambda, f) \mid i, \Phi_i(f)\right) = \frac{1}{2\pi |\boldsymbol{C}_i(f)|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{o}(\lambda, f) - \boldsymbol{\mu}_i(f))^T \boldsymbol{C}_i^{-1}(f)(\boldsymbol{o}(\lambda, f) - \boldsymbol{\mu}_i(f))} \tag{6}$$