Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 76 (2016) 201–217

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Computational methods for underdetermined convolutive speech localization and separation via model-based sparse component analysis

Afsaneh Asaei [a,*], Hervé Bourlard [a,b], Mohammad J. Taghizadeh [a,b], Volkan Cevher [b]

[a] Idiap Research Institute, Martigny, Switzerland
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Received 11 January 2015; received in revised form 16 June 2015; accepted 14 July 2015
Available online 21 July 2015

Abstract

In this paper, the problem of speech source localization and separation from recordings of convolutive underdetermined mixtures is addressed. This problem is cast as recovering the spatio-spectral speech information embedded in a microphone array compressed measurements of the acoustic field. A model-based sparse component analysis framework is formulated for sparse reconstruction of the speech spectra in a reverberant acoustic resulting in joint localization and separation of the individual sources. We compare and contrast the algorithmic approaches to model-based sparse recovery exploiting spatial sparsity as well as spectral structures underlying spectrographic representation of speech signals. In this context, we explore identification of the sparsity structures at the auditory and acoustic representation spaces. The audiory structures are formulated upon the principles of structural grouping based on proximity, autoregressive correlation and harmonicity of the spectral coefficients and they are incoporated for sparse reconstruction. The acoustic structures are formulated upon the image model of multipath propagation and they are exploited to characterize the compressive measurement matrix associated with microphone array recordings.

Three approaches to sparse recovery relying on combinatorial optimization, convex relaxation and sparse Bayesian learning are studied and evaluated on thorough experiments. The sparse Bayesian learning method is shown to yield better perception quality while the interference suppression is also achieved using the combinatorial approach with the advantage of offering the most efficient computational cost. Furthermore, it is demonstrated that an average autoregressive model can be learned for speech localization while exploiting the proximity structure in the form of block sparse coefficients enables accurate localization and high quality speech separation. Throughout the extensive empirical evaluation, we confirm that a large and random placement of the microphones enables significant improvement in source localization and separation performance.

Keywords: Structured sparse representation; Model-based sparse recovery; Reverberation; Source localization and separation; Sparse component analysis; Computational auditory scene analysis

1. Introduction

Source localization and separation are central problems in various microphone array applications. This work takes place at the intersection of sparse component analysis and computational auditory scene analysis (CASA). Motivated by the commonalities between these two approaches to

* Corresponding author.
E-mail addresses: afsaneh.asaei@idiap.ch (A. Asaei), herve.bourlard@idiap.ch (H. Bourlard), mohammad.taghizadeh@idiap.ch (M.J. Taghizadeh), volkan.cevher@epfl.ch (V. Cevher).

http://dx.doi.org/10.1016/j.specom.2015.07.002
0167-6393/

source separation, we consider structural dependencies influencing our auditory system for structured sparse recovery to develop a model-based sparse component analysis framework.

We consider the microphone array acquisition model as a linear convolutive mixing process stated as

$$x_m = \sum_{n=1}^{N} h_{mn} \circledast s_n, \quad \forall m \in \{1, \ldots, M\} \tag{1}$$

where the signal of each microphone x_m is characterized as a superposition of the signal of individual sources s_n, $\forall n \in \{1, \ldots, N\}$ convolved with the acoustic channel, h_{mn}, between the position of source n and microphone m; N and M denote the number of sources and microphones, respectively.

This formulation is stated in time domain. To exploit the sparsity structures of sound as we shall see in this paper, the frequency domain representation is considered. Hence, the short time Fourier transform is applied on the microphone array signals. The convolution-multiplication property of the Fourier transform leads to the following mixing model

$$X_m = \sum_{n=1}^{N} H_{mn} S_n, \quad \forall m \in \{1, \ldots, M\} \tag{2}$$

where X_m, H_{mn} and S_n are the frequency domain representations of x_m, h_{mn} and s_n respectively.

The goal is to recover the individual source signals from M recorded mixtures. There is no prior knowledge about N, M and the acoustic channels H_{mn} and the estimation of the signals can only be achieved under assumptions about the signal or channel characteristics. Furthermore, the linear system expressed in (2) is underdetermined if $N \geqslant M$. Hence, additional assumptions are required to circumvent the ill-posedness of the separation problem. In the next section, we overview some of the prior works on multichannel techniques for speech separation.

1.1. Prior work

The signal of individual sources can be recovered through multichannel linear filtering. These techniques can be grouped in two categories: *independent component analysis* and *beamforming*. The alternative non-linear strategies to demixing rely on extraction of the descriptions of individual sounds within the framework of *computational auditory scene analysis* and *sparse component analysis*. In the following, we study the assumptions underlying each approach and the scope of their application.

Independent component analysis (ICA) relies on the assumption that the signals are statistically independent. Hence, the objective is formulated to estimate an inverse/demixing filter such that the recovered source signals are statistically independent (Comon and Jutten, 2010). This approach typically requires the number of source and microphones to be known in advance. In addition, the system has to be (over-)determined, i.e. $M \geqslant N$ so that an inverse filter exists. Furthermore, the mixing matrix must remain the same (stationary acoustic assumption) for a relatively long period of time to provide a reasonable estimate of a large number of model parameters. This assumption is difficult to fulfill in the realistic scenarios in which speakers turn their heads or move around.

Buchner et al. proposed to incorporate characterization of the room acoustics in the separation process (Buchner et al., 2007). Their approach exploits the statistical independence assumption of the sources to perform joint deconvolution and separation of the signals in overdetermined scenarios. Nesta et al. proposed an extension for underdetermined scenario where multiple complex valued ICA adaptations jointly estimate the mixing matrix and the temporal activities of multiple sources in each frequency band to exploit the spectral sparsity of speech signals (Nesta and Omologo, 2012). The method does not explicitly rely on identification of the acoustic channel and recovery of the desired source imposes a permutation problem due to mis-alignment of the individual source components (Nesta and Omologo, 2012; Wang et al., 2011). Other extensions of ICA for the underdetermined scenarios consist in integration with sparse masking techniques within a hierarchical separation framework (Araki et al., 2004; Davies and Mitianoudis, 2004).

Beamforming is a geometric method to speech recovery that relies on steering/forming the beam pattern of the microphone array towards the desired source. This process can spatially filter out interferences from other directions regardless of the signal nature. Due to the spatial directivity, it can also mitigate the effect of reverberation which causes a field of dispersed signals. The limitation of beamforming is that separation is not possible when multiple sounds come from directions that are the same or near to each other (Wolfel and McDonough, 2009; Parra and Alvino, 2002).

Unlike the ICA approach, the beamforming requires information about the microphone array configuration and the sources (such as the direction of the desired source). However, there is no need to determine the number of spatially spread and reverberant interferences. It has been shown that beamforming can attain excellent separation performance in determined or overdetermined time-invariant demixing problems (Kumatani et al., 2011; Taghizadeh et al., 2012). However, only partial interference suppression is possible in underdetermined cases. Recent work considers non-linear combination of beamformers which incorporate sparsity of the spectro-temporal coefficients to address the underdetermined demising (Dmour and Davies, 2011). The application of this method is however limited to the anechoic scenarios and the performance is degraded in reverberant condition.

Huang proposed to exploit the acoustic channel to achieve speech separation and dereverberation (Huang et al., 2005). Their method applies a blind channel