Underdetermined Blind Separation of Speech Signals with Delays in Different Time-Frequency Domains

Alessandro Bastari, Stefano Squartini, and Francesco Piazza

Dipartimento di Elettronica, Intelligenza Artificiale e Telecomunicazioni-Università Politecnica delle Marche Via Brecce Bianche 12, I-60121 Ancona, Italy
a.bastari@univpm.it, {sts,upf}@deit.univpm.it

Abstract. This paper is devoted to the problem of speech signal separation from a set of observables, when the mixing system is underdetermined and static with unknown delays. The approaches appeared in the literature so far have shown that algorithms based on the property of sparsity of the original signals (effectively satisfied by speech sources) can be successfully applied to such a problem, specially if implemented in the time-frequency domain. Here, a survey on the usage of different time-frequency transforms within the already available three-step procedure for the addressed separation problem is carried out. The novelty of the contribution can be seen from this perspective: Wavelet, Complex Wavelet and Stockwell Transforms are the new transforms used in our problem, in substitution of the usual Short Time Fourier Transform (STFT). Their performances are analyzed and compared to those attainable through the STFT, evaluating how much different is the influence that their sparseness and spectral disjointness properties on the algorithm behavior.

1 Introduction

Recovering information from recordings where several audio sources are mixed is a challenging problem in digital signal processing. In particular it is often required to separate speech signals from the available observables, to get a certain level of intelligibility or listening quality. A common way followed in the literature to face this task is to see it as a problem of blind source separation (BSS) [1]. Actually, there is a big interest for BSS; this is mainly justified by its large applicability in several scientific fields, like bioinformatics, communications, speech and audio processing of course, imaging, and so on. In fact in all of them we are often asked to recover unknown sources from a certain set of observables (namely mixtures) detected by the sensors. This is what surely happens when we think of the well-known “cocktail party problem”, that is commonly used to introduce beginners to the scientific area related to BSS. In fact, in a cocktail party there are many audio sources (speech, music, various kinds of noise), contemporarily present at the sensors (the listener’s ears) and the goal is to be able to focus the attention on a single talker, minimizing the disturbing effect of the other signals.

In the BSS scenario we are supposed to separate the original signals by the only means of such mixtures, that is in general more than tracking a certain source of interest. The problem is completely blind when we do not have knowledge of the original signals and the system that performs the mixing. In order to face the large variety of possible applications, different formulations of the BSS problem have been proposed and studied, in accordance with the nature of the mixing [2], [3], [4]. First of all we
can distinguish between linear and nonlinear models. Within both these categories we can have static or convolutive mixing, mathematically represented by matrices of numbers and matrices of filters. The case of static mixing with delays (i.e. the sources are delayed, weighted and summed up to yield the mixture signals) has been considered too. Looking at the number of sources ($N$) or sensors ($M$) involved we can differentiate between overdetermined and underdetermined models, with $N \leq M$, $N > M$ respectively. Diverse techniques have been proposed to handle these formulation of BSS; they generally move from different hypotheses and exploit different characteristics of the signals involved. Therefore, they cannot be applied in all situations. In the following we shall take into account those approaches that are well suited to deal with separation of speech signals.

2 Separating Speech Signals Through BSS Models

In this section we shall describe the BSS model we are going to consider in the following for the separation of speech signals, and justify its choice in accordance with the assumptions we can make on the nature of sources. Moreover we will be able to exploit some characteristics of the signals involved to find out adequate separating techniques. Regarding speech signals, we can say that they generally superpose linearly at sensor level, like all acoustic signals. Neglecting the presence of echoes and delays (for the moment) in the model, we are allowed to consider the mixture values at time instant $t$ as a real coefficient linear combination of the source values at the same time instant. Therefore, let $x_i(t)$, $i = 1, 2, \ldots, M$ be the $M$ known signal mixtures (the sensor signals), yielded as the linear combination of $N$ unknown sources $s_j(t)$, $j = 1, 2, \ldots, N$, through an unknown matrix $A$ (dimensions: $M \times N$), namely mixing matrix, and eventually corrupted by the additive noise $\xi(t)$ . In formula:

$$
\begin{bmatrix}
  x_1(t) \\
  \vdots \\
  x_M(t)
\end{bmatrix}
= A
\begin{bmatrix}
  s_1(t) \\
  \vdots \\
  s_N(t)
\end{bmatrix} + \xi(t), \ \forall t.
$$

(1)

As aforementioned, the objective is to recover $s_j(t)$ by the only means of mixtures $x_i(t)$ . Eq.(1) can be written in the following compact form

$$
x(t) = As(t) + \xi(t)
$$

(2)

where $x(t) = [x_1(t) \ldots x_M(t)]^T$ and $s(t) = [s_1(t) \ldots s_N(t)]^T$ are the mixture and source vectors respectively at time instant $t$ . Assuming the involved signals to be of length $T$ ($t = 1, 2, \ldots, T$), we shall denote

$$
X = \begin{bmatrix}
  x_1(1) & x_1(2) & \cdots & x_1(T) \\
  \vdots & \vdots & \ddots & \vdots \\
  x_M(1) & x_M(2) & \cdots & x_M(T)
\end{bmatrix}
$$

(3)

as the $M \times T$ matrix having the $i$-th sensor sequence as its $i$-th row, and similarly: