A Clustering Approach to On-line Audio Source Separation

Julien Bourgeois

DaimlerChrysler Research and Technology
P.O. Box 23 60 - 89013 Ulm - Germany
julien.bourgeois@daimlerchrysler.com

Abstract

We have developed an on-line separation method for audio signals. The adopted approach makes use of the time-frequency transform of the signals as a sparse decomposition. Since the sources for the most part do not overlap in the time-frequency domain, we get raw estimates of their individual mixing parameters with an analysis of the mixture ratios. We then obtain reliable mixing parameters by dynamically clustering these instantaneous estimates. The mixing parameters are used to separate the mixtures, even at time-frequency points where the sources overlap. In addition, even when the mixing parameters change over time, our approach is able to separate signals with only one pass through the data. We have evaluated this approach first on computer generated anechoic mixtures and then on real echoic mixtures recorded in a car.

1. Introduction

When several persons speak at the same time in front of a speech-based human-machine interface, the separation of their individual speech becomes necessary since only one signal should be passed as input to a speech recognition component. This problem, sometimes referred to as the Cocktail Party Problem or Blind Source Separation, has been considered from many different points of view. Because humans have the ability to concentrate on a single voice in a babble of multiple speakers, the problem has been first studied from the psychological and perceptual perspective in the field of Auditory Scene Analysis [1]. Beamforming is another research area of interest. With an array of microphones, beamforming techniques allow one to separate signals which originate from different known spatial directions [2]. More recently, methods based on statistical principles have been investigated [3]. Since speech signals stem from independent physical processes, it seems appropriate to assume that the sources are independent. This criterion is used to estimate the inverse mixing matrix and the separation requires at least as many sensors as sources. Unfortunately, these methods are computationally expensive because they necessitate the estimation of high-order statistics and the minimization of a non-linear independence measure.

Another approach considers sources that are sparse (or have disjoint supports) in a transformed domain. This has been investigated by Zibulevsky using the complex spectrogram as a sparse decomposition on instantaneous mixtures [4]. Several researches have focused on the demixing of degenerate mixtures, i.e. when there are more sources than mixtures, using only two sensors [5,6,7]. Most of these sparse decomposition methods perform a first run on the data in order to gather statistics and to cluster the mixing parameters. Once the demixing parameters have been identified, a second run on the data performs the actual separation. The delay introduced by this process constitutes a critical drawback for some applications like dialog systems, since the user expects an immediate answer, or hearing aids. An on-line solution that applies when sources have disjoint short-term Fourier Transforms has been proposed [9]. In practice however, for some time-frequency points \((t, \omega)\) the sparsity assumption does not hold and the sources overlap. This causes audible distortion because of unnatural zeros in the Short Time Fourier Transform of the estimated sources [8]. Indeed, in the case of degenerate mixtures, no linear demixing can separate the sources.

In this paper, we focus on the overdetermined case and present a demixing scheme that continuously estimates the mixing parameters using a sparsity assumption. The demixing procedure performs even at time-frequency points where several sources overlap. In section 2, we describe the anechoic acoustic mixing model. In section 3, we propose to estimate the mixing parameters by clustering their instantaneous estimates. Section 4 presents how these parameters are used for separation in both anechoic and echoic conditions. In section 5, we report on the separation performance of our method. Finally, in section 6, we conclude and outline future work.

2. Mixing Model

We consider a linear array of \(N\) equally spaced microphones and we denote \(x_n(t), n = 1, \ldots, N\) their measurements. We are interested in recovering \(M\) sources \(s_m(t), m = 1, \ldots, M\) as they are received at the first mi-
crophone, so that we set
\[ x_1(t) = \sum_{m=1}^{M} s_m(t). \] (1)
The sources are assumed to be in far field and to propagate without reflection: only direct paths to the microphones exist. Under these conditions, each microphone receives one delayed copy of each source. Their readings are given by
\[ x_i(t) = \sum_{m=1}^{M} s_m(t - (i-1)\delta_m) \] (2)
where the delay \( \delta_m \) depends on the Direction Of Arrival (DOA) of the source \( s_m \). Since the delays can be fractional values of the sampling period, it is convenient to work in the frequency domain. To account for the short-time stationarity scale of speech, we consider the short-time windowed Fourier Transform (or spectrogram)
\[ X(t, \omega) = \text{FFT}\{h_t x\}(\omega) \]
of the signal \( x \), where \( h_t \) is a window function with finite support centered at \( t \). We can then rewrite (1-2) in the time-frequency domain as
\[ X(t, \omega) = A(\omega)S(t, \omega) \]
with \( A(\omega) = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ e^{-i\omega(N-1)\delta_1} & \cdots & e^{-i\omega(N-1)\delta_M} \end{pmatrix} \). (3)

3. Estimation of the Mixing Parameters

3.1. Instantaneous Estimation

We suppose that the time-frequency domain constitutes a sparse representation of the sources, i.e. at a given time-frequency point \((t, \omega)\), only one source \( s_k \) is non zero. We can evaluate the delay \( \delta_k \) using
\[ \hat{\delta}^{(n)}(t, \omega) = \varphi \left( \frac{X_1(t, \omega)}{X_n(t, \omega)} \right) / (n-1)\omega \]
for \( n = 2, \ldots, N \) where \( \varphi(z) \) denotes the phase of \( z \in C \). We use the mean over \( n = 2, \ldots, N \) of \( \hat{\delta}^{(n)}(t, \omega) \) as an instantaneous estimator \( \hat{\delta}(t, \omega) \) of \( \delta_k \). Because of the periodicity of the complex exponential, the estimation holds only if \( |\omega/\delta_k| < \pi \). Moreover, it performs better to estimate the delay on a frequency band that suits the microphone spacing \( \Delta \). That is why \( \hat{\delta}(t, \omega) \) is evaluated for \( \omega \in \frac{2\pi}{2\Delta} \{0, 1\}, \) where \( c = 342 \text{ m/s} \) denotes the speed of sound. If the sparsity assumption holds in a strict sense, and if there is no modeling error, \( \hat{\delta}(t, \omega) \) would exactly cover the \( M \) values \( \delta_1, \ldots, \delta_M \). In practice however, the sources overlap on a unknown subset of the time-frequency plane.

3.2. Clustering the Instantaneous Estimates

In order to improve the estimation of the delays \( \delta_k \), we gather the values of \( \hat{\delta}(t, \omega) \) in a FIFO buffer \( B \). \( B \) is then segmented into clusters, whose centers \( \hat{\delta}_k \) are the seeked estimates of \( \delta_k \). If prior information on the source locations is available, the cluster centers are initialized accordingly. Otherwise, on can distribute the initial centers uniformly on \([-\Delta, \Delta]/c\).

After each new frame, we recover the number \( M_{\text{loc}} \) of active sources in that frame using the estimates \( \hat{S}_k \) given in (8). We define a source \( \hat{S}_k \) as active in the frame \( t \) if its power is greater than a threshold \( \alpha \), i.e. if
\[ \sum_{\omega} |\hat{S}_k(t, \omega)|^2 \geq \alpha. \] (5)
The K-means algorithm is then applied on \( B \) to obtain \( M_{\text{loc}} \) new cluster centers \( \hat{\delta}_{k(\text{loc})} \). To reduce the number of iterations needed by the K-means, it is useful to initialize the centers to the current values \( \hat{\delta}_k, k = 1, \ldots, M_{\text{loc}} \). The new centers \( \hat{\delta}_{k(\text{loc})} \) are further used to update the estimates \( \hat{\delta}_k \) with the following rule
\[ \hat{\delta}_k = (1 - \mu)\hat{\delta}_k + \mu\hat{\delta}_{k(\text{loc})} \] (6)
where \( i(k) \) denotes the index \( m \) of the closest \( \hat{\delta}_{k(\text{loc})} \) of \( \hat{\delta}_k \). \( \mu \in (0, 1) \) is an adaption factor close to zero. This adaptive procedure provides reliable estimates of the delays \( \delta_k \).

4. Demixing

We can continuously separate the mixture \( X(t, \omega) = A(\omega)S(t, \omega) \) using the estimate \( A(\omega) \) built according to (3) using the current estimates \( \hat{\delta}_k \).

4.1. Anechoic environment

The mixing model (1) holds and if \( M < N \), it constitutes an underdetermined linear system. In order to account for the information given from all microphones, a convenient solution consists of considering the pseudo-inverse of \( A(\omega) \)
\[ A^+(\omega) = (A^H(\omega)A(\omega))^{-1}A^H(\omega). \] (7)
Then the source estimate
\[ \hat{S}(t, \omega) = A^+(\omega)X(t, \omega) \] (8)
minimizes the observation error \( \|X(t, \omega) - A(\omega)S(t, \omega)\|_2 \). Note that it is equivalent to a delay-sum preprocessing: the output of a delay-sum beamformer is given by \( A^H(\omega)X(t, \omega) = A^H(\omega)A(\omega)S(t, \omega) \), so that \( \hat{S}(t, \omega) = A^+(\omega)X(t, \omega) \) inverses the acoustic mixing and the delay-sum preprocessing.

4.2. Echoic environment

The propagation model (1-2) is not accurate enough to allow a clean separation. Demixing according to (8) can-
cells the direct path of interfering sources but their reflections remain. However, the direct path in most cases carries more power than the reflections. Thus, the estimated sources with low power can actually originate from reflections. Therefore, we found it useful to post process the source estimates $\hat{S}(t, \omega)$ given in (8). We keep only the sources with magnitudes greater than a relative threshold, i.e. we set $\hat{S}_i(t, \omega) = 0$ if
\[ |\hat{S}_i(t, \omega)| < \beta \max_{m=1,\ldots,M} |\hat{S}_m(t, \omega)| \]  
(9)
where $\beta \in (0, 1)$. The method was found to be more robust for $\beta$ close to 1. However, it yields a higher distortion since the points $(t, \omega)$ where $m$ sources overlap will be treated as if only one were active: there will be $m-1$ missing sources.

5. Tests and Results

We tested the adaptive demixing method on synthetic mixtures simulating an anechoic environment and on echoic mixtures recorded in a car interior as well. All signals were excerpts from the TIMIT speech database and had 4 second durations. In both cases, the sampling rate was 12 kHz and the spectrogram was computed using half-overlapping frames of 512 samples weighted with the Hanning window. The buffer $B$ had length 100 and the adaptation factor was set to $\mu = 1/10$. For our tests, $\alpha$ was set to 50. Microphone spacing was set to $\Delta = 3$ cm. As the separated signals are permuted estimates of the original sources, one must re-order the $\hat{S}_i$, $i = 1, \ldots, M$. This is done by assigning
\[ \hat{S}_k = \arg \max_{\hat{S}_i} \langle \hat{S}_i, \hat{S}_k \rangle \]  
(10)
where $\langle X, Y \rangle = \sum_{t, \omega} X(t, \omega)\overline{Y}(t, \omega)$ is the scalar product in $L^2(\mathbb{Z}^2)$. To recover the contributions of each source and the distortion due to the algorithm, we project $\hat{S}_k$ on the space spanned by the original sources $\hat{s}_i$, the projection error being the distortion [8]. Formally, we assume that the sources are orthogonal and we consider the decomposition
\[ \hat{S}_k = \langle \hat{S}_k, \hat{s}_k \rangle \hat{s}_k + \sum_{l \neq k} \langle \hat{S}_k, \hat{s}_l \rangle \hat{s}_l > \hat{S}_l + D_k. \]  
(11)
The Signal to Interference Ratio (SIR) is defined as the mean over $k = 1, \ldots, M$ of
\[ -10 \log_{10} \frac{\sum_{l \neq k} \langle \hat{S}_k, \hat{s}_l \rangle^2}{\langle \hat{S}_k, \hat{s}_k \rangle^2} \]  
(12)
and the Signal to Distortion Ratio (SDR) is defined as the mean over $k = 1, \ldots, M$ of
\[ -10 \log_{10} \frac{\|D_k\|^2}{\|\hat{S}_k - D_k\|^2}. \]  
(13)
The results reported below show the gain provided by our method, i.e. the differences of SIR/SDR with and without (considering $\hat{S}_k = X_1$ for all $k$) processing. Each of them is averaged over 20 separation tests. We do not include the first second of data in order to give the method time to learn the mixing parameters. We report here the median and the standard deviation (std) values.

5.1. Anechoic environment

Using the model (1-2), we simulated acoustic anechoic conditions for a number of mixing configurations. Firstly, we assess the adaption ability of the presented method by simulating two moving sources. Three mixtures were generated with the model (3) and varying DOAs. Figure 1 shows the evolution of the actual and estimated DOAs over time. Table 1 gives demixing performances for 2, 3 and 4 sources. The sources are placed so that their DOA have a constant spacing of $\pi/4$. In order to maintain a relatively consistent separation difficulty, the $M$ speakers are separated with $N = M + 1$ microphones. The results show a better separation with fewer sources. We also studied the influence of the DOA difference between two sources. The results shown in Table 2 exhibit the SIR gains for one source located at $-\pi/4$ and the second source arriving from $-\pi/4 + \pi/2, \ldots, -\pi/4 + \pi/12$. Three microphone are used. Lastly, table 3 shows separation results for a varying number of microphones using two sources located at $-\pi/3$ and $\pi/3$. As expected, tables 2 and 3 demonstrate that the best SIR reduction is obtained when the sources are far away from each other and when more microphones are available.

Over all tests, the SDR exhibited little variation with $-2.56$ dB median and 0.88 standard deviation.
Table 2: SIR improvement vs. DOA difference.

<table>
<thead>
<tr>
<th>SIR Gain (dB)</th>
<th>$\pi/2$</th>
<th>$\pi/3$</th>
<th>$\pi/4$</th>
<th>$\pi/6$</th>
<th>$\pi/12$</th>
</tr>
</thead>
<tbody>
<tr>
<td>median</td>
<td>36.2</td>
<td>28.5</td>
<td>25.6</td>
<td>28.6</td>
<td>25.9</td>
</tr>
<tr>
<td>std</td>
<td>7.6</td>
<td>7.2</td>
<td>6.1</td>
<td>4.6</td>
<td>2.7</td>
</tr>
</tbody>
</table>

Table 3: SIR improvement vs. Number of Microphones.

<table>
<thead>
<tr>
<th>SIR Gain (dB)</th>
<th>$N = 3$</th>
<th>$N = 4$</th>
<th>$N = 5$</th>
<th>$N = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>median</td>
<td>36.4</td>
<td>34.4</td>
<td>37.8</td>
<td>38.1</td>
</tr>
<tr>
<td>std</td>
<td>7.2</td>
<td>5.6</td>
<td>5.1</td>
<td>7.4</td>
</tr>
</tbody>
</table>

5.2. Echoic Environment

We tested the method on 20 mixtures of two sources played by artificial heads in a car. The shrinkage factor was set to $\beta = 1/2$. The two heads were placed at the driver and front passenger seat positions. An array of 6 microphones was positioned at the rearview mirror with a distance of $\Delta = 3$ cm between each microphone. Since our mixing model does not account for reflections, the separation performance is worse than in the anechoic case. Table 4 shows that in real conditions, more microphones provide better separation results. Because of the shrinkage (9), the distortion is higher than in anechoic conditions with $-12.9$ dB median and 0.54 standard deviation.

Table 4: SIR improvement vs. Number of Microphones.

<table>
<thead>
<tr>
<th>SIR Gain (dB)</th>
<th>$N = 3$</th>
<th>$N = 4$</th>
<th>$N = 5$</th>
<th>$N = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>median</td>
<td>17.0</td>
<td>17.8</td>
<td>17.7</td>
<td>19.9</td>
</tr>
<tr>
<td>std</td>
<td>4.4</td>
<td>2.3</td>
<td>4.3</td>
<td>3.4</td>
</tr>
</tbody>
</table>

6. Conclusion

We have presented an efficient on-line algorithm for separation of audio sources. The identification of the mixing parameters lies on a sparsity assumption of the time-frequency representation of the sources. Our experiments with digitally mixed speech sources and real recordings demonstrate a good separation quality. In contrast to other sparse decomposition methods, even at time-frequency points where they overlap, the sources can be separated. However, our algorithm can not handle degenerate mixtures. Future works include automatic tuning of the adaptation and threshold parameters and the improvement of the instantaneous delay estimates with a weaker assumption on the sources.

7. Acknowledgements

We would like to thank Dr. Klaus Linhard and Prof. Wolfgang Minker for numerous discussions during the preparation of this work. This research was partially supported by the european project Hoarse.

8. References