14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

TOWARDS UNDERDETERMINED SOURCE RECONSTRUCTION FROM A CLAP–AND–PLAY BINAURAL LIVE RECORDING

Pau Bofill, Enric Monte

Dept. d'Arquitectura de Computadors, UPC
Campus Nord, Mòdul D6, Jordi Girona, 1-3
08034 Barcelona, Spain
pau@ac.upc.edu

Dept. de Teoria del Senyal i Comunicacions, UPC
Campus Nord, Mòdul D5, Jordi Girona, 1-3
08034 Barcelona, Spain
enric@gps.tsc.upc.edu

ABSTRACT

The goal of our current research is to be able to separate a few audio sources from the signals of two microphones, using a separate recording of each player clapping their hands. The separation is performed in the frequency domain, where speech and music signals are mostly sparse. Being underdetermined, the separation is performed in two steps. In the first step, the clapping is used to estimate the transfer function from each source to each microphone. In the second step, the sources are reconstructed using Second Order Cone Programming (SOCP).

Our experiments show moderatly good results for synthetic mixtures (11.5dB average SNR) and poor results for the real case (2.2dB). This paper points out some of the issues that make this task a difficult one, and shows some experimental analysis of why this is so.

1. THE CLAP–AND–PLAY SEPARATION APPROACH

A binaural live recording of $N > 2$ audio sources is an instance of an underdetermined convolutive mixture,

$$\mathbf{x}(t) = \mathbf{h}_1(t) * s_1(t) + \cdots + \mathbf{h}_N(t) * s_N(t) = \mathbf{h}(t) * \mathbf{s}(t), \quad (1)$$

with $\mathbf{x}(t) = [x_l(t)\ x_r(t)]'$ the left and right channels of the mixture, $s_j(t)$ source signal j, $\mathbf{h}_j(t) = [h_{lj}(t)\ h_{rj}(t)]'$ the impulse response of the room from source j to each microphone, and $*$ the convolution operator. That is, the full matrix of impulse responses $\mathbf{h}(t)$ matricially convolved with the vector $\mathbf{s}(t)$ of source signals.

In the short time frequency domain the above equation can be approximated by

$$\begin{aligned} \mathbf{X}(k,\tau) &\approx \mathbf{H}_1(k)S_1(k,\tau) + \cdots + \mathbf{H}_N(k)S_N(k,\tau) \quad (2)\\ &= \mathbf{H}(k)\mathbf{S}(k,\tau), \quad (3)\end{aligned}$$

with k the frequency bin and τ the frame number. Then, $\mathbf{S}(k,\tau)$ is the column vector of sources and $\mathbf{H}(k)$ is the whole mixing matrix. The above system is a linear mixture for each frequency bin. Hereafter, we omit the bin and frame indices, except when required.

The Blind Source Separation (BSS) problem [1] consists of finding an estimate $\hat{\mathbf{H}}$ of $\mathbf{H}$ and an estimate $\hat{\mathbf{S}}$ of $\mathbf{S}$ using only the information in $\mathbf{X}$. BSS in the frequency domain is intrinsically subject to the scaling and permutation ambiguities. When

$$\hat{\mathbf{H}}\hat{\mathbf{S}} = \mathbf{H}\mathbf{S} = \mathbf{X} \quad (4)$$

holds, the scaling ambiguity is avoided by observing the estimated sources at the microphones [2]. That is, for source j, $\hat{\mathbf{H}}_j\hat{S}_j \approx \mathbf{H}_j S_j$ if the estimates $\hat{\mathbf{H}}$ and $\hat{\mathbf{S}}$ are good. We therefore define $\mathbf{Y}_j = \mathbf{H}_j S_j$, the stereophonic version of source j as recorded at the microphones and $\hat{\mathbf{Y}}_j = \hat{\mathbf{H}}_j\hat{S}_j$ our estimate of $\mathbf{Y}_j$. Our goal, then, is to make $\hat{\mathbf{Y}}_j$ as close as possible to $\mathbf{Y}_j$.

For the underdetermined case (more sources than microphones), the separation procedure can be formulated in two steps: the inference of the mixing matrix, and the reconstruction of the sources.

In the first step, in order to estimate $\hat{\mathbf{H}}$ we ask each player to clap their hands in turn, one at a time, thus avoiding the permutation ambiguity. For each source j, the resulting signal $\mathbf{Y}_j^{clap} = \mathbf{H}_j S_j^{clap}$ is a stereophonic clapping sound. Due to the nature of the clapping sound, $\mathbf{Y}_j^{clap}$ is rich in frequency content and it can be used to estimate the impulse response by dividing the right and left channels, as follows:

$$Z = \frac{Y_{rj}^{clap}}{Y_{lj}^{clap}} = \frac{H_{rj}S_j^{clap}}{H_{lj}S_j^{clap}} = \frac{H_{rj}}{H_{lj}}. \quad (5)$$

Then, we define $\hat{\mathbf{H}}_\mathbf{j}$ as

$$Mag(\hat{H}_{lj}) = \frac{1}{\sqrt{1+\alpha^2}} \quad Phase(\hat{H}_{lj}) = \delta/2 \quad (6)$$

$$Mag(\hat{H}_{rj}) = \frac{\alpha}{\sqrt{1+\alpha^2}} \quad Phase(\hat{H}_{rj}) = -\delta/2, \quad (7)$$

whith $\alpha = Mag(Z)$, the magnitude, and $\delta = Phase(Z)$, the phase. In this way, without loss of generality, $\hat{\mathbf{H}}_\mathbf{j}$ is defined to have unit length, and balanced, zero-average phase. In the following the above operation is called *normalization*, and we write

$$\hat{\mathbf{H}}_j = \mathscr{N}(\mathbf{Y}_j^{clap}). \quad (8)$$

In the second step, given $\hat{\mathbf{H}}$, the reconstruction of the sources needs some additional assumption because the system is underdetermined. The usual assumption is the sparsity of the sources. In our case, we assume that the magnitudes of the sources are laplacian and their phases uniform, and a maximum likelihood formulation leads to the following optimization problem,

$$\begin{aligned} &min_{\hat{\mathbf{S}}} \textstyle\sum_j Mag(\hat{\mathbf{S}}_j)\\ &\text{subject to} \quad \mathbf{X} = \hat{\mathbf{H}}\hat{\mathbf{S}}. \end{aligned} \quad (9)$$

Supported by the Ministry of Education of Spain and the European Union (FEDER funds) under contract TIN2004-07739-C02-01
Supported by contract TIN2005-08852

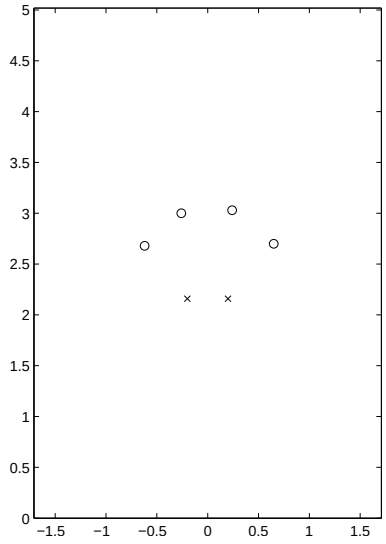

Figure 1: Room size and location of the sources (circles) and microphones (crosses), in meters. Sources and microphones are 0.9m and 1.03m high, respectively. The room is 3m high

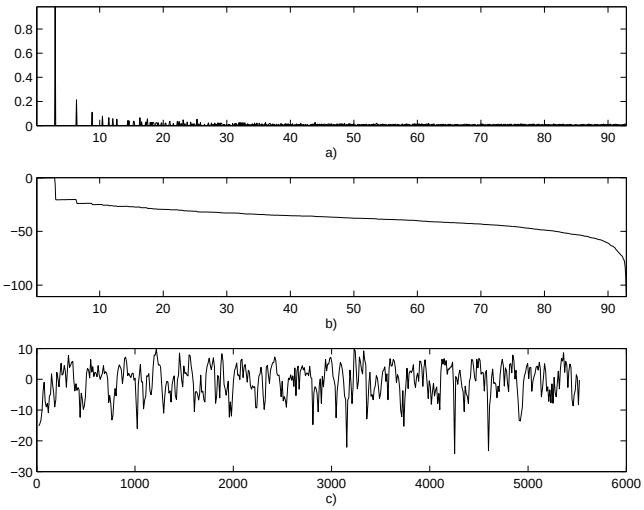

Figure 2: Synthetic impulse response from source $j = 4$ to microphone $i = r$. a) $h_{ij}(t)$ (t in sec.) b) Reverberation curve. $T_{60} = 89$ms. c) $20 * log(Mag(H_{ij}(f))$ (f in Hz)

This problem is an instance of Second Order Cone Programming (SOCP) [3]. Details of the maximum likelihood formulation can be found in [4, 5], and the particularization of SOCP to magnitude minimization is described in [6].

Our approach stems from [7, 4], and in some respects it is similar to the work in [8] and references within.

2. SEPARATION EXPERIMENTS: THE SYNTHETIC VS THE REAL CASE

2.1 Experiment I: four speakers in a simulated office room

We first ran a synthetic experiment, encouraged by the results in [5]. We used the impulse response simulator in [9] to synthetize $\mathbf{h}(t)$. The setting of the synthetic room is shown in Fig. 1 (the actual dimensions of our office). We used a reflection coefficient $r = 0.99$ that yielded a reververation time $T_{60} \approx 90$ms. The impulsional response was then truncated to $l_h = 93ms$ (1024 samples at $fs = 11.025Hz$). The resulting signal is shown in Figure 2.

The unmixed sources were 4 speech utterances (1 male, 3 female) recorded at close distance with an unexpensive dynamic microphone and a standard sound card, downsampled to $fs = 11025$Hz. Time domain convolution was used to generate the mixture $\mathbf{x}(t)$, and an $l_w = l_h$ point short time FFT transform was used with a Hanning window to produce $\mathbf{X}$, with a 70% time overlap between adjacent frames.

The mixing matrix was then normalized $\hat{\mathbf{H}} = \mathcal{N}(\mathbf{H})$ to simulate the loss of scale information and the sources were reconstructed by SOCP. Back in the time domain, results were compared in sensor space using the following signal to error ratio (SER),

$$SER(\hat{\mathbf{y}}, \mathbf{y}) = min_{\beta,t_0} 10 \log \frac{||\beta \hat{\mathbf{y}}(t - t_0)||^2}{||\beta \hat{\mathbf{y}}(t - t_0) - \mathbf{y}(t)||^2}. \quad (10)$$

Results are shown in Table 1. The subjective intelligibility of the separated signals was good, since the background noise was mainly a distorted, unintelligible mix of the other sources.

Signal $\hat{y}_{r4}$ yielded the worst results. Througout the analysis experiments of section 3, we will use this signal as the Worst Case representative of the Synthetic setting (WCS).

2.2 Experiments II and III: four speakers live in the actual office room

We then proceeded to the real case. Experiment II consisted of four people (3 male, 1 female) speaking simultaneously in a setting geometrically equal to Fig. 1 (that is, *in* our office). Speakers and microphones were actually set at the locations signaled in the figure, and a couple of condenser mycrophones were used with a good quality sound card. Again, signals were downsampled to $fs = 11025Hz$. The analysis and resynthesis parameters were the same as in Experiment I.

The clapping sounds were produced with a couple of plastic school rules, because clapping hands saturated the microphones, and $\hat{\mathbf{H}}$ was estimated as $\hat{\mathbf{H}}_j = \mathcal{N}(\mathbf{Y}_j^{clap})$ (Eqn. 8). T_{60} measured directly on the clapping sounds was again around 90ms. The resulting signal is shown in Figure 3. When we ran the SOCP software, the separation was very poor.

Experiment III was the same as Experiment II, but each speaker was recorded separately, thus providing the $\mathbf{y}_j(t)$'s for a numerical comparison of the results, with $\mathbf{x}(t) = \sum_j \mathbf{y}_j(t)$. According to the superposition principle, this artificial mixture should be equivalent to the simultaneous recording. Results are shown in Table 1. For signal $\hat{y}_{l2}$ the performance seems quite good, but still the intelligibility is quite bad. We call this signal the Best Case representative of the Live setting (BCL).

In the next section we report the experiments performed to analyze the reason for these poor results.

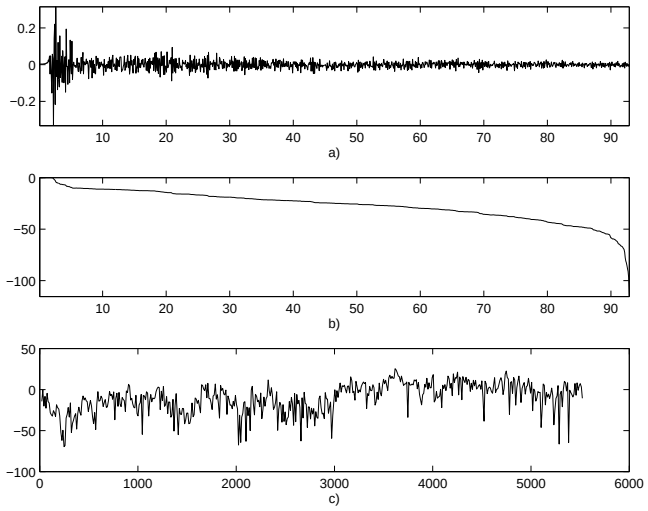

Figure 3: Clapping sound from source $j = 2$ to microphone $i = 1$. a) $y_{ij}^{clap}(t)$ (t in sec.) b) Reververation curve. $T_{60} =$ 89ms c) $20 * log(Mag(Y_{ij}^{clap}(f))$ (f in Hz)

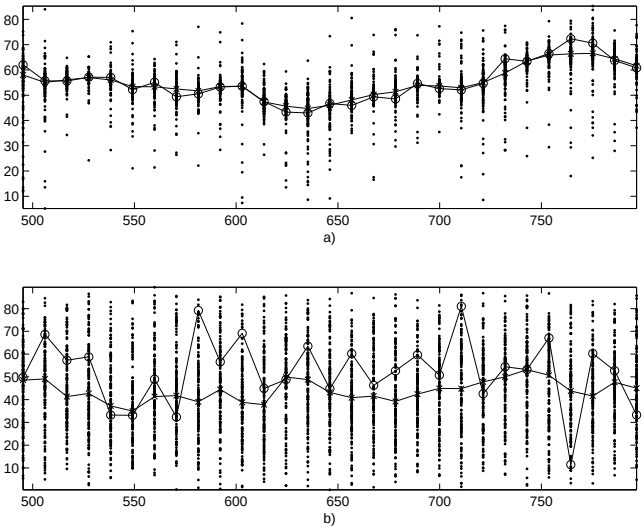

Figure 4: Variability of the local mixing angles $\gamma_j(k,\tau)$ (in degrees), for j corresponding to a) the WCS signal in Experiment IV and b) the BCL signal in Experiment V, for the central frequency range (in Hz). Circled lines correspond to $\hat{\mathbf{H}}$, scattered dots correspond to frames for all k, and crossed lines show the median of the latter

3. EXPERIMENTAL ANALYSIS

3.1 Analysis and resynthesis

The analysis and resynthesis procedure was checked by transforming $\mathbf{x}(t)$ into $\mathbf{X}(k,\tau)$ and back to the time domain, $\tilde{\mathbf{x}}(t)$. The SER was 706.5dB in the context of Experiment I and 707.1dB in the context of Experiment III. Thus, the analysis and resynthesis procedure worked fine.

The problem with the short-time fourier transform of convolved sources is that adjacent frames overlap, thus introducing time aliasing. In order to measure this effect, we took the transforms of the source and the impulse response separately and multiplied them in the frequency domain, $\breve{\mathbf{X}} = \mathbf{HS}$, using a zero padded window of length $\breve{l}_w = l_w + l_h$ (cyclic convolution). Back in the time domain, when comparing $\breve{\mathbf{x}}(t)$ with $\mathbf{x}(t)$ the average SER was only 13.4dB. The impact of the time-domain aliasing was large indeed, leaving little gain margin for the separation procedure. This measure could not be repeated for Experiment III, since the original sources are not available.

3.2 Experiments IV and V

In order to see the best possible performance of the reconstuction, we then fed the SOCP optimizer with the *local* mixing matrix *computed at each frame*. That is, in Experiment IV we used $\hat{\mathbf{H}}_j(k,\tau) = \mathcal{N}(\mathbf{Y}(k,\tau))$ instead of the normalized synthetic mixing matrix, and in Experiment V we used $\hat{\mathbf{H}}_j(k,\tau)$ instead of the normalized clapping sound. Results are shown in Table 1. (11.8 and 5.2dB, respectively).

No improvement was observed in Experiment IV (11.8dB) because the synthetic mixture is static and the local mixing matrices shouldn't differ much from $\hat{\mathbf{H}}$. But in Experiment V the improvement was significant (2.2 to 5.2dB). Notice, in particular, that the BCL signal reached 14.3dB but, of course, this result is artificial because in a real situation the $\mathbf{Y}(k,\tau)$ are not available.

3.3 Variability of the local mixing matrices

In order to study the variability of the local mixing matrices, for source j we define

$$\gamma_j = \tan^{-1}\left(\frac{H_{rj}}{H_{lj}}\right). \tag{11}$$

Figure 4 is a plot of $\gamma_j(k,\tau)$ when j corresponds to the WCS and BCL signals, respectively.

As expected, the contrast between the two is quite clear, but the variability of the synthetic case was larger than expected, probably due to the time-domain convolution and windowing. For Experiment IV, the curve of the median follows closely the curve of the synthetic mixing matrix, whereas for Experiment V the mixing matrix produced by the clapping sound is much more erratic.

Quantitatively, with $\mu(k)$ the mean of $\gamma_j(k,\tau)$ over all frames τ and $\sigma(k)$ the corresponding deviation, Table 1 shows the average μ and σ over all frequency bins k for the WCS and BCL signals. The mean is related to the geometry of the mixing, but there is an obvious difference in deviation from one case to the other.

4. DISCUSSION

In this paper we have presented our current approach towards the separation of underdetermined live recorded sources. The separation procedure has two steps: the estimate of the mixing matrix and the reconstuction of the sources. On the one side, this paper analyses the feasibility of estimating the mixing matrix using a separate clapping sound, from the location of each source. On the other side, the paper evaluates the performance of SOCP for the reconstruction of the sources for different mixing matrix estimates.

Table 1: Min, max and average SER(dB) for the different experiments (see text)

	Synthetic			Live		
	min	max	avg	min	max	avg
	Exp. I			Exp. III		
$SER(\hat{\mathbf{y}},\mathbf{y})$	5.7	20.8	11.5	0.2	8.2	2.2
$SER(\tilde{\mathbf{x}},\mathbf{x})$			706.5			707.1
$SER(\breve{\mathbf{x}},\mathbf{x})$			13.4			
	Exp. IV			Exp. V		
$SER(\hat{\mathbf{y}},\mathbf{y})$	8.3	17.4	11.8	0.8	14.3	5.2
$\mu(k)$			53.5			38.6
$\sigma(k)$			9.7			18.8

The experiments presented here show a comparison between a synthetic context based on a simulated impulse response, and a live context based on a clapping sound. Unfortunately, although the synthetic results were good, the live results were really poor.

The source separation of live recordings is difficult for the following reasons:

1. Sources are not punctual, and the location of the clapping is not exact, which leads to estimates of the wrong impulse response.
2. Sources are not static (the speakers move). The mixing matrix is different from frame to frame, a fact that was actually demonstrated in section 3.3.
3. There is background noise, affecting both the clapping and the actual recordings. This should explain some of the variability of the live local mixing matrices.
4. For short time FFT, the convolution theorem is just an approximation, $X \approx HS$, because of time aliasing between adjacent frames. This was shown in section 3.1.
5. For underdetermined systems, even though Eqn. 4 holds, SOCP doesn't guarantee that $\hat{\mathbf{H}}_j\hat{S}_j \approx \mathbf{H}_j S_j$, not even when $\hat{\mathbf{H}}_j \propto \mathbf{H}_j$, because the decomposition is not unique. That means that even with a good estimate of the mixing matrix, the reconstruction may not be so good. This is ilustrated by the results in the synthetic context.
6. Analysis and resynthesis must be done with care both to avoid clicking at the frame junctions and to reduce musical noise. As shown in section 3.1 this was successfully done.
7. The finite precision of the computations may introduce additional variability.

Experiments VI and V, as compared to I and III, make use of the source signals, which would be unavailable in the real situation, but they show how good the separation might be using the local mixing matrices. The conclusion, though, is that the wrong estimate of the clap-based mixing matrix is not the only reason for the lack of success.

Finally, we hope that the issues presented in this paper will provide a fruitfull discussion at the special session on underdetermined sparse audio source separation. Further results will be presented at the conference.

REFERENCES

[1] Jutten C. and Herault J., "Blind Separation of Sources, an Adaptive Algorithm Based on Neuromimetic Architecture", in *Signal Processing*, Vol 24, No 1, pp 1-10, 1991.

[2] Asano F. and Ikeda S., "Evaluation and Real-Time Implementation of Blind Source Separation System using Time-Delayed Decorrelation", in *Proc. ICA'2000*, pp 411-415, 2000.

[3] Lobo M.S., Vandenberghe L., Boyd S. and Lebret H., "Applications of Second-order Cone Programming", in *Linear Algebra and its Applications*, 284, pp 193-228, 1998.

[4] Bofill P., "Underdetermined Blind Separation of Delayed Sound Sources in the Frequency Domain", in *Neurocomputing, Special issue: Evolving Solution with Neural Networks*, Fanni A. and Uncini A. (Eds), Vol 55, Issues 3-4, pp 627-641, October 2003.

[5] Bofill P. and Monte E., "Underdetermined Convoluted Source Reconstruction using LP and SOCP, and a Neural Approximator of the Optimizer", in *Proc. ICA'2006*, LNCS, vol. 3889, Springer-Verlag, pp. 569-576, March 2006.

[6] Winter S., Hiroshi S. and Makino S., "On Real and Complex Valued l_1-norm minimization for overcomplete blind source separation", in *IEEE Workshop on Apps. of Sig. Proc. to Audio and Acoustics*, pp. 86-89, October 2005.

[7] Bofill P. and Zibulevsky M., "Underdetermined Blind Source Separation using Sparse Representations", in *Signal Processing*, 81 (2001), pp 2353-2362, 2001.

[8] A. Blin, S. Araki, and S. Makino, "Underdetermined blind separation of convolutive mixtures of speech using time-frequency mask and mixing matrix estimation", in *IEICE Trans. Fundamentals*, vol. E88-A, no. 7, pp. 1693-1700, July 2005.

[9] McGovern S.G., http://www.steve-m.us/code/fconv.m, paper available at http://www.steve-m.us/rir.html