GEOMETRIC MULTICHANNEL COMMON SIGNAL SEPARATION WITH APPLICATION TO MUSIC AND EFFECTS EXTRACTION FROM FILM SOUNDTRACKS

Juan José Burred, Pierre Leveau

Audionamix
114, Avenue de Flandre
75019 Paris, France
{juan.jose.burred, pierre.leveau}@audionamix.com

ABSTRACT

We address the task of separation of music and effects from dialogs in film or television soundtracks. This is of interest for film studios wanting to release films in new, previously unavailable languages when the original separated music and effects track is not available. For this purpose, we propose several methods for common signal extraction from a set of soundtracks in different languages, which are multichannel extensions of previous methods for center signal extraction from stereo signals. The proposed methods are simple, effective, and have an intuitive geometrical interpretation. Experiments show that the proposed methods improve the results provided by our previously proposed methods based on basic filtering techniques.

Index Terms— source separation, common signal extraction

1. INTRODUCTION

Having access to the separated audio tracks of an audio mixture is essential for the exploitation of the audio content items. Here, we are concerned with the analysis and separation of film, television or video soundtracks, which are the composite of several well-identified tracks: the dialogs, the sound effects (background noises, foley) and the music. To release foreign-language versions of a given film or television series, the music and effects (M+E) track is provided to the dubbing studios, where the local dialogs are recorded. Unfortunately, these M+E tracks are sometimes not available, or have been lost altogether. This is especially the case for old film material. In this context, there is an interest in extracting the M+E track from an existing master.

We address the problem of extracting the M+E track from a set of soundtracks of the same film in different languages. The goal is to allow studios to release films in new, previously unavailable languages when the separated M+E track is not available. From a signal processing point of view, this problem is equivalent to extracting the common signal between several input signals. We suppose that the masters are available only in mono. The problem is formulated here as a multichannel (more than two input channels), underdetermined (more sources than mixtures) source separation problem. Apart from our previous work on the subject [1], we believe this is a new application of multichannel source separation.

Note that an alternative technique for our purposes would be to apply a voice extraction algorithm to the individual input mixtures. However, this kind of approach is arguably not mature enough to provide usable results. Instead, we take advantage of the multiplicity of already available foreign versions to recast the task as a common signal extraction problem.

In the stereo case, the extraction of the common (center) signal between the two channels has been widely addressed. Approaches like DUET [2], or ℓ_1-norm minimization via the *shortest-path* (SP) algorithm [3] provide good results, and have applications in singing voice or lead instrument separation. The approach proposed in this paper is to extend this kind of approaches to the common signal extraction between more than two input tracks. We present three new algorithms that rely on a geometric analysis of the inter-channel spectral amplitude ratios, and compare them with our own previous approaches based on basic filtering techniques. In Section 2, we will formalize the problem of common signal extraction. Then, in Section 3, the algorithms will be introduced. The first approach (termed "cone", Sect. 3.2) is inspired by the energy allocation method of DUET. The second approach (N-SP, Sect. 3.3) is an N-dimensional generalization of the SP algorithm, and the third one (N-SP-SUB, Sect. 3.4) is a redefinition of bi-dimensional SP into subspaces of a larger-dimensional space. All algorithms are simple, effective and fully deterministic. The two latter ones allow, additionally, the direct extraction of the dialogs. Experiments and results will be described in Section 4.

2. PROBLEM STATEMENT

We formulate the common signal extraction as a linear multichannel source separation problem. This is an underdetermined separation problem: the separation system has N input mixtures, corresponding to the N input soundtracks, denoted by $m_1(t), \dots, m_N(t)$, generated from $N+1$ sources, corresponding to N dialog (voice) tracks $v_1(t), \dots, v_N(t)$ and one M+E (background) track $b(t)$. The generative model is thus the following:

$$\left\{ \begin{array}{ccccc} m_1(t) & = & k_1 b(t) & + & v_1(t) \\ \dots & & & & \\ m_N(t) & = & k_N b(t) & + & v_N(t) \end{array} \right. , \qquad (1)$$

which can be recast as the following linear mixture model in matrix form:

$$\begin{pmatrix} m_1(t) \\ m_2(t) \\ \dots \\ m_N(t) \end{pmatrix} = \begin{pmatrix} k_1 & 1 & 0 & \dots & 0 \\ k_2 & 0 & 1 & \dots & 0 \\ \vdots & & & \ddots & \\ k_N & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} b(t) \\ v_1(t) \\ v_2(t) \\ \dots \\ v_N(t) \end{pmatrix} . \qquad (2)$$

The gain coefficients $k_1, \dots, k_N$ are straightforward to estimate in a preprocessing step by measuring the amplitudes of the input channels in portions where no dialog is present in any language. Thus, we can assume for simplicity $k_1 = k_2 = \ldots = k_N = 1$ and we can define the mixing matrix $\mathbf{A}$ of size $N \times (N+1)$ as

$$\mathbf{A} = (\mathbf{w}|\mathbf{I}), \tag{3}$$

where $\mathbf{w} = (1, 1, \ldots, 1)^T$ denotes an N-element column unit vector, $\mathbf{I}$ denotes an $N \times N$ identity matrix and | denotes horizontal concatenation. This formulation differs from a general linear mixture model in the fact that one source, $b(t)$, is always assumed to be located at the center, and in the sparsity of the mixing matrix.

If we state the problem in a sparse representation context, we suppose that the sources are represented in a space where a few of their elements are non-zero. In that case, the elements of each source have a weak probability to be common to one of the other sources. The elements of the representation that are close to $\mathbf{w}$ will be components most probably belonging to the common signal, the others will be allocated to the channel-specific signals. This principle will be exploited in the methods presented here.

The use of the linear model of Eq. 1 implies that the M+E tracks have to be identical among all foreign language versions, up to a gain factor. This is a simplification, since this is not always the case in reality. This study does not address the equalization differences that can be applied to the M+E tracks, nor the temporal desynchronization issues that occur when the soundtracks have been digitalized from magnetic tapes. It is thus supposed that a preprocessing has been applied to compensate these issues and to approximate the linearity of the mixing. A preprocessing block aimed at aligning and equalizing input tracks was presented in our previous work [1].

3. ALGORITHMS

In this work we present the comparative evaluation of 5 methods for extracting the common signal $b(t)$, and in some cases the dialog signals $v_i(t)$, from the linear mixture model of Eq. 2. Two of the methods were already presented in a previous work [1] and will be summarized in the following section. The three other approaches are novel: the first (cone) is based on an extension of the DUET algorithm [2] and the two others (N-SP, N-SP-SUB) on an extension of the SP algorithm [3]. Other variants have been developed using other signal processing frameworks [4, 5].

For all algorithms, a time-frequency (t-f) representation, in our case a Short-Time Fourier Transform (STFT), is performed on every input signal. It is widely known that the STFT increases the representation sparsity of speech and music signals, and is thus beneficial for geometric approaches such as ours, relying on the clustering of t-f bins around certain directions. More specifically, the STFT increases the disjointness of the mixture, which can be measured in terms of approximate W-Disjoint Orthogonality (WDO) [2], meaning that most t-f bins are assumed to be contributed mainly by a single source. The STFT yields a tensor representation of the multichannel input signal of dimension $T \times F \times N$, where T is the number of time frames and F the number of frequency bins. The algorithms that are described here aim at generating the t-f representation of the common signal, and eventually the channel-specific signals. Because of the linearity of the STFT, the mixture model of Eq. 2 is still valid in the t-f domain. The proposed algorithms operate on the collection of $F \times T$ vectors of N-dimensional t-f points.

The geometric methods will be graphically illustrated for $N = 3$, in which case the second to last columns of matrix $\mathbf{A}$ (Eq. 3) are the canonical vectors defining the orthogonal axes $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ spanning the space and the vector $\mathbf{w}$ (first column) corresponds to the bisector vector of the first quadrant, equidistant from all three space axes (see Figs. 1(a) - 1(c)).

3.1. Previous approaches: minimum and median filtering

In [1], we presented two methods for common signal extraction based on basic filtering techniques. Both were based on assuming a high sparsity and statistical independence of the speech signals. This means that, with a high probability, a certain bin of the mixture at a certain frame will belong to the M+E track. The probability distribution among the channels will be very peaky around the M+E value, with outliers when a dialog is present. A median filter of the STFT amplitudes, robust to outliers, is thus an appropriate choice, which constitutes our first method.

An alternative is to simply take the minimum among the STFT amplitudes. For that, one can assume that, given several observations of the same t-f bin, the most likely to correspond to the common M+E track will be the one with the smaller energy. In spite of the simplicity of this assumption, the minimum method works reasonably well, especially for lower number of channels, as was confirmed by the experiments.

3.2. Closest source energy allocation (“cone method”)

This method relies on the full allocation of each t-f point of the N-dimensional input signal to the closest signal based on their relative amplitudes on each channel. This relates to the DUET method when only the amplitudes are taken into account to compute the direction of the point, the delay being discarded [6]. In DUET, a certain maximum angle around the central vector $\mathbf{w}$ defines the tolerance about the non-orthogonality of the t-f support of the common signal. Here, we extend that notion replacing the tolerance angle by a tolerance *solid angle* between the amplitude vector of a given t-f point $\mathbf{v}$ and the central vector $\mathbf{w}$. In other words, the t-f points that are allocated to the common signal lay inside a cone (for $N = 3$) or inside a hypercone (for $N > 3$) with a given aperture angle θ and whose center axis is defined by $\mathbf{w}$ in the N-dimensional space (Fig. 1(a)). Analytically, a t-f point $\mathbf{v}$ is assigned to the common signal given an aperture angle θ if

$$\arccos \frac{\langle \mathbf{w}, \mathbf{v} \rangle}{\|\mathbf{w}\| \|\mathbf{v}\|} < \theta. \tag{4}$$

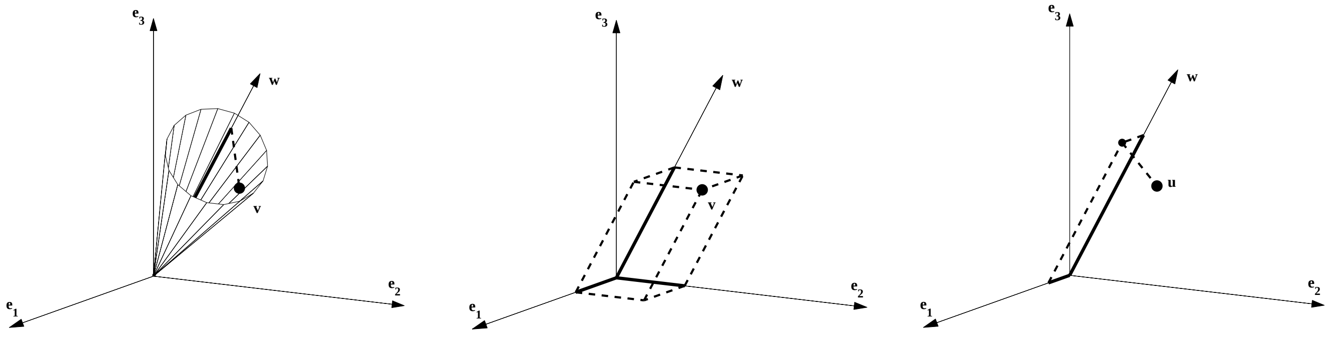
It should be noted that the maximum angle that can be set is $\arccos \frac{1}{\sqrt{N}}$, which is the angle between a canonical vector and $\mathbf{w}$. Obviously, the performance of this method will depend on the aperture angle θ, which will be tested as a parameter in the experimental evaluations of Sect. 4.

Once selected by Eq. 4, instead of taking the magnitude of vector $\mathbf{v}$ as the contribution to the common source, as was done in the original DUET method [6], we take the orthogonal projection of $\mathbf{v}$ upon the central axis of the cone $\mathbf{w}$, denoted by $P_\mathbf{w}\{\mathbf{v}\}$, in order for the distance to the center to have an influence on the energy allocation:

$$P_\mathbf{w}\{\mathbf{v}\} = \langle \mathbf{w}, \mathbf{v} \rangle \frac{\mathbf{w}}{\|\mathbf{w}\|} = \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} u_i \right) \mathbf{w}. \tag{5}$$

3.3. N-dimensional shortest path (N-SP)

This method relies on an extension of the SP algorithm as proposed in [3]. In the original method, every bi-dimensional t-f bin point is allocated to the 2 closest sources, and the contribution to each source



(a) Cone method. The cone section denotes the maximum tolerance angle for a bin to be assigned to the center source.

(b) N-SP: N-dimensional shortest path decomposition.

(c) N-SP-SUB: subspace projection onto a plane, followed by a bi-dimensional shortest path decomposition.

Fig. 1. Illustration of the geometric separation methods for $N = 3$. Vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ represent the input mixtures and output dialogs, central vector $\mathbf{w}$ represents the output M+E track. The contribution of the t-f bin to each output source is denoted by the bold segments on the axes.

is computed using a shortest-path decomposition along the axes defined by those two sources. A shortest-path decomposition is a non-orthogonal projection consisting on finding the set of vectors along the considered axes that sum up to the point $\mathbf{v}$ under consideration.

In its N-dimensional version (N-SP), the first step of this method is to create the vector set S composed of the $N-1$ vectors from the set $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ closest to $\mathbf{v}$ plus the $\mathbf{w}$ vector, creating the $N \times N$ reduced mixing matrix $\mathbf{A}_\rho$. Once selected, w is decomposed onto the N selected vectors, by inverting the determined linear separation sub-problem:

$$\begin{cases} \mathbf{c}_{i \in S} & = & \mathbf{A}_\rho^{-1}\mathbf{v} \\ \mathbf{c}_{i \notin S} & = & 0 \end{cases} \tag{6}$$

where $\mathbf{c}_{i \in S}$ is the vector of contributions to the sources whose index i correspond to the axes selected in set S. The projection corresponding to $\mathbf{w}$ is the amplitude of the common signal for the given t-f point, and the values of the other vector components are the amplitude of the channel-specific signals. Thus, this method outputs both the M+E and the dialogs.

Geometrically, the N-SP method corresponds to decomposing a vector $\mathbf{v}$ into the contributions along the edges of a parallelepiped for $N = 3$ (see Fig. 1(b)) and of a parallelotope for $N > 3$. It is easy to show that, when extracting the M+E track, the N-SP method is equivalent to the min method.

3.4. N-dimensional shortest path with subspace projections (N-SP-SUB)

The last proposed method is also related to the bi-dimensional SP algorithm, but from a different point of view. Instead of generalizing the dimensionality of the decomposition, we propose as an alternative to recast an N-dimensional SP decomposition into a lower-dimensional SP decomposition by projecting the original data into subspaces of lower dimensionality $2 \leq M \leq N-1$. By choosing an appropriate target dimensionality M, we gain control of the trade-off between interferences from other sources and amount of artifacts introduced by the zeros in Eq. 6. The orthogonal projection of an N-dimensional vector $\mathbf{v}$ onto a lower-dimensional subspace $\mathcal{T}$ of dimension M spanned by the columns of the $N \times M$ matrix $\mathbf{S}_\rho$ is given by

$$P_{\mathcal{T}}\{\mathbf{v}\} = \mathbf{S}_\rho(\mathbf{S}_\rho^T\mathbf{S}_\rho)^{-1}\mathbf{S}_\rho^T\mathbf{v}^T. \tag{7}$$

An illustration of this idea is shown in Fig. 1(c), where the original 3-dimensional bin has been orthogonally projected onto the bi-dimensional subspace (plane) spanned by its two closest vectors, and then subjected to a SP decomposition. In general, after subspace projection, the vector will be subjected to an M-SP decomposition, as described in the previous section.

4. RESULTS

In order to evaluate the 5 presented methods, a collection of 15 sets of synchronized soundtrack excerpts were created: 5 of them containing 3 languages, 5 containing 4 languages and 5 containing 5 languages (Spanish, French, Italian, Japanese and German). Each set of mixtures was created by linearly mixing a short M+E fragment with each of the dialog fragments. All sound files were sampled at 48 kHz and, for the STFT analysis, a Hamming window of 80 ms with an overlap factor of 75% was used. A website with sound examples of separation results is available[1].

Evaluation is based on a set of well-known objective measures given the separated sources, namely the Source to Distortion Ratio (SDR), which can be considered the overall performance measure, the Source to Interferences Ratio (SIR), which measures the leakage of the unwanted sources into the desired sources, and the Source to Artifacts Ratio (SAR), which measures the distortion not due to interferences, as implemented in the BSS_EVAL toolbox [7].

The average results across mixture sets and, in the case of the extracted dialogs, across the different languages, are shown in Tables 1 to 3, which show the mentioned measures in dB ± standard deviation. In addition, the performance of the cone method was tested as a function of the aperture angle θ (Fig. 2). For the cone method, the values shown in the tables correspond to the maximum SDR attained in the aperture angle tests, as shown by the curves. For the N-SP-SUB method, the target dimensionality chosen was $M = N-1$.

A first general observation from the results is that the gain in performance of the geometric methods, and in particular of the N-

[1] `http://research.audionamix.com/gcss_icassp2011`

$N = 3$			
Method	SDR (dB)	SIR (dB)	SAR (dB)
Music and effects			
median	3.04 ± 3.18	11.41 ± 4.24	4.34 ± 2.74
cone	7.22 ± 2.03	20.75 ± 2.74	7.48 ± 2.00
N-SP / min	7.30 ± 2.49	20.68 ± 1.87	7.55 ± 2.51
N-SP-SUB	5.28 ± 3.45	12.65 ± 3.99	6.50 ± 3.08
Dialog			
N-SP	15.08 ± 2.33	21.95 ± 2.24	16.15 ± 2.34
N-SP-SUB	10.70 ± 2.34	24.46 ± 1.09	10.92 ± 2.41

Table 1. Average results for $N = 3$ input channels.

$N = 4$			
Method	SDR (dB)	SIR (dB)	SAR (dB)
Music and effects			
median	5.96 ± 3.32	13.72 ± 4.14	7.11 ± 2.97
cone	7.80 ± 1.63	23.06 ± 2.36	7.97 ± 1.65
N-SP / min	8.17 ± 1.86	23.71 ± 1.36	8.31 ± 1.88
N-SP-SUB	9.10 ± 3.13	20.68 ± 3.19	9.47 ± 3.10
Dialog			
N-SP	16.08 ± 2.29	20.96 ± 2.28	17.87 ± 2.27
N-SP-SUB	15.70 ± 2.10	24.81 ± 1.78	16.35 ± 2.15

Table 2. Average results for $N = 4$ input channels.

$N = 5$			
Method	SDR (dB)	SIR (dB)	SAR (dB)
Music and effects			
median	6.35 ± 3.51	17.70 ± 4.58	6.89 ± 3.32
cone	8.33 ± 2.02	22.15 ± 2.20	8.56 ± 2.04
N-SP / min	8.52 ± 1.62	25.06 ± 1.58	8.64 ± 1.64
N-SP-SUB	10.78 ± 2.72	24.48 ± 2.81	10.99 ± 2.71
Dialog			
N-SP	15.91 ± 2.42	19.80 ± 2.42	18.28 ± 2.39
N-SP-SUB	17.19 ± 2.25	23.35 ± 2.13	18.49 ± 2.29

Table 3. Average results for $N = 5$ input channels.

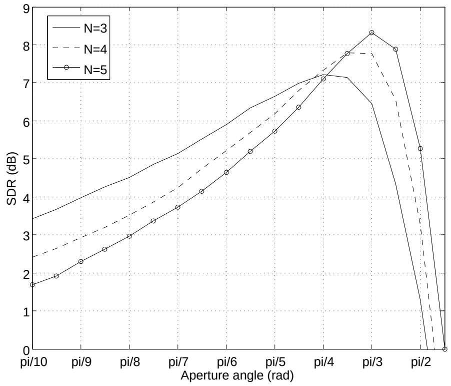


Fig. 2. Average SDR for the cone method, as a function of aperture angle θ.

SP-SUB method, increases with the number of channels N. For $N = 5$, N-SP-SUB clearly performs best, both for M+E and dialog extraction, attaining an SDR of 10.78 dB for M+E and of 17.19 dB for dialog. For $N = 3$ and $N = 4$, performances are more balanced, and the method of choice will depend on the measure to optimize. E.g., for dialog extraction in low dimensionalities, N-SP consistently maximizes SDR and SAR, and N-SP-SUB maximizes SIR.

5. CONCLUSIONS AND OUTLOOK

We have presented three new methods for the extraction of a common signal from an ensemble of linearly mixed signals, and demonstrated their application to the extraction of the music and effect track from film soundtracks. The three new methods (cone, N-SP, N-SP-SUB) are simple and rely on intuitive geometric principles, and were favorably compared to our own previous methods (min, median). For sets of 5 input mixtures, the best performing method is N-SP-SUB. For sets of 3 or 4 input mixtures, the method of choice should be either N-SP or N-SP-SUB, depending on which performance measure is to be maximized: SIR or SAR.

The methods are efficient and can be used to compute large amounts of input data. However, they assume that the mixtures are perfectly linear. In other words, the input soundtracks have to be perfectly synchronized and mutually frequency-equalized. If this is not the case, the input signals can be aligned and equalized with a preprocessing block that was presented in our previous work [1]. However, such compensation only works if the tracks are desynchronized by a fixed delay, and if the compensation filters are time-invariant. This is sometimes not the case, and both delay and equalization can vary in time, giving rise to the problem of *drifting*. Solving this problem is the main goal of our future research, and it will probably require the use of convolutive mixing models.

Another possible future improvement concerns the evaluation method. We have used purely squared-error-based measures. It has been shown that perceptual measures are far more adequate for assessing the quality of source separation and better correlated with human judgements [8]. We plan to use such measures in all our future evaluations.

6. REFERENCES

[1] A. Liutkus and P. Leveau, “Separation of Music+Effects Sound Track from Several International Versions of the Same Movie,” in *128th Convention of the Audio Engineering Society*, London, UK, May 2010.

[2] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[3] P. Bofill and M. Zibulevsky, “Underdetermined blind source separation using sparse representations,” *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

[4] R. Gribonval and S. Lesage, “A survey of sparse component analysis for blind source separation: principles, perspectives, and new challenges,” in *Proceedings of ESANN*, 2006.

[5] T. Melia, S. Rickard, and C. Fearon, “Histogram-based blind source separation of more sources than sensors using a DUET-ESPRIT technique,” in *Proceedings of the 13th European Signal Processing Conference (EUSIPCO05)*, 2005.

[6] H. Shim, J. Abel, and K.-M. Sung, “Stereo music source separation for 3D upmixing,” in *Proc. 127th Convention of the Audio Engineering Society*, New York, USA, October 2009.

[7] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Speech and Audio Processing*, vol. 14 (4), 2006.

[8] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, “Multicriteria subjective and objective evaluation of audio source separation,” in *AES 38th Int. Conf. on Sound Quality Evaluation*, Pitea, Sweden, June 2010.