CONVOLUTIVE COMMON AUDIO SIGNAL EXTRACTION

Pierre Leveau, Simon Maller, Juan José Burred and Xabier Jaureguiberry

Audionamix
114, Avenue de Flandre,
75019 Paris, France
firstname.middlename.lastname@audionamix.com

ABSTRACT

This paper addresses the extraction of a common signal among several mono audio tracks when this common signal undergoes a track-specific filtering. This problem arises in the extraction of a common music and effects track from a set of soundtracks in different languages. To this aim, a novel approach is proposed. The method is based on the dictionary modeling of track-specific and common signals, and is compared to a previous one proposed by the authors based on geometric considerations. The approach is integrated into a Non-Negative Matrix Factorization framework using the Itakura-Saito divergence. The method is evaluated on a synthetic database composed of filtered music and effects tracks, the filters being track-specific, and track-specific dialogs. The results show that this task becomes tractable, while the previously introduced method could not handle track-specific filtering.

Index Terms— audio source separation, common signal extraction, Non-Negative Matrix Factorization

1. INTRODUCTION

For several applications in the content industry, having access to the separated audio tracks of an audio mixture is fundamental. Here, we are concerned with the analysis and separation of film, television or video soundtracks. They are the composite of several well-identified tracks: the dialogs, the sound effects and the music, for which there are specific treatments and uses. For example, the composite track of sound effects mixed with the music (often called the Music and Effects track, or MNE) is interesting to have, in order to release foreign-language versions of a given film or television series. In this case, local dialogs are recorded and mixed with the MNE. For old movies, the MNE tracks are often unavailable, damaged or lost. In this context, there is an interest in extracting the MNE track from an existing master.

We address the problem of extracting the MNE track from a set of soundtracks of the same film in different languages. In this case, the MNE track is common to the several versions, apart from several treatments (equalization, processing) and defects (missynchronization or drifting, noise, clicks, etc.). Following previous work [1, 2], the goal is to achieve realistic applications of this extraction process. In our previous approach [2], the MNE tracks are assumed to be identical (up to a gain factor) among the versions, the dialog tracks being signals specific to the versions. The assumption of the strict equality of the common signal among the versions makes the extraction of the common signal fail in more realistic cases because of the aforementioned treatments and defects. Indeed, a different filtering is often applied to the common signal by the local mix engineer, which is often the case when the MNE track is mixed with the language-specific signal. This effect can be somewhat compensated [1] using pre-computation of filters on signal parts when only the common signal is active, but such method can lack in robustness especially when this type of data is missing. In this study, we integrate the filtering of the common signal as part of the signal model, so that it can be computed in the global estimation process. The new signal model is ascribed to a Non Negative Factorization framework, whose latest developments enable to perform the joint optimization of the track-specific signal model (the version-specific dialogs) and the common signal (the MNE track with version-specific filtering), and does not require any preprocessing step.

The paper is organized as follows. In Section 2, the signal model will be presented. Then in Section 3, an algorithm to estimate the parameters will be proposed. Finally, in Section 4, the results of experimental evaluations will be discussed.

2. SIGNAL MODEL

The proposed signal model involves two main components: the common part, and the specific parts for a number of input channels. For the application case of a MNE extraction problem, the common part is the MNE and the specific parts are the version dialogs.

2.1. Problem statement

In this study, the problem of extracting the common signal, up to a filtering, from several language-specific versions can be formulated as a multichannel convolutive source separation problem [3]. The input signal s of the source separation system is a J-channel signal (J being the number of mono versions). From this signal, we want to extract a set of J MNE signals $\mathbf{s}_{ci}$, which can be interpreted as the result of filtering an original common signal s_c by the channel-specific filters $\mathbf{g}_i$. The channel-specific signals (the version-specific dialogs) $\mathbf{s}_{si}$ are also extracted. Thus, $2J$ sources have to be output from J input channels, which corresponds to an underdetermined source separation problem. The flowchart representing the mixing process corresponding to such problem formulation is shown on Figure 1. This study proposes to estimate all its variables.

2.2. Background: Non-negative Factorization modeling of power spectrograms

The signal model is based on a Non-Negative Factorization (NMF) framework which is briefly summarized here. The chosen metric to evaluate the distortion is the Itakura-Saito (IS) divergence [4]. This model involves the transformation of the input signal s into the time-frequency domain by means of a Short Time Fourier Transform (STFT), yielding a matrix $\mathbf{S}$. The squared modulus of each

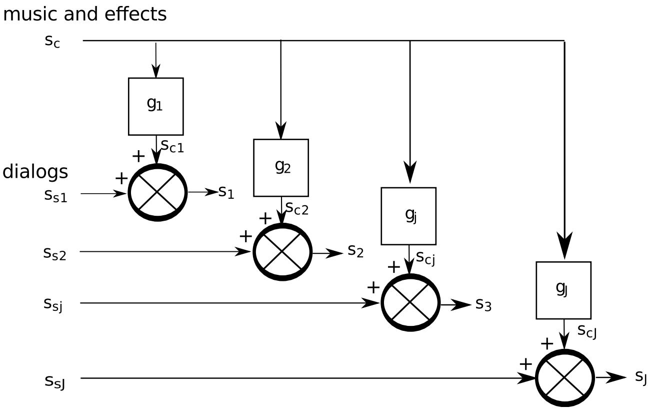

Figure 1: Flowchart of the mixing process. $\mathbf{g}_i$ are the channel-specific filters, $\mathbf{s}_c$ the unfiltered common signal, $\mathbf{s}_{ci}$ are the set of filtered common signals, $\mathbf{s}_{si}$ are the unmixed dialogs in different languages and $\mathbf{s}_i$ are the mixes input to the separation system.

element is then computed to obtain a Power Spectral Density (PSD) observation matrix $\mathbf{X}$, which in this context can be interpreted as variances. The problem of NMF is to find the matrices $\mathbf{W}$ and $\mathbf{H}$ such that

$$\mathbf{X} \simeq \mathbf{M} = \mathbf{WH}, \tag{1}$$

where $\mathbf{M}$ is the PSD matrix of the model. $\mathbf{W}$ and $\mathbf{H}$ have dimensions $F \times K$ and $K \times N$, respectively, and it is desirable that $F \times K + K \times N \ll FN$. Audio signals can often be described by such model, since it adequately accounts for information redundancy in the frequency domain resulting from repetitive events typical in speech and music signals.

The factorization is formulated as the following minimization problem:

$$\{\mathbf{W}, \mathbf{H}\} = \underset{\mathbf{W},\mathbf{H} \geq 0}{\operatorname{argmin}} D_{IS}(\mathbf{X}|\mathbf{M}), \tag{2}$$

where D_{IS} is a matrix cost function involving the IS divergence d_{IS}:

$$D_{IS}(\mathbf{X}|\mathbf{M}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d_{IS}(\mathbf{X}_{(f,n)}|\mathbf{M}_{(f,n)}). \tag{3}$$

The IS divergence is defined as:

$$d_{IS}(x, y) = \frac{x}{y} - \log\frac{x}{y} - 1. \tag{4}$$

This divergence is a good measure for the perceptual difference between two signals, which is explained by its scale invariance: $d_{IS}(\gamma x|\gamma y) = d_{IS}(x|y)$, for a given scalar γ.

The matrix $\mathbf{W}$ obtained after an approximation following this model contains a set of PSDs as its columns, and is commonly called the *dictionary*, whereas $\mathbf{H}$ contains *activations* (weights) of these PSDs across time as its rows. If K is carefully chosen, the PSDs constitute a good characterization of the audio sources involved in the mixture. The choice of K depends on the complexity of the modeled source, and can be optimized using source separation quality evaluation methods.

2.3. Common signal model

The assumptions on which the common signal model are based are the following:

- It is redundant in the time-frequency domain.
- It is common to all the input channels up to a linear filtering. This assumption is less restrictive than in [2] (where only a gain factor was allowed) and enables to make the system more flexible against filtering operations applied to the common signal. This filtering is normally applied by sound engineers and needs to be estimated.

These assumptions yield a multi-channel convolutive NMF model for the variance of the common signal $\mathbf{M}_c$, for which the filters $\mathbf{g}_i$ are specific to each channel. The PSD dictionary $\mathbf{W}$ and the activations $\mathbf{H}$ are common to all filtered MNE signals. The variance model writes, for each channel i:

$$\mathbf{M}_{ci} = \operatorname{diag}(\mathbf{g}_i)\mathbf{W}_c\mathbf{H}_c, \tag{5}$$

where $\mathbf{M}_{ci}$ is the variance of the model of the source in channel i, $\operatorname{diag}(\mathbf{g}_i)$ is a diagonal matrix with $\mathbf{g}_i$ as the diagonal vector and $\mathbf{W}_c$ and $\mathbf{H}_c$ are, respectively, the PSD dictionaries and activations of the common signal.

This filtering can either be seen as a convolutive mix of sources [3] (in which case the filtering is a part of the mixing system) or as a source adaptation filter [5] (in which case the filtering is a part of the source). In the evaluation part, the second option will be considered: the estimation of the $\mathbf{s}_{ci}$ signals will be compared to the known references.

2.4. Specific signal model

The following assumptions concern the model for the language-specific signals:

- It is redundant in the time-frequency domain. In the case of the dialog signals, this is also a reasonable assumption since the voice signals are composed of a limited number of spectral patterns.
- The PSDs at different time frames are different across the channels. This assumption is less restrictive and more realistic than approximate W-disjoint Orthogonality (WDO), exploited in [2], since it implies that the PSD dictionary supports of the modeled sources are disjoint, rather than the time-frequency supports as in WDO.

The variance model writes, for each channel i:

$$\mathbf{M}_{si} = \mathbf{W}_{si}\mathbf{H}_{si}. \tag{6}$$

The disjointness of the dictionary support is not explicitly expressed in the model, but rather implicitly: the dictionaries for each version are not linked.

2.5. Whole signal model

The whole signal variance model is thus the following for each channel:

$$\mathbf{M}_i = \mathbf{M}_{si} + \mathbf{M}_{ci} = \mathbf{W}_{si}\mathbf{H}_{si} + \operatorname{diag}(\mathbf{g}_i)\mathbf{W}_c\mathbf{H}_c. \tag{7}$$

Note that this definition assumes additivity of PSDs, and thus statistical independence between the common signal and the specific signal models.

3. ALGORITHM

3.1. Optimization function

The algorithm involves a minimization of the cost function $\mathcal{J}$ defined as follows:

$$\begin{aligned} \mathcal{J} &= \sum_i D_{IS}(\mathbf{X}_i|\mathbf{M}_{si} + \mathbf{M}_{ci}) && (8) \\ &= \sum_i D_{IS}(\mathbf{X}_i|\mathbf{W}_{si}\mathbf{H}_{si} + \text{diag}(\mathbf{g}_i)\mathbf{W}_c\mathbf{H}_c). && (9) \end{aligned}$$

The optimization algorithm is based on a gradient descent algorithm. The multiplicative update rule framework is used here. Update rules for respective variables are obtained on the basis of the work done in [6] for the channel-specific NMF parts and in [3, 5] for the convolutive NMF part.

For the channel-specific NMF part, the update rules are:

$$\begin{aligned} \mathbf{W}_{si} &\leftarrow \mathbf{W}_{si}.* \frac{(\mathbf{M}_{s_i}^{.-2}.*\mathbf{X}_i)\mathbf{H}_{s_i}^T}{\mathbf{M}_{s_i}^{.-1}\mathbf{H}_{s_i}^T} && (10) \\ \mathbf{H}_{si} &\leftarrow \mathbf{H}_{si}.* \frac{\mathbf{W}_{s_i}^T(\mathbf{M}_{s_i}^{.-2}.*\mathbf{X}_i)}{\mathbf{W}_{s_i}^T\mathbf{M}_{s_i}^{.-1}}. && (11) \end{aligned}$$

For the convolutive NMF-related part, the update rules are:

$$\begin{aligned} \mathbf{W}_c &\leftarrow \mathbf{W}_c.* \frac{\sum_i(\mathbf{M}_{c_i}^{.-2}.*\mathbf{X}_i)\mathbf{H}_c^T}{\sum_i \mathbf{M}_{c_i}^{.-1}\mathbf{H}_c^T} && (12) \\ \mathbf{H}_c &\leftarrow \mathbf{H}_c.* \frac{\sum_i(\text{diag}(\mathbf{g}_i)\mathbf{W}_c)^T\mathbf{M}_{c_i}^{.-2}.*\mathbf{X}_i}{\sum_i(\text{diag}(\mathbf{g}_i)\mathbf{W}_c)^T\mathbf{M}_{c_i}^{.-1}} && (13) \\ \mathbf{g}_i &\leftarrow \mathbf{g}_i.* \frac{\sum_i(\mathbf{X}_i.*(\mathbf{W}_c\mathbf{H}_c))./\mathbf{M}_{c_i}^{.2}}{\sum_i(\mathbf{W}_c\mathbf{H}_c./\mathbf{M}_{ci})}. && (14) \end{aligned}$$

The $.*$ operator denotes an element-wise product. All divisions and exponentiations are also element-wise. The update rules are called successively until a given number of iterations is reached.

3.2. Wiener mask-based separation

Once the variance of each source is estimated (channel-specific sources and common source with adapted filters), the STFT of the sources are obtained using Wiener filtering of the mix STFT with the help of Wiener masks $\mathbf{C}$ on each channel:

$$\begin{aligned} \mathbf{S}_{si} &= \mathbf{C}_{si}.*\mathbf{S}_i = \mathbf{W}_{si}\mathbf{H}_{si}./\mathbf{M}_i.*\mathbf{S}_i && (15) \\ \mathbf{S}_{ci} &= \mathbf{C}_{ci}.*\mathbf{S}_i = (\text{diag}(\mathbf{g}_i)\mathbf{W}_c\mathbf{H}_c)./\mathbf{M}_i.*\mathbf{S}_i. && (16) \end{aligned}$$

The temporal signals s_{si} and s_{ci} of the sources are then obtained through an overlap-add operation applied to the $\mathbf{S}_{si}$ and $\mathbf{S}_{ci}$ STFTs.

4. EXPERIMENTAL STUDY

4.1. Datasets and metrics

The extraction of the common signal has been tested on the dataset used in [2], and also with a dataset composed of filtered MNE tracks, with different filters for each track. The original dataset in [2] is a collection of 15 soundtrack mixes: 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 MNE fragment with each of the dialog fragments. This corresponds to an instantaneous mix in which the MNE tracks are identical and remain unfiltered. The second dataset contains convolutive mixes. To generate them, Butterworth low-pass (cutting respectively at 1000, 2000, and 3000 Hz) and high-pass filters of order 10 (cutting respectively at 500 and 800 Hz) were applied. It should be noted that such filters introduce heavy modifications to the frequency contents of the signal, and thus produced a demanding evaluation scenario. All sound files were sampled at 48 kHz.

Evaluation is based on a well-known objective metric given the separated sources, namely the Source to Distortion Ratio (SDR). This metric gives an overall performance of the source separation algorithm. For the sake of conciseness we will not present the results for 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. The SDR is implemented in the BSS_EVAL toolbox [7].

4.2. Experiments

Several methods have been tested for the performance comparison:

- The two best geometric separation algorithms presented in [2]. The first one (called N-SP) is a multichannel extension of the *shortest-path* (SP) algorithm. The second one (called N-SP-SUB) is a variation thereof which performs a usual 2-dimensional SP after a projection into a subspace. It is expected that these methods will give low results for convolutive mixes: the filtering breaks the assumption that the common signal is theoretically on the bisector vector of the channel space (see [2] for more details).
- The Wiener oracle algorithm, which gives a very optimistic upper bound of what can be achieved if the Wiener masks were optimally estimated.
- The algorithm presented in this paper without the filters $\mathbf{g}_i$ (CNMF, for Common NMF).
- The full algorithm presented in this paper (CCNMF, for Convolutive Common NMF).

For the STFT analysis involved in each of these algorithms, a Hamming window of 80 ms with an overlap factor of 75% was used for all the tested algorithms. A website with sound examples of separation results is available[1].

4.3. Results

The results are presented in Table 1. Results corresponding to the instantaneous dataset are in the columns labeled "SDR inst", and results corresponding to the convolutive dataset in the columns marked "SDR conv". From the results, we can draw the following conclusions:

- For the instantaneous mixes, the CNMF method performs around 3-4 dB below the geometric methods N-SP and N-SP-SUB, which are close to the oracle performance. The quality of extraction of CNMF increases as the number of available languages increases. CCNMF shows even lower results (3dB to 5dB less): it seems that optimizing a filter adds degrees of

[1] `http://research.audionamix.com/ccase_waspaa2011`

	3 versions		4 versions		5 versions	
Method	SDR inst (dB)	SDR conv (dB)	SDR inst (dB)	SDR conv (dB)	SDR inst (dB)	SDR conv (dB)
Music and Effects (MNE)						
N-SP	**7.30 ± 2.49**	-5.53 ± 4.16	8.17 ± 1.86	-14.11 ± 2.95	8.53± 1.62	-20.28 ±5.94
N-SP-SUB	5.28 ± 3.45	-5.15 ± 4.21	**9.10 ± 3.13**	-9.7 ± 4.84	**10.78± 2.72**	-16.03 ± 1.49
CNMF	5.19± 1.78	**5.09 ± 3.8**	5.64 ± 1.99	3.15 ± 3.11	6.04± 2.05	-0.35 ± 2.16
CCNMF	0.08 ± 3.61	2.56 ± 3.34	2.11 ± 3.23	**3.17 ± 3.36**	3.35 ± 2.29	**3.23 ± 3.41**
Oracle	*11.44 ± 1.93*	*17.51 ± 0.56*	*11.44 ± 1.93*	*18.59 ± 2.20*	*11.67 ± 2.07*	*19.43 ± 2.67*
Dialogs						
N-SP	**15.08 ± 8.50**	8.50 ± 3.09	**16.08 ± 2.29**	8.59 ± 3.85	15.91 ± 2.42	9.38 ± 3.88
N-SP-SUB	10.70 ± 2.34	8.49 ± 2.74	15.70 ± 2.10	10.39 ± 3.23	**17.19 ± 2.25**	9.91 ± 3.74
CNMF	12.17 ± 0.61	**12.39 ± 0.1**	12.72 ± 0.74	**12.12 ± 1.61**	12.72 ± 0.83	**11.76 ± 2.22**
CCNMF	1.38 ± 3.99	7.34 ± 2.04	4.22 ± 4.03	8.71 ± 2.74	4.88 ± 2.94	10.28 ± 3.3
Oracle	*16.97 ± 0.36*	*17.51 ± 0.56*	*17.27 ± 0.67*	*18.59 ± 2.20*	*17.10 ± 0.69*	*19.43 ± 2.67*

Table 1: Results (mean ± standard deviation) of Music and Effects and Dialog extraction using several common signal extraction techniques. The N-SP and N-SP-SUB techniques have been presented in [2] and are geometric common signal extraction methods. The CNMF and CCNMF are the methods presented in this paper: CNMF (resp. CCNMF) involves an NMF (resp. filtered NMF) modeling of the common signal.

freedom that make the model converge towards the wrong values.

- For the convolutive mixes, the results of the geometric methods degrade considerably, especially for the MNE track extraction (20 to 40 dB below oracle). Moreover, the performance decreases when the number of versions increases. Indeed, the filters degrade the sparsity assumption taken for these methods.
- The CCNMF method provides MNE signal extraction results that are acceptable for convolutive mixes, and that are slightly increasing when the number of versions available increases. The improvement over geometric methods for convolutive MNE extraction is important: from 7 dB to 20 dB, increasing with the number of versions. Surprisingly, the CNMF method provides the best results, except for the convolutive database case with 5 input versions.
- The dialog extraction task is less sensitive to the used method (geometric or NMF-based); the best results are provided by the CNMF method. The good results of the geometric methods can be explained by the fact that the dialogs components in the time-frequency domain stay close to the component axes [2] when the convolution is applied to the common signal.

The experimental study shows that the proposed algorithm succeeds in handling MNE extraction with convolutive mixes, in which case the geometric methods fail as expected. The filter adaptation does not always provides the best results, but the NMF-based methods always show a significant improvement. There is still a quality gap to fill to get exploitable results (more than 10 dB SDR are required in most applicative contexts).

5. CONCLUSION

In this study, a novel algorithm for common audio signal extraction has been introduced. Its specific feature is to handle the case in which the common signal of interest (a music and effects track) is mixed convolutively with the signal-specific sources (the dialogs). It involves the modeling of the signal-specific sources with NMF models, and the common signal as a possibly filtered NMF model. Experimental results show that this method succeeds when convolutive mixes are processed, while previously introduced geometric methods fail. However there is still a quality gap to bridge to get results appropriate for applicative use, for example for remixing with a new dialog track.

Further studies will add constraints to the filter in order to improve its estimation. The compensation of the synchronization between the different MNE signals will also be addressed to allow processing more realistic cases involving inter-channel drifting. Models for multi-channel versions will also be investigated.

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] J. J. Burred and P. Leveau, “Geometric multichannel common signal separation with application to music and effects extraction from film soundtracks,” in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, Prague, May 2011.

[3] A. Ozerov and C. Févotte, “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[4] C. Févotte, N. Bertin, and J. Durrieu, “Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis,” *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[5] X. Jaureguiberry, P. Leveau, S. Maller, and J. J. Burred, “Adaptation of source-specific dictionaries in non-negative matrix factorization for source separation,” in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, Prague, May 2011.

[6] D. Lee and H. Seung, “Algorithms for non-negative matrix factorization,” *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[7] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.