Multimodal Soft Nonnegative Matrix Co-Factorization for Convolutive Source Separation

Farnaz Sedighin, Massoud Babaie-Zadeh, *Senior Member, IEEE*, Bertrand Rivet, and Christian Jutten, *Fellow, IEEE*

***Abstract*—In this paper, the problem of convolutive source separation via multimodal soft Nonnegative Matrix Co-Factorization (NMCF) is addressed. Different aspects of a phenomenon may be recorded by sensors of different types (e.g., audio and video of human speech), and each of these recorded signals is called a modality. Since the underlying phenomenon of the modalities is the same, they have some similarities. Especially, they usually have similar time changes. It means that changes in one of them usually correspond to changes in the other one. So their active or inactive periods are usually similar. Assuming this similarity, it is expected that the activation coefficient matrices of their Nonnegative Matrix Factorization (NMF) have a similar form. In this paper, the similarity of the activation coefficient matrices between the modalities is considered for co-factorization. This similarity is used for separation procedure in a soft manner by using penalty terms. This results in more flexibility in the separation procedure. Simulation results and comparison with state-of-the-art algorithms show the effectiveness of the proposed algorithm.**

***Index Terms*—Multimodality, blind source separation, nonnegative matrix co-factorization, convolutive mixture, audio-visual speech separation.**

I. Introduction

BLIND Source Separation (BSS) is a challenging problem in signal processing which aims to separate original sources from their mixtures where no information is available about the mixing matrix or the sources except the statistical independence of the original sources and the structure of the mixtures (linear, time instantaneous, convolutive, ...) [1]. An $M \times M$ convolutive mixture, where M is the number of sources and sensors, is modeled as [1]

$$x_i(t) = \sum_{j=1}^{M} \tilde{a}_{ij}(t) * s_j(t), \quad i = 1, 2, ..., M \tag{1}$$

where $x_i(t)$ is the i-th mixture, $s_j(t)$ is the j-th source, $\tilde{a}_{ij}(t)$ is the impulse response filter from the j-th source to the i-th mixture and '$*$' denotes the convolution operator. A usual approach for solving the convolutive source separation problem is to resort to the frequency domain using Short Time Fourier Transform (STFT) [2]–[4]. STFT is usually arranged in a matrix such that the n-th column of this STFT matrix is the Fourier transform of the n-th frame of the signal. The n-th frame of $x_i(t)$, denoted by $x_{i,n}(t)$, is defined as

$$x_{i,n}(t) = x_i(t + (n-1)\tau')\mathscr{W}(t), \quad t \in [0 : \tau] \tag{2}$$

where $\mathscr{W}(t)$ is a finite length window of length τ, and τ' is the amount of the window shift.

Since STFT is a linear transform and by assuming that the filter ($\tilde{a}_{ij}(t)$) time duration is much less than the STFT window length (τ), (1) can be written in the STFT domain as [5]

$$\forall (f, n) \quad x_i(f, n) = \sum_{j=1}^{M} \tilde{a}_{ij}(f) s_j(f, n), \tag{3}$$

where $x_i(f, n)$ is the (f, n)-th element of the STFT matrix of $x_i(t)$, $s_j(f, n)$ is the (f, n)-th element of the STFT matrix of $s_j(t)$, and $\tilde{a}_{ij}(f)$ is the Fourier transform of $\tilde{a}_{ij}(t)$. Different approaches have been proposed for convolutive source separation in the STFT domain, e.g., [1]–[7].

Multimodal nature of natural phenomena can also be exploited for convolutive source separation. Different aspects of a multimodal phenomenon are measured by using different instruments. Each of these measurements is called a modality. For example, human talk is a multimodal phenomenon, with basically two main modalities: audio signal received by ears or microphones, and video signal received by eyes or cameras. Indeed, modalities provide different (but related) signals coming from a single phenomenon [8]. A review on separating the acoustic part of the speech using the corresponding video modality can be found in [7], [9].

Since the modalities are the different recordings of the same phenomenon, they usually have some similarities. Therefore, the joint analysis of the modalities is a powerful tool—in fact a particular approach related to data fusion [8]—for exploiting their similarities in solving different problems. Due to the mentioned similarity among the modalities with the same physical origin, the Nonnegative Matrix Factorization (NMF) of the modalities can have similar parameters, called shared factors [10]. NMF is a decomposition approach in which a matrix $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ with nonnegative elements is factorized as the product of two matrices with nonnegative elements [11] as

$$\mathbf{V} \simeq \mathbf{WH}, \tag{4}$$

Manuscript received May 19, 2016; revised October 31, 2016 and January 22, 2017; accepted February 8, 2017. Date of publication March 8, 2017; date of current version April 19, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Joao Xavier. This work was supported by the Center for International Scientific Studies and Collaboration (CISSC), and also by European Project ERC-2012-AdG-320684-CHESS.

F. Sedighin and M. Babaie-Zadeh are with the Department of Electrical Engineering, Sharif University of Technology, Tehran 1458889694, Iran (e-mail: f_sedighin@ee.sharif.edu; mbzadeh@yahoo.com).

B. Rivet is with the GIPSA-lab, CNRS, Université Grenoble Alpes, Grenoble INP, Grenoble 38000, France (e-mail: bertrand.rivet@gipsa-lab.grenoble-inp.fr).

C. Jutten is with the GIPSA-Lab, CNRS, Université Grenoble Alpes, Grenoble INP, Grenoble 38000, France, and also with the Institut Universitaire de France, Paris 75005, France (e-mail: christian.jutten@gipsa-lab.grenoble-inp.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2017.2679692

where $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K \times N}$. K is usually chosen less than F and N [11]. NMF can be achieved for example by solving [11]

$$\min_{\mathbf{W} \geq 0, \mathbf{H} \geq 0} D(\mathbf{V} \| \mathbf{W}\mathbf{H}), \tag{5}$$

where D measures the divergence between $\mathbf{V}$ and $\mathbf{WH}$.

Simultaneous nonnegative factorization of recordings of multimodal or multichannel[1] datasets with shared factors is a common approach in data fusion and is called Nonnegative Matrix Co-Factorization (NMCF). For multimodal or multichannel datasets with two recordings, NMCF can be achieved by solving [5], [12], [13]

$$\min_{\mathbf{W}_1 \geq 0, \mathbf{W}_2 \geq 0, \mathbf{H} \geq 0} \lambda_1 D(\mathbf{V}_1 \| \mathbf{W}_1 \mathbf{H}) + \lambda_2 D(\mathbf{V}_2 \| \mathbf{W}_2 \mathbf{H}), \tag{6}$$

where λ_i is the weight of the i-th term, $\mathbf{V}_1 \in \mathbb{R}_+^{F \times N}$ and $\mathbf{V}_2 \in \mathbb{R}_+^{F \times N}$ are the recordings of the dataset, $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ is the shared factor that is identical in both $\mathbf{V}_1$ and $\mathbf{V}_2$ and $\mathbf{W}_1 \in \mathbb{R}_+^{F \times K}$ and $\mathbf{W}_2 \in \mathbb{R}_+^{F \times K}$ are the unshared factors.

In [5], NMCF of multichannel audio dataset is used for convolutive source separation. Mixtures that are recorded by different microphones are simultaneously factorized into matrices which are then used for convolutive source separation. In that paper, the shared factors of the recordings of the observed multichannel dataset are assumed to be equal. It leads to a method based on a cost function like (6), in which there is a single matrix $\mathbf{H}$ in both of the two terms[2] (this is called *hard coupling*). In [10], [14], the equality constraint of the shared factors is replaced by the similarity constraint of the shared factors: it leads to methods based on a cost function with an additional penalty term measuring the similarity between the shared factors (this is called *soft coupling*). In [15], this method, called soft NMCF, is also used for convolutive source separation. But the algorithm of [15], to prevent the convergence of the shared factors to zero, requires a normalization of $\mathbf{W}_1$ and $\mathbf{W}_2$ and additional matrix factors.

The algorithms proposed in [5], [15] for convolutive source separation are only based on audio recordings, i.e., only the similarities of the recordings of a *multichannel* audio dataset are considered for convolutive source separation.

Conversely, in this paper, a *multimodal* soft NMCF approach is proposed for convolutive source separation by exploiting both audio and video signals. It is assumed that the audio sources are mixed together convolutively at each microphone and the videos of the speakers are also recorded such that the video information of each speaker is available separately (i.e., non-mixed). The information provided by each video modality is then exploited in the separation of the audio mixtures. Actually, in this paper, the only information that we exploit from the video modality is the surface of the lip opening of the corresponding speaker, extracted from the actual video signal [16]. Since the lip opening of a speaker can have non-zero values during the silence periods (because the lips of the speaker can be open during the silence periods), the derivative of the lip opening signal is used as the second modality. So, we have two modalities: the audio and the derivative of the lip opening signal which for simplicity we call "lip surface" signal throughout this paper.

These two modalities of a single speaker usually have similar time changes. It means that changes in one of these modalities usually correspond to changes in the other modality. Especially, these modalities have nearly the same active or inactive periods. Therefore, it is expected that their activation coefficient matrices have zero elements in nearly the same indices. In this paper, we use this similarity for convolutive source separation. The lip surface modalities are factorized first and then the resulting parameters are used for separating the audio signals. As we show later, in this approach, the problem of the convergence of the shared factors to zero (which exists in the algorithm of [15]) no longer occurs.

The remainder of this article is organized as follows. In Section II, NMCF for convolutive source separation is reviewed. Soft NMCF will be reviewed in Section III. In Section IV, the proposed multimodal soft NMCF algorithm for convolutive source separation is presented and finally, Section V is devoted to experimental results.

II. A Review on NMCF for Convolutive Source Separation

Separating convolutive mixtures using NMCF of multichannel audio recordings has been introduced in [5]. The STFT matrix of the j-th source, denoted by $\mathbf{S}_j$, is a matrix of size $F \times N$ where F is the number of frequency bins and N is the number of time frames. The power spectrogram matrix of the j-th source is defined as $\mathbf{V}_j^s \in \mathbb{R}_+^{F \times N}$ whose elements are

$$v_j^s(f,n) = |s_j(f,n)|^2, \tag{7}$$

where $s_j(f,n)$ and $v_j^s(f,n)$ are the (f,n)-th elements of $\mathbf{S}_j$ and $\mathbf{V}_j^s$, respectively. In [5], it is assumed that the power spectrogram matrix of each individual source can be factorized as

$$\mathbf{V}_j^s \approx \mathbf{W}_j \mathbf{H}_j, \tag{8}$$

where $\mathbf{W}_j \in \mathbb{R}_+^{F \times K}$ is the basis dictionary matrix and $\mathbf{H}_j \in \mathbb{R}_+^{K \times N}$ is the activation coefficient matrix of the NMF of the power spectrogram matrix of the j-th source ($\mathbf{V}_j^s$). It is also assumed that the (f,n)-th element of $\mathbf{S}_j$ (i.e., $s_j(f,n)$) has a complex Gaussian distribution as

$$s_j(f,n) \sim \mathscr{N}_c\left(0,\ \sum_{k=1}^{K} w_j(f,k) h_j(k,n)\right), \tag{9}$$

where $w_j(f,k)$ and $h_j(k,n)$ are the elements of $\mathbf{W}_j$ and $\mathbf{H}_j$, respectively. It is shown in [17] that under the above assumption, and by assuming the mutual independence of the elements across the frequency bins and the time frames, the Maximum Likelihood (ML) estimation of $\mathbf{W}_j$ and $\mathbf{H}_j$ from $\mathbf{S}_j$ is achieved

[1] Multichannel dataset consists of recordings of a phenomenon with several sensors of the *same type* while multimodal dataset consists of recordings of a phenomenon with sensors of *different types*. Several audio recordings of a human speech with different microphones is an example of a multichannel dataset, while audio and video recordings of a human speech is an examples of a multimodal dataset.

[2] Details on NMCF for convolutive source separation will be reviewed in Section II.

by minimizing

$$\begin{aligned}
&-\log P(\mathbf{S}_j|\mathbf{W}_j\mathbf{H}_j)\\
&= -\sum_{n=1}^{N}\sum_{f=1}^{F}\log P\left(s_j(f,n)|0,\sum_{k=1}^{K} w_j(f,k)h_j(k,n)\right)\\
&= NF\log(\pi)+\sum_{n=1}^{N}\sum_{f=1}^{F}\log\left(\sum_{k=1}^{K} w_j(f,k)h_j(k,n)\right)\\
&\qquad + \frac{|s_j(f,n)|^2}{\sum_{k=1}^{K} w_j(f,k)h_j(k,n)}\\
&= \sum_{n=1}^{N}\sum_{f=1}^{F} d_{\text{IS}}\left(|s_j(f,n)|^2\|\sum_{k=1}^{K} w_j(f,k)h_j(k,n)\right)+\text{cst}\\
&= \sum_{n=1}^{N}\sum_{f=1}^{F} d_{\text{IS}}\left(v_j^s(f,n)\|\sum_{k=1}^{K} w_j(f,k)h_j(k,n)\right)+\text{cst}\\
&= D_{\text{IS}}(\mathbf{V}_j^s\|\mathbf{W}_j\mathbf{H}_j)+\text{cst},
\end{aligned}\tag{10}$$

where P denotes the Probability Density Function (PDF), "cst" denotes the terms which are independent of $\mathbf{W}_j$ or $\mathbf{H}_j$ and D_{IS} denotes the Itakura-Saito Divergence between two matrices, defined as

$$\begin{aligned}
D_{\text{IS}}(\mathbf{Y}\|\hat{\mathbf{Y}}) &= \sum_{i,j} d_{\text{IS}}(y(i,j)\|\hat{y}(i,j))\\
&= \sum_{i,j}\frac{y(i,j)}{\hat{y}(i,j)}-\log\frac{y(i,j)}{\hat{y}(i,j)}-1,
\end{aligned}\tag{11}$$

where $y(i,j)$ and $\hat{y}(i,j)$ are the (i,j)-th elements of $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, respectively, and $d_{\text{IS}}(y(i,j)\|\hat{y}(i,j))$ denotes the element-wise Itakura-Saito divergence. Based on (10) and by considering (5), it is clear that the ML estimation of $\mathbf{W}_j$ and $\mathbf{H}_j$ for the j-th source ($\mathbf{S}_j$) is the NMF of its power spectrogram matrix using the Itakura-Saito divergence.

Similar to the above discussion and by assuming that the time durations of the impulse responses of the filters ($\tilde{a}_{ij}(t)$) are much smaller than the STFT window size, the ML estimation of the parameters from the STFT matrix of the i-th mixture, denoted by $\mathbf{X}_i$ of size $F\times N$, is achieved by minimizing (as detailed in [5])

$$\begin{aligned}
&-\log P(\mathbf{X}_i\|\sum_{j=1}^{M}\mathbf{A}_{ij}\mathbf{W}_j\mathbf{H}_j)\\
&= \sum_{n=1}^{N}\sum_{f=1}^{F} d_{\text{IS}}\left(|x_i(f,n)|^2\|\sum_{j=1}^{M}|\tilde{a}_{ij}(f)|^2\sum_{k=1}^{K} w_j(f,k)h_j(k,n)\right)\\
&\quad+\text{cst} = D_{\text{IS}}\left(\mathbf{V}_i^x\|\sum_{j=1}^{M}\mathbf{A}_{ij}\mathbf{W}_j\mathbf{H}_j\right)+\text{cst},
\end{aligned}\tag{12}$$

where M is the number of sources (and also the number of mixtures), $x_i(f,n)$ is the (f,n)-th element of $\mathbf{X}_i$, the matrix $\mathbf{V}_i^x\in\mathbb{R}_+^{F\times N}$ is the power spectrogram matrix of the i-th mixture whose (f,n)-th element is equal to $|x_i(f,n)|^2$, $\tilde{a}_{ij}(f)$ is the Fourier transform of $\tilde{a}_{ij}(t)$ and $\mathbf{A}_{ij}\in\mathbb{R}_+^{F\times F}$ is a diagonal matrix whose (f,f)-th element is $a_{ij}(f,f)=|\tilde{a}_{ij}(f)|^2$.

It worth emphasizing that $\mathbf{W}_j$ and $\mathbf{H}_j$ (the factorization parameters of the j-th source) are the same for all of the mixtures, but $\mathbf{A}_{ij}$'s have different values for each mixture. In [5], convolutive source separation is then achieved by minimizing $\sum_{i=1}^{M}-\log P(\mathbf{X}_i\|\sum_{j=1}^{M}\mathbf{A}_{ij}\mathbf{W}_j\mathbf{H}_j)$ which is equivalent to minimizing the following cost function

$$C=\sum_{i=1}^{M} D_{\text{IS}}\left(\mathbf{V}_i^x\|\sum_{j=1}^{M}\mathbf{A}_{ij}\mathbf{W}_j\mathbf{H}_j\right).\tag{13}$$

The above equation is an NMCF problem of a multichannel dataset for convolutive source separation in which $\mathbf{W}_j$'s and $\mathbf{H}_j$'s are the shared factors and $\mathbf{A}_{ij}$'s are the unshared factors. The parameters $w_j(f,k)$, $h_j(k,n)$ and $a_{ij}(f,f)$ are estimated by minimizing the cost function (13) with respect to $w_j(f,k)$, $h_j(k,n)$ and $a_{ij}(f,f)$. Finally, the j-th source in the i-th mixture is reconstructed using Wiener filtering as [5]

$$\hat{s}_{ij}(f,n)=\frac{a_{ij}(f,f)\left(\sum_{k=1}^{K} w_j(f,k)h_j(k,n)\right)x_i(f,n)}{\hat{v}_i(f,n)},\tag{14}$$

where $\hat{s}_{ij}(f,n)$ is the (f,n)-th element of the reconstructed j-th source in the i-th mixture and $\hat{v}_i(f,n)$ is the (f,n)-th element of $\hat{\mathbf{V}}_i=\sum_{j=1}^{M}\mathbf{A}_{ij}\mathbf{W}_j\mathbf{H}_j$ (note that $\hat{\mathbf{V}}_i\in\mathbb{R}_+^{F\times N}$).

III. A Review on Soft NMCF for Convolutive Source Separation

In [15], the hard coupling approach of the previous section has been modified to a soft coupling, and so a soft NMCF has been proposed for separating multichannel (in fact, stereo) audio datasets. In that paper, the first and the second mixtures (the signals received in the left and the right microphones) are modeled as

$$\begin{aligned}
x_l(t)&=s_1(t)+s_2(t),\\
x_r(t)&=\tilde{a}_1(t)*s_1(t)+\tilde{a}_2(t)*s_2(t),
\end{aligned}\tag{15}$$

where $\tilde{a}_1(t)$ and $\tilde{a}_2(t)$ are the time domain filters for the first and the second sources. So the elements of the STFT matrices of the mixtures are

$$\begin{aligned}
x_l(f,n)&=s_1(f,n)+s_2(f,n),\\
x_r(f,n)&=\tilde{a}_1(f)s_1(f,n)+\tilde{a}_2(f)s_2(f,n),
\end{aligned}\tag{16}$$

where $x_l(f,n)$ and $x_r(f,n)$ are the (f,n)-th elements of the STFT matrices of the mixtures received in the left and the right microphones, respectively, and $\tilde{a}_1(f)$ and $\tilde{a}_2(f)$ are the Fourier transforms of $\tilde{a}_1(t)$ and $\tilde{a}_2(t)$, respectively. Convolutive source separation using soft NMCF is then achieved by minimizing the following cost function with respect to $\boldsymbol{\theta}\triangleq(\mathbf{W}_l,\mathbf{H}_l,\mathbf{W}_r,\mathbf{H}_r,\mathbf{A}_1,\mathbf{A}_2)$ [15]

$$\begin{aligned}
C(\boldsymbol{\theta}) &= D_{\text{IS}}(\mathbf{V}_l^x\|\mathbf{W}_{1l}\mathbf{H}_{1l}+\mathbf{W}_{2l}\mathbf{H}_{2l})\\
&\quad+D_{\text{IS}}(\mathbf{V}_r^x\|\mathbf{A}_1\mathbf{W}_{1r}\mathbf{H}_{1r}+\mathbf{A}_2\mathbf{W}_{2r}\mathbf{H}_{2r})+\lambda_s p(\mathbf{H}_l,\mathbf{H}_r),
\end{aligned}\tag{17}$$

where $\mathbf{V}_l^x\in\mathbb{R}_+^{F\times N}$ and $\mathbf{V}_r^x\in\mathbb{R}_+^{F\times N}$ are the power spectrogram matrices of the mixtures received in the left and the right microphones, $\mathbf{W}_l = [\mathbf{W}_{1l}, \mathbf{W}_{2l}] \in \mathbb{R}_+^{F\times 2K}$, $\mathbf{H}_l = [\mathbf{H}_{1l}^T, \mathbf{H}_{2l}^T]^T \in \mathbb{R}_+^{2K\times N}$, $\mathbf{W}_r = [\mathbf{W}_{1r}, \mathbf{W}_{2r}] \in \mathbb{R}_+^{F\times 2K}$ and $\mathbf{H}_r = [\mathbf{H}_{1r}^T, \mathbf{H}_{2r}^T]^T \in \mathbb{R}_+^{2K\times N}$ where $\mathbf{W}_{1l}$, $\mathbf{W}_{2l}$, $\mathbf{W}_{1r}$, $\mathbf{W}_{2r}$ are of size $F \times K$ and $\mathbf{H}_{1l}$, $\mathbf{H}_{2l}$, $\mathbf{H}_{1r}$, $\mathbf{H}_{2r}$ are $K \times N$ matrices and $\mathbf{A}_1$ and $\mathbf{A}_2$ are diagonal $F \times F$ matrices whose diagonal elements are $a_1(f,f) = |\tilde{a}_1(f)|^2$ and $a_2(f,f) = |\tilde{a}_2(f)|^2$, respectively [15]. $\mathbf{W}_{il}$ and $\mathbf{H}_{il}$ $(i = 1, 2)$ are the NMF parameters of the i-th source received in the left microphone, $\mathbf{W}_{ir}$ and $\mathbf{H}_{ir}$ $(i = 1, 2)$ are the NMF parameters of the i-th source received in the right microphone and $(.)^T$ denotes the matrix transpose operator. $p(\mathbf{H}_l, \mathbf{H}_r)$ is the penalty term which controls the similarity of $\mathbf{H}_l$ and $\mathbf{H}_r$ and λ_s is the weight of the penalty term. The value of λ_s highly affects the performance of the soft coupling algorithm. So in [15] the soft coupling algorithm is executed for different values for λ_s and finally the best result is selected.

As it is clear from (17), the equality constraint (i.e., hard coupling) of $\mathbf{H}_l$ and $\mathbf{H}_r$ of the two mixtures, which was assumed in [5], is replaced by the similarity constraint (i.e., soft coupling) of $\mathbf{H}_l$ and $\mathbf{H}_r$ (it should be noted that in [15], as in [5], it is assumed that $\mathbf{W}_l = \mathbf{W}_r$). In other words, it is no more required that $\mathbf{H}_l$ and $\mathbf{H}_r$ are equal, they only have to be similar. The similarity of the parameters is controlled by the penalty term, $p(\mathbf{H}_l, \mathbf{H}_r)$, which is added to the cost function.

The penalty terms that are used in [15] are $\|\mathbf{H}_l - \mathbf{H}_r\|_1$ and $\|\mathbf{H}_l - \mathbf{H}_r\|_F^2$, where $\|.\|_1$ denotes the sum of the absolute values of a matrix (ℓ_1 penalty term) and $\|.\|_F$ is the Frobenius norm of a matrix (ℓ_2 penalty term). For any $0 < \alpha < 1$, the cost function (17) satisfies the following property: $C(\frac{1}{\alpha}\mathbf{W}_l, \alpha\mathbf{H}_l, \frac{1}{\alpha}\mathbf{W}_r, \alpha\mathbf{H}_r) < C(\mathbf{W}_l, \mathbf{H}_l, \mathbf{W}_r, \mathbf{H}_r)$ [15]. So without any additional constraint, $\mathbf{H}_l$ and $\mathbf{H}_r$ will converge to zero, i.e., to a trivial solution. To avoid this problem, in [15], each column of $\mathbf{W}_l$ and $\mathbf{W}_r$ is normalized to have a unit ℓ_1 norm and then to compensate for the effect of the normalization, the k-th rows of $\mathbf{H}_l$ and $\mathbf{H}_r$ are multiplied by additional parameters $b_{lk} = \sum_f w_l(f,k)$ and $b_{rk} = \sum_f w_r(f,k)$, respectively. In addition, due to possibly different scalings of the activation coefficient matrices, especially when the activation coefficient matrices are extracted from modalities of different types (e.g., audio and video modalities of a speech), a diagonal matrix is multiplied to one of the matrices in the penalty term. So the ℓ_2 penalty term in [15] is

$$p(\mathbf{H}_l, \mathbf{H}_r) = \|\mathbf{B}_l\mathbf{H}_l - \mathbf{S}\mathbf{B}_r\mathbf{H}_r\|_F^2, \tag{18}$$

where $\mathbf{B}_l$ and $\mathbf{B}_r$ are diagonal matrices of size $K \times K$, whose (k,k)-th elements are b_{lk} and b_{rk}, respectively, and $\mathbf{S} \in \mathbb{R}_+^{K\times K}$ is a diagonal matrix to compensate for the potentially scale difference between $\mathbf{H}_l$ and $\mathbf{H}_r$. More details about $\mathbf{B}_l$, $\mathbf{B}_r$ and $\mathbf{S}$ can be found in [15].

Finally, after the estimation of the parameters by minimizing the cost function of (17), the sources are reconstructed using Wiener filtering.

IV. The Proposed Multimodal Soft NMCF Algorithm for Convolutive Source Separation

As mentioned before, audio and lip surface modalities coming from a single speech have some similarities. In particular, changes in one of them usually correspond to changes in the other one. This similarity is shown in Fig. 1. Due to this similarity, it is expected that the activation coefficient matrices resulted from the NMF of the modalities are similar, especially for the entries close to zero (silence periods). This similarity is used in papers such as [14] for speaker diarization.

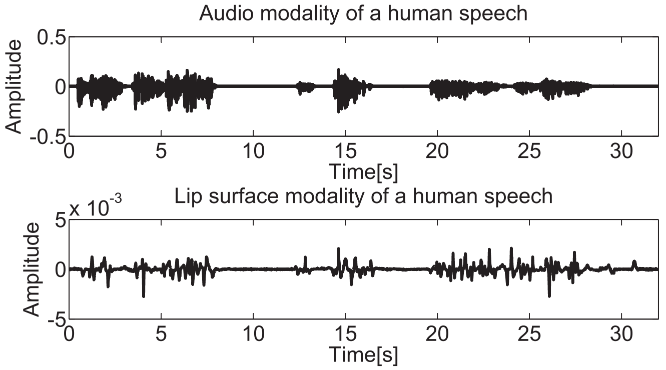

Fig. 1. Time variations of human speech viewed by the two modalities: (top) the audio modality, (down) the lip surface modality of the speaker.

In this paper, we use this similarity along with soft NMCF presented in [15] for convolutive source separation. In the first step of the proposed algorithm, the activation coefficient matrices of the lip surface modalities are extracted by the NMF of the power spectrogram matrices of the lip surface modalities as (it should be noted again that the lip surface modality of each speaker is available separately)

$$\min_{\mathbf{W}_j^v \geq 0, \mathbf{H}_j^v \geq 0} D_{\text{IS}}(\mathbf{V}_j^v \| \mathbf{W}_j^v \mathbf{H}_j^v), \tag{19}$$

where $\mathbf{V}_j^v \in \mathbb{R}_+^{F\times N}$ is the power spectrogram matrix of the lip surface modality of the j-th speaker and $\mathbf{W}_j^v \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H}_j^v \in \mathbb{R}_+^{K\times N}$ are the factorization parameters.

The penalty term in (17), which controls the similarity of the activation coefficient matrices, is broken into the two following terms

$$p(\mathbf{H}_l, \mathbf{H}_r) \rightarrow p_l(\mathbf{H}_l, \mathbf{H}^v) + p_r(\mathbf{H}_r, \mathbf{H}^v),$$

where for a 2×2 mixture, $\mathbf{H}^v = [\mathbf{H}_1^{vT}, \mathbf{H}_2^{vT}]^T$. Each of the rows of $\mathbf{H}^v$ is normalized to have unit ℓ_1 norm. The above penalty terms can also be written as

$$p_l(\mathbf{H}_{1l}, \mathbf{H}_1^v) + p_l(\mathbf{H}_{2l}, \mathbf{H}_2^v) + p_r(\mathbf{H}_{1r}, \mathbf{H}_1^v) + p_r(\mathbf{H}_{2r}, \mathbf{H}_2^v). \tag{20}$$

By the above discussion, in the second step of the proposed algorithm, the following cost function is proposed for convolutive source separation via multimodal soft NMCF

$$\begin{aligned} C_{\text{NEW}}(\boldsymbol{\theta}) = {} & D_{\text{IS}}(\mathbf{V}_l^x \| \mathbf{W}_{1l}\mathbf{H}_{1l} + \mathbf{W}_{2l}\mathbf{H}_{2l}) \\ & + D_{\text{IS}}(\mathbf{V}_r^x \| \mathbf{A}_1\mathbf{W}_{1r}\mathbf{H}_{1r} + \mathbf{A}_2\mathbf{W}_{2r}\mathbf{H}_{2r}) \\ & + \lambda_l p_l(\mathbf{H}_l, \mathbf{H}^v) + \lambda_r p_r(\mathbf{H}_r, \mathbf{H}^v), \end{aligned} \tag{21}$$

where $\boldsymbol{\theta} \triangleq (\mathbf{W}_l, \mathbf{W}_r, \mathbf{H}_l, \mathbf{H}_r, \mathbf{A}_1, \mathbf{A}_2)$. It is assumed that $\mathbf{W}_{1l} = \mathbf{W}_{1r}$ and $\mathbf{W}_{2l} = \mathbf{W}_{2r}$ (as in [5] and [15]), so in the rest of the paper $\mathbf{W}_1$ is used instead of $\mathbf{W}_{1l}$ and $\mathbf{W}_{1r}$, and $\mathbf{W}_2$ is used instead of $\mathbf{W}_{2l}$ and $\mathbf{W}_{2r}$. λ_l and λ_r are the weighs of the penalty terms. The dimensions of the matrices are the same as in Section III. In this paper, it is assumed that $p_l(.) = p_r(.) = p(.)$.

The extension of the above cost function to more than two sources and sensors is straightforward.

Suppose that for the NMF of each of the lip surface modalities, K is set to a positive integer such as κ, so for 2×2 mixtures, $\mathbf{H}^v = [\mathbf{H}_1^{vT}, \mathbf{H}_2^{vT}]^T \in \mathbb{R}_+^{2\kappa \times N}$. Since $\mathbf{H}_1^v$ and $\mathbf{H}_2^v$ correspond to the lip surface modalities of the first and the second speakers, respectively, and by considering (20), it is expected that after the update procedure $\mathbf{H}_{1l}$ and $\mathbf{H}_{1r}$ correspond to the activation coefficient matrix of the first source and $\mathbf{H}_{2l}$ and $\mathbf{H}_{2r}$ correspond to the activation coefficient matrix of the second source. Consequently, after the update procedure, $\mathbf{W}_1$ corresponds to the first source and $\mathbf{W}_2$ corresponds to the second source. Finally, the sources are reconstructed using Wiener filtering (14). It should be noted that in this paper we consider the same κ for both audio and lip surface modalities.

The details of the proposed penalty term and the update rules are discussed in the following subsections.

A. The Proposed Penalty Term

In this paper, the following penalty term is proposed for controlling the similarity of the activation coefficient matrices of the corresponding audio and lip surface modalities:

$$p(\mathbf{H}, \mathbf{H}^v) = D_{\text{MM}}(\mathbf{H}^v \| \mathbf{H}) = \sum_{k,n} \frac{d_{\text{IS}}(h^v(k,n) \| h(k,n))}{h^v(k,n)} =$$

$$\sum_{k,n} \frac{1}{h(k,n)} - \frac{\log h^v(k,n)}{h^v(k,n)} + \frac{\log h(k,n)}{h^v(k,n)} - \frac{1}{h^v(k,n)}, \quad (22)$$

where $D_{\text{MM}}(\mathbf{H}^v \| \mathbf{H})$ is the proposed multimodal penalty term (where MM stands for MultiModal), d_{IS} is the element-wise Itakura-Saito divergence defined in (11), $\mathbf{H}$ denotes $\mathbf{H}_l$ or $\mathbf{H}_r$ and $h(k,n)$ and $h^v(k,n)$ are the (k,n)-th elements of $\mathbf{H}$ and $\mathbf{H}^v$, respectively. The zero valued elements of $\mathbf{H}^v$ are replaced by a very small positive constant ϵ to prevent division by zero.

As mentioned before, the zero valued (very small valued) indices of the activation coefficient matrices of the audio and the lips surface modalities are nearly the same. Thus, the penalty term must take into account this similarity, especially during the silence periods, i.e., for the small valued indices of the activation coefficient matrix of the lip surface modality ($\mathbf{H}^v$). This is done in (22) by weighting each term $d_{\text{IS}}(h^v(k,n) \| h(k,n))$ by $\frac{1}{h^v(k,n)}$. This ensures that $\mathbf{H}$ and $\mathbf{H}^v$ are very similar for the small values, but can be far from similarity for the larger values of $h^v(k,n)$.

Similarity of $\mathbf{H}_l$ and $\mathbf{H}_r$ to $\mathbf{H}^v$ guarantees the similarity of $\mathbf{H}_l$ and $\mathbf{H}_r$. So the proposed cost function for separating 2×2 convolutive mixtures, i.e., (21), can be written as

$$\begin{aligned} C_{\text{NEW}}(\boldsymbol{\theta}) = &D_{\text{IS}}(\mathbf{V}_l^x \| \mathbf{W}_1\mathbf{H}_{1l} + \mathbf{W}_2\mathbf{H}_{2l}) \\ &+ D_{\text{IS}}(\mathbf{V}_r^x \| \mathbf{A}_1\mathbf{W}_1\mathbf{H}_{1r} + \mathbf{A}_2\mathbf{W}_2\mathbf{H}_{2r}) \\ &+ \lambda_l D_{\text{MM}}(\mathbf{H}^v \| \mathbf{H}_l) + \lambda_r D_{\text{MM}}(\mathbf{H}^v \| \mathbf{H}_r). \quad (23) \end{aligned}$$

In addition, since $\mathbf{H}^v$ has been computed in advance and is kept fixed during the update procedure, the problem of the convergence of $\mathbf{H}_l$ and $\mathbf{H}_r$ to zero, which was noted in [15], no longer occurs. Consequently, the matrices $\mathbf{B}_l$ and $\mathbf{B}_r$, used in [15] to compensate for the effect of the normalization of the basis vectors, is not needed in the proposed penalty term.

B. The Update Rules

Similar to [11], [18], in this paper the update rules are derived using a majorization-minimization approach and exploiting auxiliary functions [11]. $G(\mathbf{H}, \mathbf{H}^t)$ is an auxiliary function for $F(\mathbf{H})$ if the following conditions hold [11]

$$G(\mathbf{H}, \mathbf{H}^t) \geq F(\mathbf{H}) \quad G(\mathbf{H}^t, \mathbf{H}^t) = F(\mathbf{H}^t),$$

where $\mathbf{H}^t$ is the point at which the values of $G(\mathbf{H}^t, \mathbf{H}^t)$ and $F(\mathbf{H}^t)$ are the same. Thus $F(\mathbf{H})$ is non-increasing under the update

$$\mathbf{H}^{t+1} = \underset{\mathbf{H}}{\operatorname{argmin}}\, G(\mathbf{H}, \mathbf{H}^t).$$

This is because [11]:

$$F(\mathbf{H}^{t+1}) \leq G(\mathbf{H}^{t+1}, \mathbf{H}^t) \leq G(\mathbf{H}^t, \mathbf{H}^t) = F(\mathbf{H}^t).$$

It means that minimizing the auxiliary function results in minimizing $F(\mathbf{H})$. So finding a proper convex auxiliary function is an important step for deriving the update rules.

Suppose that $\hat{\mathbf{V}}_l = \mathbf{W}_1\mathbf{H}_{1l} + \mathbf{W}_2\mathbf{H}_{2l} \in \mathbb{R}_+^{F \times N}$ and $\hat{\mathbf{V}}_r = \mathbf{A}_1\mathbf{W}_1\mathbf{H}_{1r} + \mathbf{A}_2\mathbf{W}_2\mathbf{H}_{2r} \in \mathbb{R}_+^{F \times N}$, whose (f,n)-th elements are $\hat{v}_l(f,n)$ and $\hat{v}_r(f,n)$, respectively and $v_l^x(f,n)$ and $v_r^x(f,n)$ are the (f,n)-th elements of $\mathbf{V}_l^x$ and $\mathbf{V}_r^x$, respectively. By using a majorization-minimization approach and finding proper auxiliary functions, the update rules are derived as (details are deferred to Appendix A)

$$w_q(f,k) \leftarrow w_q(f,k) \times$$

$$\sqrt{\frac{\sum_n h_{ql}(k,n) \frac{v_l^x(f,n)}{\hat{v}_l^2(f,n)} + \sum_n a_q(f,f) h_{qr}(k,n) \frac{v_r^x(f,n)}{\hat{v}_r^2(f,n)}}{\sum_n \frac{h_{ql}(k,n)}{\hat{v}_l(f,n)} + \sum_n \frac{a_q(f,f) h_{qr}(k,n)}{\hat{v}_r(f,n)}}}, \quad (24)$$

$$h_{ql}(k,n) \leftarrow \sqrt{\frac{h_{ql}^2(k,n) \sum_f w_q(f,k) \frac{v_l^x(f,n)}{\hat{v}_l^2(f,n)} + \lambda_l}{\sum_f \frac{w_q(f,k)}{\hat{v}_l(f,n)} + \frac{\lambda_l}{h_q^v(k,n) h_{ql}(k,n)}}}, \quad (25)$$

$$h_{qr}(k,n) \leftarrow \sqrt{\frac{h_{qr}^2(k,n) \sum_f a_q(f,f) w_q(f,k) \frac{v_r^x(f,n)}{\hat{v}_r^2(f,n)} + \lambda_r}{\sum_f \frac{a_q(f,f) w_q(f,k)}{\hat{v}_r(f,n)} + \frac{\lambda_r}{h_q^v(k,n) h_{qr}(k,n)}}}, \quad (26)$$

$$a_1(f,f) \leftarrow a_1(f,f) \sqrt{\frac{\sum_n h'(f,n) \frac{v_r^x(f,n)}{\hat{v}_r^2(f,n)}}{\sum_n \frac{h'(f,n)}{\hat{v}_r(f,n)}}}, \quad (27)$$

$$a_2(f,f) \leftarrow a_2(f,f) \sqrt{\frac{\sum_n h''(f,n) \frac{v_r^x(f,n)}{\hat{v}_r^2(f,n)}}{\sum_n \frac{h''(f,n)}{\hat{v}_r(f,n)}}}, \quad (28)$$

where $\mathbf{H}' \triangleq \mathbf{W}_1\mathbf{H}_{1r} \in \mathbb{R}_+^{F \times N}$ and $\mathbf{H}'' \triangleq \mathbf{W}_2\mathbf{H}_{2r} \in \mathbb{R}_+^{F \times N}$ with elements $h'(f,n)$ and $h''(f,n)$, respectively and $w_q(f,k)$, $h_{ql}(k,n)$, $h_{qr}(k,n)$, $a_q(f,f)$, $h_q^v(k,n)$ are the elements of $\mathbf{W}_q$, $\mathbf{H}_{ql}$, $\mathbf{H}_{qr}$, $\mathbf{A}_q$, $\mathbf{H}_q^v$ $(q = 1, 2)$, respectively. The parameters are updated sequentially in each iteration until convergence.

The proposed multimodal soft NMCF algorithm for convolutive source separation (for 2×2 mixtures) is summarized in Algorithm 1.

Algorithm 1: Proposed Multimodal Soft NMCF Algorithm.

```
Compute the power spectrogram of each of the lip
  surface modalities ($\mathbf{V}_j^v \quad (j = 1, 2)$).
Compute $\mathbf{H}_j^v (j = 1, 2)$ for a predetermined $\kappa$. ($\kappa$ is the
  number of the rows of $\mathbf{H}_j^v$)
for $k = 1 : \kappa$ do
    $h_j^v(k, n) = \frac{h_j^v(k,n)}{\sum_n h_j^v(k,n)}$
end for
$\mathbf{H}^v = [\mathbf{H}_1^{vT}, \mathbf{H}_2^{vT}]^T$
if $h^v(k, n) = 0$ then
    $h^v(k, n) \leftarrow \epsilon$
end if
Update the parameters sequentially in each iteration
  using (24)–(28) until convergence.
Reconstruct the original sources using (14).
```

V. Experimental Results

In this section, the validity of the proposed algorithm is investigated via experimental results. Pairs of audio and lip surface modalities extracted from human speeches are used for the simulations. The details about extracting the lip surface modalities and recording the audio signals can be found in [16]. Since the sampling frequency of the audio modalities is 16 kHz and the sampling frequency of the lip surface signals is 50 Hz, the lip surface signals are up-sampled by rate of 320 using the "interp.m" function of Matlab. The audio mixtures are artificially created as

$$x_l(t) = s_1(t) + s_2(t)$$
$$x_r(t) = s_1(t - 0.0019) + s_2(t - 0.0031).$$

The duration of the audio signals is 32 sec (512000 samples). It should be noted again that the lip surface signals are not mixed, so each lip surface modality corresponds to a speaker and consequently to an audio signal.

In the first step, the similarity of the activation coefficient matrices of the two modalities coming from a single speech, which is the basic assumption of this paper, is studied. Activation coefficient matrices of audio and lip surface modalities corresponding to a single speech are computed separately. For the simplicity of the comparison, in this experiment κ (i.e., the number of the rows of $\mathbf{H}_i^v, i = 1, 2$) is set equal to 1. So each of the estimated matrices has only one row. In Fig. 2, the estimated activation coefficient matrices of the audio (top) and the lip surface (down) modalities of a speech are shown. As it is seen, the estimated activation coefficient matrices of the audio and the lip surface modalities are similar especially in their zero indices. The simulations regarding the proposed algorithm are presented in the following subsections.

A. Convergence of the Proposed Algorithm

In this subsection, the convergence of the proposed cost function is experimentally investigated. In this experiment, the proposed algorithm is used for separating sources from 2×2 mixtures. The number of the iterations of the proposed algorithm is set equal to 200, κ (the number of the rows of $\mathbf{H}_i^v$) is set equal to 10 and λ_l and λ_r (the weights of the penalty terms) are set equal to 1. The window length of STFT is 1000 samples which is equal to 0.0625 sec. For numerical comparison, the following Signal to Noise Ratio (SNR) is used

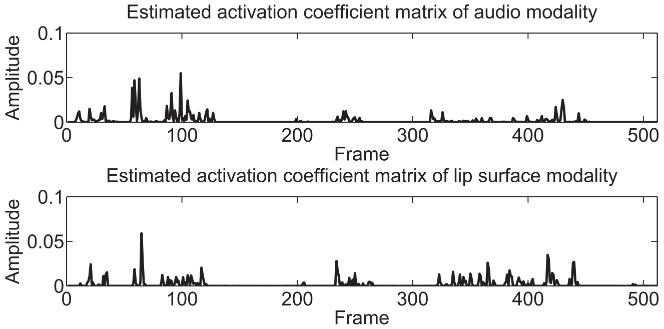

Fig. 2. The estimated activation coefficient matrices of the audio modality (top) and the lip surface modality (down) of a human speech with $\kappa = 1$ (note that there is no mixture for the audio signal and by $\kappa = 1$ the activation coefficient matrices reduce to a simple row).

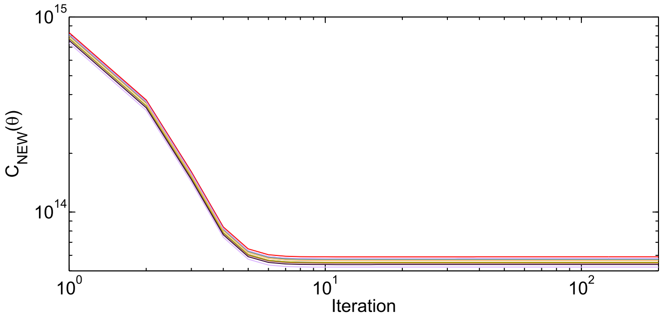

Fig. 3. The proposed cost function for 10 executions of the proposed algorithm.

$$\mathrm{SNR} = 10 \log_{10} \Big(\frac{\sum_{f,n} |s(f,n)|^2}{\sum_{f,n} (|s(f,n)| - |\hat{s}(f,n)|)^2} \Big), \tag{29}$$

where $s(f, n)$ is the (f, n)-th element of the STFT matrix of the original signal and $\hat{s}(f, n)$ is the (f, n)-th element of the STFT matrix of the estimated signal. We also define the average SNR for the first and the second separated sources from 2×2 mixtures as

$$\mathrm{SNR}_{\mathrm{avg}} = \frac{\mathrm{SNR}_1 + \mathrm{SNR}_2}{2}, \tag{30}$$

where SNR_i $(i = 1, 2)$ is the average of the separating SNRs of the i-th separated source from each mixture. The value of the cost function and $\mathrm{SNR}_{\mathrm{avg}}$ for 10 executions of the proposed algorithm for the separation of the sources from 2×2 mixtures are shown in Figs. 3 and 4. It should be noted that the mixtures are the same in each execution but the parameters of the algorithm are initialized randomly. It is clear that, for these executions of the proposed algorithm, the cost function (Fig. 3) and the $\mathrm{SNR}_{\mathrm{avg}}$ curves (Fig. 4) converge to different values depending on the initializations of the parameters.

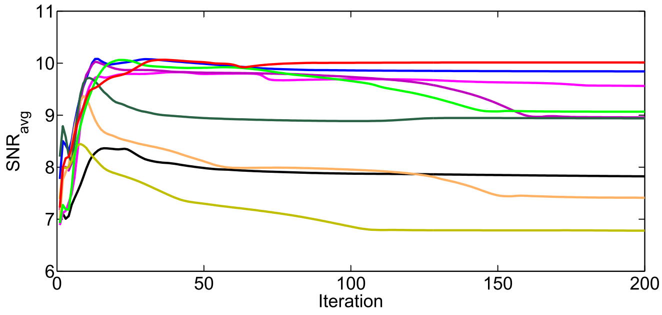

Fig. 4. $\mathrm{SNR}_{\mathrm{avg}}$ (in dB) for 10 executions of the proposed algorithm.

B. Comparison of the Proposed Algorithm with Other Separating Algorithms

To have a comparison between the proposed multimodal algorithm and the algorithms proposed in [5] and [15], several 2×2 convolutive mixtures are separated using the three mentioned algorithms. In this paper, we refer to the algorithm proposed in [5] as the *hard coupling* algorithm and the algorithm proposed in [15] as the *soft coupling* algorithm. The window length of STFT is 1000 samples which is equal to 0.0625 sec, the penalty coefficients λ_l and λ_r are set equal to 1 and κ is set equal to 10 (since κ somehow determines the structure of the algorithm, for a fair comparison, we choose the same κ for all of the algorithms). As mentioned earlier, choosing a proper penalty coefficient (λ_s in (17)) highly affects the performance of the soft coupling algorithm. So, for each mixture, the soft coupling algorithm is executed for $\lambda_s = [0.5, 1, 1.5, 2, 2.5]$ and finally the best result is selected. All of the algorithms are initialized randomly with nonnegative elements. The resulting SNR_1's and SNR_2's are given in Table I, for our proposed algorithm (multimodal soft), the algorithm of [15] (soft coupling) and the algorithm of [5] (hard coupling). Clearly, the proposed multimodal soft NMCF algorithm outperforms the other audio-only algorithms.

For visually demonstrating the quality of the proposed algorithm in source separation, a 2×2 mixture is separated using the proposed algorithm. λ_l and λ_r are set equal to 1 and κ is set equal to 10. The original and the separated signals are shown in Figs. 5 and 6, respectively. The quality of the proposed algorithm in source separation is clear from the results.

The efficiency of the proposed algorithm when the lip surface information is only available for one of the sources (say s_1) is investigated in Table II. For this propose, the coefficients of the penalty terms corresponding to the second source (the second and the fourth terms of (20)) are set equal to zero and $\mathbf{H}_{2l} = \mathbf{H}_{2r}$. Clearly, the SNR performance, although smaller than when the lip surface modality is available for all speakers, is more than the performances of the hard and the soft coupling algorithms, for most mixtures. It should be noted that since the mixtures in Tables I and II are the same and the best results are selected for the soft coupling algorithm, the results for the soft coupling algorithm in both tables are the same.

The efficiency of the proposed algorithm for separating 3×3 mixtures is compared with the hard coupling algorithm. The results are presented in Table III. It is clear from the results that the performance of the proposed algorithm for separating 3×3 mixtures is less than the performance of the proposed algorithm for separating 2×2 mixtures. But the performance of the proposed algorithm, in most cases, is better than the performance of the hard coupling algorithm.

TABLE I
SEPARATION SNR (dB) FOR THE PROPOSED MULTIMODAL SOFT NMCF, THE SOFT AND THE HARD COUPLING ALGORITHMS FOR SEPARATING 20 MIXTURES

#	multimodal soft		soft coupling		hard coupling	
	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$
1	**5.59**	**8.06**	-0.75	3.26	-1.66	1.13
2	**5.36**	**6.85**	-0.22	2.60	-0.77	0.80
3	**3.86**	**4.78**	-0.39	1.44	-1.24	-0.34
4	**9.05**	**7.35**	2.52	-0.04	2.31	-0.44
5	**4.28**	**2.45**	1.42	0.86	1.85	0.33
6	**4.36**	**3.60**	1.51	-0.76	2.46	0.39
7	**6.69**	**2.83**	2.65	-1.21	5.48	-2.40
8	**3.23**	**3.02**	2.13	2.00	0.54	-1.99
9	**5.65**	**1.02**	4.03	-1.76	5.30	-2.82
10	**5.63**	**1.35**	4.79	0.06	4.55	-3.07
11	**6.19**	**5.35**	0.58	1.40	0.94	-2.66
12	-0.21	**4.34**	**0.01**	2.75	-1.18	0.67
13	**4.71**	**6.22**	0.37	1.78	-1.11	-1.80
14	**4.45**	**1.74**	0.03	0.53	2.91	-2.96
15	0.10	**3.58**	**1.51**	2.47	1.32	2.61
16	**7.33**	**7.19**	1.95	0.01	2.51	-2.67
17	**4.80**	**3.27**	2.64	-0.25	3.69	-5.50
18	**4.89**	**4.09**	2.10	0.71	3.20	-1.36
19	**7.20**	**5.51**	2.36	-0.55	2.62	0.90
20	**3.96**	**6.04**	-0.57	1.38	-2.24	-1.47
avg	**4.856**	**4.432**	1.433	0.834	1.574	-1.132

C. Investigating the Effect of λ_l and λ_r

The effect of λ_l and λ_r on the quality of the proposed algorithm is studied in Table IV. The results are the averaged $\mathrm{SNR}_{\mathrm{avg}}$ for 5 different mixtures. It is clear from these results that, generally, the performance of the proposed algorithm is first improved by increasing λ_l and λ_r, but further increase of these penalty coefficients results in a reduction of the performance of the proposed algorithm.

D. Investigating the Effect of κ

In this section, the effect of κ (the number of the rows of $\mathbf{H}_i^v$) on the performance of the proposed algorithm is investigated. In Fig. 7, $\mathrm{SNR}_{\mathrm{avg}}$ averaged over 10 different mixtures is plotted for different values of κ. It is seen that the separation performance is increased with κ, up to $\kappa = 10$, but increasing κ to larger amounts does not highly affect the quality of the proposed algorithm.

E. Soft or Hard Coupling

For investigating the effect of the proposed multimodal soft coupling of the audio and the lip surface modalities, in Table V, the proposed method is compared with the situation when $\mathbf{H}_l = \mathbf{H}^v$ and $\mathbf{H}_r = \mathbf{H}^v$, i.e., with the hard coupling of the activation coefficient matrices of the audio and the lip surface modalities. In this approach, $\mathbf{H}_l$ and $\mathbf{H}_r$ are set equal to $\mathbf{H}^v$ and

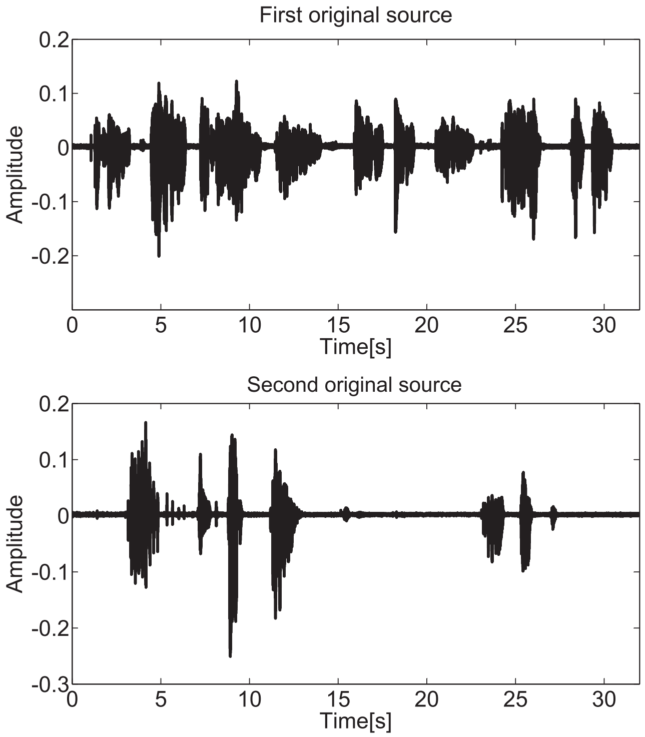

Fig. 5. The original sources of the experiment of Section V-B.

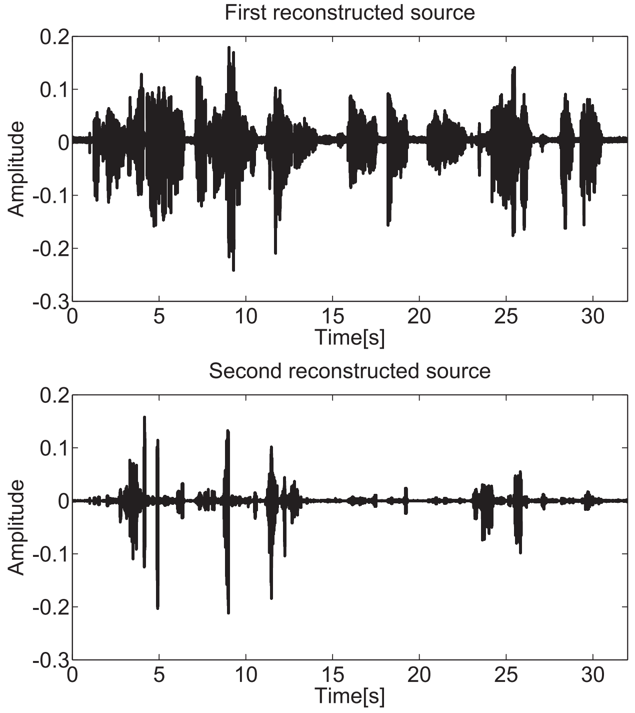

Fig. 6. The reconstructed sources of the experiment of Section V-B.

are kept fixed, that is, only $\mathbf{W}_l$ and $\mathbf{W}_r$ are updated during the update procedure. SNR_1 and SNR_2 for the separated sources using the mentioned approaches are presented in Table V. It is clear from the results that the proposed multimodal soft coupling results in a better separation performance compared with the hard coupling of the activation coefficient matrices of the

TABLE II
SEPARATION SNR (dB) FOR THE PROPOSED MULTIMODAL SOFT NMCF WHEN THE LIP SURFACE INFORMATION IS AVAILABLE ONLY FOR THE FIRST SOURCE, THE SOFT AND THE HARD COUPLING ALGORITHMS FOR SEPARATING 20 MIXTURES

#	multimodal soft		soft coupling		hard coupling	
	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$
1	-1.51	1.30	-0.75	**3.26**	**-0.35**	1.02
2	**-0.22**	1.31	**-0.22**	**2.60**	-0.84	1.04
3	-0.69	1.14	**-0.39**	**1.44**	-1.31	-0.97
4	**2.52**	1.21	**2.52**	-0.04	1.93	**2.47**
5	2.81	0.01	1.42	0.86	**3.39**	**1.55**
6	**3.04**	**1.20**	1.51	-0.76	2.07	0.24
7	**4.97**	-0.36	2.65	-1.21	4.69	**0.01**
8	**3.47**	**2.86**	2.13	2.00	1.98	-3.02
9	4.16	**0.98**	4.03	-1.76	**4.70**	-3.88
10	4.45	-1.16	**4.79**	**0.06**	4.46	-0.55
11	**3.79**	**2.64**	0.58	1.40	0.95	-2.61
12	-0.5	1.09	**0.01**	**2.75**	-1.32	-2.15
13	-0.60	**3.19**	**0.37**	1.78	-0.64	0.81
14	0.28	**1.07**	0.03	0.53	**2.72**	-3.72
15	-1.70	0.82	**1.51**	**2.47**	-0.03	-0.27
16	**3.96**	**2.93**	1.95	0.01	2.60	-1.94
17	2.99	**0.70**	2.64	-0.25	**3.39**	-1.17
18	**3.43**	**0.82**	2.10	0.71	0.43	0.21
19	0.77	**0.01**	**2.36**	-0.55	2.25	-1.78
20	**1.39**	**3.61**	-0.57	1.38	-2.3	-1.65
avg	**1.840**	**1.268**	1.433	0.834	1.438	-0.818

TABLE III
SEPARATION SNR (dB) FOR THE PROPOSED MULTIMODAL SOFT NMCF ALGORITHM AND THE HARD COUPLING ALGORITHM FOR SEPARATING 3 × 3 MIXTURES

#	multimodal soft			hard coupling		
	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\hat{\mathbf{S}}_3$	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\hat{\mathbf{S}}_3$
1	**-0.06**	**2.36**	0.40	-3.43	0.45	**1.66**
2	1.12	1.16	0.15	**1.52**	**2.32**	**1.18**
3	**0.30**	**1.64**	0.01	-1.22	1.45	**1.32**
4	**-1.22**	**1.92**	**-0.32**	-3.69	0.02	-0.35
5	0.53	1.64	-0.24	**0.98**	**3.06**	**1.10**
6	2.12	**0.07**	0.33	**3.23**	-3.31	**1.70**
7	1.50	**0.68**	**0.1**	**2.86**	-2.51	-0.87
8	**2.10**	**1.91**	**0.14**	1.76	-0.44	-0.85
9	**-1.05**	**2.04**	**0.04**	-3.32	-1.08	-0.39
10	**-1.85**	**1.14**	-0.13	-2.76	0.75	**0.76**
avg	**0.349**	**1.456**	0.048	-0.407	0.071	**0.526**

TABLE IV
SNR_{avg} (dB) FOR THE PROPOSED MULTIMODAL SOFT NMCF ALGORITHM VERSUS λ_l AND λ_r AVERAGED OVER 5 MIXTURES

penalty coefficients		λ_r				
		1	1.5	2	2.5	3
λ_l	1	5.607	5.735	5.724	5.707	5.653
	1.5	5.577	5.705	5.690	5.683	5.667
	2	5.580	5.680	5.646	5.616	5.593
	2.5	5.573	5.647	5.599	5.575	5.523
	3	5.548	5.627	5.578	5.564	5.510

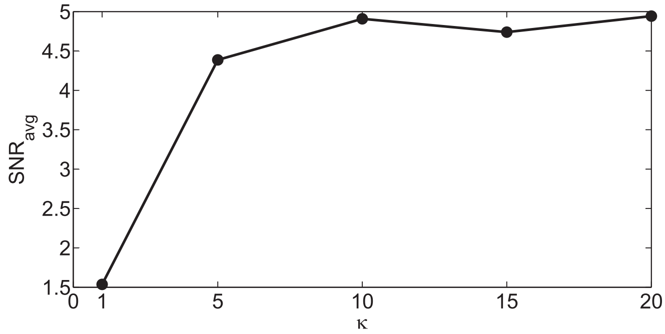

Fig. 7. The averaged SNR$_{\text{avg}}$ for separating 10 different mixtures versus κ (the number of the rows of $\mathbf{H}_i^v$).

TABLE V
SEPARATION SNR (dB) FOR THE PROPOSED MULTIMODAL SOFT NMCF ALGORITHM AND THE APPROACH BASED ON THE HARD COUPLING OF THE ACTIVATION COEFFICIENT MATRICES OF THE AUDIO AND THE LIP SURFACE MODALITIES

#	multimodal soft		hard coupling of $\mathbf{H}_l, \mathbf{H}_r$ and $\mathbf{H}^v$	
	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$
1	4.10	6.6	**6.00**	**8.50**
2	**5.99**	**7.43**	3.92	5.55
3	**4.01**	**4.82**	3.43	4.38
4	**10.27**	**8.56**	7.19	5.55
5	4.18	**2.68**	**4.22**	2.62
6	**4.2**	**3.31**	3.96	3.11
7	6.86	**2.72**	**7.04**	2.63
8	2.70	2.63	**2.84**	**3.05**
9	**5.79**	**0.68**	4.39	0.29
10	**6.10**	**1.76**	5.34	1.03
avg	**5.42**	**4.119**	4.833	3.671

audio and the lip surface modalities. As mentioned before, the activation coefficient matrices of the corresponding audio and lip surface modalities are similar, but they are not necessarily equal. So, the proposed soft coupling between $\mathbf{H}_l$, $\mathbf{H}_r$ and $\mathbf{H}^v$, which is able to preserve their difference, results in a better source separation performance compared with the hard coupling situation in which $\mathbf{H}_l = \mathbf{H}^v$ and $\mathbf{H}_r = \mathbf{H}^v$.

In the next simulation, the proposed method is compared with the situation where $\mathbf{H}_l = \mathbf{H}_r$ but they are coupled in a soft manner to $\mathbf{H}^v$. The results are presented in Table VI. It is seen that, for most mixtures, the proposed multimodal soft NMCF algorithm achieves a better source separation performance.

F. Investigating the Penalty Term

In this section, the effect of the proposed penalty term on the performance of the proposed algorithm is investigated. For this purpose, the proposed penalty term is compared with the ℓ_2 penalty term defined as (where $\|.\|_F$ stands for Frobenius norm)

$$\|\mathbf{H}_l - \mathbf{H}^v\|_F^2 + \|\mathbf{H}_r - \mathbf{H}^v\|_F^2, \tag{31}$$

and also compared with the ℓ_1 penalty term defined as

$$\|\mathbf{H}_l - \mathbf{H}^v\|_1 + \|\mathbf{H}_r - \mathbf{H}^v\|_1. \tag{32}$$

TABLE VI
SEPARATION SNR (dB) FOR THE PROPOSED MULTIMODAL SOFT NMCF ALGORITHM AND THE SITUATION WHEN $\mathbf{H}_l$ AND $\mathbf{H}_r$ ARE ASSUMED TO BE EQUAL

#	multimodal soft		multimodal with hard coupling of $\mathbf{H}_l, \mathbf{H}_r$	
	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$
1	**3.56**	**5.24**	3.30	3.67
2	**6.80**	**8.31**	6.65	8.03
3	**3.23**	**4.28**	2.15	3.26
4	10.37	8.64	**11.55**	**9.85**
5	**4.37**	**3.53**	4.28	3.19
6	**4.16**	**3.25**	3.97	3.15
7	6.86	2.64	**8.37**	**3.67**
8	**3.06**	**2.84**	2.98	2.73
9	5.74	**1.32**	**5.95**	1.04
10	**5.67**	**1.23**	5.45	0.79
avg	5.382	**4.128**	**5.465**	3.938

TABLE VII
SEPARATION SNR (dB) FOR THE PROPOSED MULTIMODAL SOFT NMCF ALGORITHM AND FOR THE COST FUNCTIONS WITH ℓ_2 AND ℓ_1 COUPLINGS

#	multimodal soft		ℓ_2 coupling		ℓ_1 coupling	
	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$
1	**5.66**	**8.30**	-0.42	2.77	-1.17	2.65
2	**4.74**	**6.23**	0.80	3.42	1.88	3.55
3	**3.53**	**4.54**	0.99	3.77	0.44	3.04
4	**10.61**	**8.84**	2.34	1.23	1.18	0.84
5	**3.95**	**3.08**	2.69	1.34	2.64	1.3
6	**4.64**	**3.69**	2.74	2.09	2.89	2.15
7	**6.70**	**2.53**	2.86	-1.28	2.78	-0.83
8	**3.35**	**3.22**	1.88	1.95	1.06	1.69
9	**5.12**	**0.58**	3.29	-0.03	3.16	0.19
10	**6.31**	**1.84**	3.23	-0.85	2.58	-1.41
avg	**5.461**	**4.285**	2.04	1.441	1.744	1.317

The proposed cost function (23) and the cost functions obtained from (31) and (32) are compared with each other. The penalty coefficients are set equal to 1. The results are presented in Table VII. It is clear that the proposed penalty term results in a better separation performance compared with the ℓ_2 and the ℓ_1 penalty terms.

G. Simultaneous Factorization of the Audio and the Lip Surface Modalities

In this part, the proposed algorithm is compared with an approach in which the audio mixtures and the lip surface modalities are factorized simultaneously. In this situation, the cost function (21) changes to

$$\begin{aligned} C_{\text{sim}} = {} & D_{\text{IS}}(\mathbf{V}_l^x \| \mathbf{W}_{1l}\mathbf{H}_{1l} + \mathbf{W}_{2l}\mathbf{H}_{2l}) \\ & + D_{\text{IS}}(\mathbf{V}_r^x \| \mathbf{A}_1\mathbf{W}_{1r}\mathbf{H}_{1r} + \mathbf{A}_2\mathbf{W}_{2r}\mathbf{H}_{2r}) \\ & + D_{\text{IS}}(\mathbf{V}_1^v \| \mathbf{W}_1^v\mathbf{H}_1^v) + D_{\text{IS}}(\mathbf{V}_2^v \| \mathbf{W}_2^v\mathbf{H}_2^v) \\ & + \lambda_l p(\mathbf{H}_l, \mathbf{H}^v) + \lambda_r p(\mathbf{H}_r, \mathbf{H}^v), \end{aligned}$$

where $\mathbf{V}_1^v \in \mathbb{R}_+^{F \times N}$ and $\mathbf{V}_2^v \in \mathbb{R}_+^{F \times N}$ are the power spectrogram matrices of the first and the second lip surface modalities

TABLE VIII
SEPARATION SNR (dB) FOR THE PROPOSED MULTIMODAL SOFT NMCF ALGORITHM AND FOR THE SITUATION WHEN THE LIP SURFACE MODALITIES AND THE AUDIO MIXTURES ARE FACTORIZED SIMULTANEOUSLY

#	multimodal soft		simultaneous factorization of modalities	
	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$
1	**7.14**	**9.80**	5.44	8.17
2	**6.64**	**8.44**	4.64	6.79
3	**3.72**	**4.70**	3.51	4.68
4	**8.70**	**6.65**	6.3	4.20
5	**4.39**	3.32	4.31	**3.59**
6	**4.46**	**3.37**	3.93	3.21
7	6.73	**2.47**	**7.71**	2.29
8	**3.47**	**3.14**	2.81	1.92
9	**6.17**	**1.44**	5.92	0.12
10	**5.62**	**1.24**	4.03	0.56
avg	**5.704**	**4.457**	4.86	3.553

TABLE IX
SEPARATION SNR (dB) FOR THE PROPOSED MULTIMODAL SOFT NMCF, THE SOFT COUPLING AND THE EM BASED ALGORITHMS FOR A MORE COMPLICATED MIXTURE

#	multimodal soft		EM based		soft coupling	
	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$
1	**7.39**	**9.99**	-0.53	4.44	-0.77	1.93
2	**6.10**	**7.65**	-1.50	-0.12	0.82	2.49
3	**3.43**	**4.55**	0.02	3.12	-0.85	3.37
4	**11.12**	**9.47**	1.85	-0.12	0.53	0.27
5	**4.70**	**3.28**	1.4	-0.05	1.17	0.39
6	**4.48**	**3.72**	0.91	0.17	2.17	1.69
7	**7.46**	**2.63**	1.02	-3.52	3.24	-2.70
8	**3.21**	**3.19**	1.17	-0.15	2.81	2.22
9	**4.60**	**1.27**	0.80	-0.21	4.49	-1.26
10	**5.44**	**1.18**	3.15	-1.05	1.89	-1.84
avg	**5.793**	**4.693**	0.829	0.251	1.55	0.656

and $\mathbf{W}_1^v \in \mathbb{R}_+^{F \times K}$ and $\mathbf{W}_2^v \in \mathbb{R}_+^{F \times K}$ are the basis dictionary matrices resulted from the factorization of $\mathbf{V}_1^v$ and $\mathbf{V}_2^v$, respectively. Recall that $\mathbf{H}^v = [\mathbf{H}_1^{vT}, \mathbf{H}_2^{vT}]^T$. The other parameters have been defined earlier. The third and the fourth terms of C_{sim} correspond to the factorization of the lip surface modalities and the last two terms correspond to the penalty terms. The penalty coefficients are set equal to 1. The resulting SNR's are presented in Table VIII.

It is clear from the results that, most of the times, the sequential factorization of the lip surface signals and the audio mixtures, which is proposed in this paper, results in a better source separation performance, compared with the simultaneous factorization of the lip surface modalities and the audio mixtures.

H. More Complicated Mixtures

Finally, we investigate the performance of the proposed algorithm for separating the following more complicated mixing system:

$$\begin{aligned} x_l(t) &= s_1(t) + s_2(t), \\ x_r(t) &= 0.8 s_1(t - 0.0019) + 0.4 s_1(t - 0.0062) \\ &\quad + 0.2 s_1(t - 0.0094) + s_2(t - 0.0031) \\ &\quad + 0.3 s_2(t - 0.0075) + 0.1 s_2(t - 0.0125). \end{aligned}$$

The proposed algorithm is compared with the soft coupling algorithm and the Expectation Maximization (EM) based algorithm proposed in [5]. The EM based algorithm proposed in [5], is based on the maximization of the joint likelihood of the mixtures using EM algorithm. The results are presented in Table IX.

It is seen that even for more complicated mixtures, the performance of the proposed algorithm is more than the performances achieved by the soft coupling and the EM based algorithms.

VI. CONCLUSION

In this paper, a multimodal algorithm was proposed for the separation of the convolutive mixtures of the audio signals when the video signals of the speakers are also available. The proposed algorithm was focused on separating the audio sources from the stereo mixtures, and with the help of the lip surface signal of each speaker as the second modality. The similarity of the activation coefficient matrices of the audio and the lip surface modalities along with the similarity of the NMF parameters of the audio signals of the two mixtures were used for convolutive source separation using the proposed multimodal soft NMCF approach. The penalty term of the soft NMCF algorithm [15], which controls the similarity of the activation coefficient matrices of the audio modalities, was split into two penalty terms that control the similarity of the activation coefficient matrices of the audio and the lip surface modalities corresponding to a same speech. In the first step of the algorithm, the activation coefficient matrices of the lip surface modalities were extracted and in the second step, the resulting activation coefficient matrices were used for convolutive audio source separation. The proposed algorithm does not need deriving any prior probability model for the audio and the lip surface modalities and does not suffer from the permutation problem. The update rules are derived using a majorization-minimization approach and with the help of convex auxiliary functions. Although the auxiliary functions are convex, the main cost function is not convex and the algorithm usually converges to a local minimum of the cost function. The extraction of the lip surface signals of each speaker requires a good video quality, but it is a very simple feature, and the simulation results show that despite this simple feature, the multimodal algorithm proposed here outperforms the audio-only algorithms. Future works can be devoted to source separation by taking into account the delay between the audio and the lip surface modalities and considering more accurate video information of each speaker than the lip surface signal.

APPENDIX A
DERIVING THE UPDATE RULES

As mentioned earlier, the update rules are derived using a majorization-minimization approach and with the help of auxiliary functions [18]. In [18], the following auxiliary function is used for estimating $\mathbf{H}$ in (5) when D is the Itakura-Saito divergence

$$G_1(\mathbf{H},\mathbf{H}^t)=\sum_{f,n}\left\{v(f,n)\sum_k\frac{h^t(k,n)^2w(f,k)}{h(k,n)\hat{v}(f,n)^2}\right.$$
$$\left.+\sum_k h(k,n)\frac{w(f,k)}{\hat{v}(f,n)}\right\}+\text{cst},$$

where $h^t(k,n)$ is the (k,n)-th element of $\mathbf{H}^t$ and $\hat{v}(f,n)$ is the (f,n)-th element of $\hat{\mathbf{V}}=\mathbf{W}\mathbf{H}^t$. The first term of the above auxiliary function is derived by using the Jensen's inequality and the second term of the above auxiliary function is achieved by replacing the "log" function by its tangent at the point $h^t(k,n)$ [18]. In this paper, we have used the following auxiliary function for optimizing the proposed penalty term

$$G_2(\mathbf{H},\mathbf{H}^t)=\sum_{k,n}\frac{1}{h(k,n)}+\frac{h(k,n)}{h^v(k,n)h^t(k,n)}+\text{cst}.$$

Similar to [18], the above auxiliary function is achieved by replacing the "log" function by its tangent at the point $h^t(k,n)$.

By the above discussions, the following auxiliary functions are used for optimizing $\mathbf{H}_{1l}$ at the point $\mathbf{H}_{1l}^t$ and $\mathbf{H}_{2l}$ at the point $\mathbf{H}_{2l}^t$:

$$\sum_{f,n}\left\{v_l^x(f,n)\sum_k\frac{h_{1l}^t(k,n)^2w_1(f,k)}{h_{1l}(k,n)\hat{v}_l(f,n)^2}\right.$$
$$\left.+\sum_k h_{1l}(k,n)\frac{w_1(f,k)}{\hat{v}_l(f,n)}\right\}$$
$$+\lambda_l\left(\sum_{k,n}\left\{\frac{1}{h_{1l}(k,n)}+\frac{h_{1l}(k,n)}{h_1^v(k,n)h_{1l}^t(k,n)}\right\}\right)+\text{cst},$$
$$\sum_{f,n}\left\{v_l^x(f,n)\sum_k\frac{h_{2l}^t(k,n)^2w_2(f,k)}{h_{2l}(k,n)\hat{v}_l(f,n)^2}\right.$$
$$\left.+\sum_k h_{2l}(k,n)\frac{w_2(f,k)}{\hat{v}_l(f,n)}\right\}$$
$$+\lambda_l\left(\sum_{k,n}\left\{\frac{1}{h_{2l}(k,n)}+\frac{h_{2l}(k,n)}{h_2^v(k,n)h_{2l}^t(k,n)}\right\}\right)+\text{cst},$$

where $h_{1l}^t(k,n)$ and $h_{2l}^t(k,n)$ are the (k,n)-th elements of $\mathbf{H}_{1l}^t$ and $\mathbf{H}_{2l}^t$, respectively.

In a similar manner, the following auxiliary functions are used for optimizing $\mathbf{H}_{1r}$ and $\mathbf{H}_{2r}$ at the points $\mathbf{H}_{1r}^t$ and $\mathbf{H}_{2r}^t$, respectively:

$$\sum_{f,n}\left\{v_r^x(f,n)\sum_k\frac{h_{1r}^t(k,n)^2a_1(f,f)w_1(f,k)}{h_{1r}(k,n)\hat{v}_r(f,n)^2}\right.$$
$$\left.+\sum_k h_{1r}(k,n)\frac{a_1(f,f)w_1(f,k)}{\hat{v}_r(f,n)}\right\}$$
$$+\lambda_r\left(\sum_{k,n}\left\{\frac{1}{h_{1r}(k,n)}+\frac{h_{1r}(k,n)}{h_1^v(k,n)h_{1r}^t(k,n)}\right\}\right)+\text{cst},$$
$$\sum_{f,n}\left\{v_r^x(f,n)\sum_k\frac{h_{2r}^t(k,n)^2a_2(f,f)w_2(f,k)}{h_{2r}(k,n)\hat{v}_r(f,n)^2}\right.$$
$$\left.+\sum_k h_{2r}(k,n)\frac{a_2(f,f)w_2(f,k)}{\hat{v}_r(f,n)}\right\}$$
$$+\lambda_r\left(\sum_{k,n}\left\{\frac{1}{h_{2r}(k,n)}+\frac{h_{2r}(k,n)}{h_2^v(k,n)h_{2r}^t(k,n)}\right\}\right)+\text{cst}.$$

The auxiliary functions for optimizing $\mathbf{W}_1$ and $\mathbf{W}_2$ at the points $\mathbf{W}_1^t$ and $\mathbf{W}_2^t$ are

$$\sum_{f,n}\left\{v_l^x(f,n)\sum_k\frac{w_1^t(f,k)^2h_{1l}(k,n)}{w_1(f,k)\hat{v}_l(f,n)^2}\right.$$
$$\left.+\sum_k w_1(f,k)\frac{h_{1l}(k,n)}{\hat{v}_l(f,n)}\right\}$$
$$+\sum_{f,n}\left\{v_r^x(f,n)\sum_k\frac{w_1^t(f,k)^2a_1(f,f)h_{1r}(k,n)}{w_1(f,k)\hat{v}_r(f,n)^2}\right.$$
$$\left.+\sum_k w_1(f,k)\frac{a_1(f,f)h_{1r}(k,n)}{\hat{v}_r(f,n)}\right\}+\text{cst},$$
$$\sum_{f,n}\left\{v_l^x(f,n)\sum_k\frac{w_2^t(f,k)^2h_{2l}(k,n)}{w_2(f,k)\hat{v}_l(f,n)^2}\right.$$
$$\left.+\sum_k w_2(f,k)\frac{h_{2l}(k,n)}{\hat{v}_l(f,n)}\right\}$$
$$+\sum_{f,n}\left\{v_r^x(f,n)\sum_k\frac{w_2^t(f,k)^2a_2(f,f)h_{2r}(k,n)}{w_2(f,k)\hat{v}_r(f,n)^2}\right.$$
$$\left.+\sum_k w_2(f,k)\frac{a_2(f,f)h_{2r}(k,n)}{\hat{v}_r(f,n)}\right\}+\text{cst}.$$

Finally, the following auxiliary functions are used for optimizing $\mathbf{A}_1$ and $\mathbf{A}_2$ at the points $\mathbf{A}_1^t$ and $\mathbf{A}_2^t$:

$$\sum_{f,n}\left\{v_r^x(f,n)\frac{a_1^t(f,f)^2h'(f,n)}{a_1(f,f)\hat{v}_r(f,n)^2}+a_1(f,f)\frac{h'(f,n)}{\hat{v}_r(f,n)}\right\}+\text{cst},$$
$$\sum_{f,n}\left\{v_r^x(f,n)\frac{a_2^t(f,f)^2h''(f,n)}{a_2(f,f)\hat{v}_r(f,n)^2}+a_2(f,f)\frac{h''(f,n)}{\hat{v}_r(f,n)}\right\}+\text{cst},$$

where $h'(f,n)$ and $h''(f,n)$ are the elements of $\mathbf{H}'$ and $\mathbf{H}''$ defined after (28). As mentioned earlier, in the above auxiliary functions, "cst" contains the terms that do not depend on the target parameter. Setting the derivative of each of the above auxiliary functions with respect to their target parameters equal to zero and finding the nonnegative root, result in the mentioned update rules. Note that the parameters with the superscript "t" correspond to the previous iteration (t-th iteration) and the resulting parameters using the proposed update rules correspond to the current iteration (($t+1$)-th iteration).

References

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. New York, NY, USA: Academic, 2010.

[2] V. Capdevielle, C. Serviere, and J.-L. Lacoume, "Blind separation of wideband sources in the frequency domain," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1995, vol. 3, pp. 2080–2083.

[3] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.

[4] A. Dapena, M. F. Bugallo, and L. Castedo, "Separation of convolutive mixtures of temporally-white signals: A novel frequency-domain approach," in *Proc. Int. Conf. Independent Component Anal. Blind Source Separation*, San Diego, CA, USA, 2001, pp. 315–320.

[5] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[6] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook on Speech Processing and Speech Communication*, Berlin, Germany: Springer, 2007.

[7] B. Rivet, L. Girin, and C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 96–108, Jan. 2007.

[8] D. Lahat, T. Adalı, and C. Jutten, "Challenges in multimodal data fusion," in *Proc. 22nd Eur. Signal Process. Conf.*, 2014, pp. 101–105.

[9] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 125–134, May 2014.

[10] B. Rivet, M. Duda, A. Guérin-Dugué, C. Jutten, and P. Comon, "Multimodal approach to estimate the ocular movements during EEG recordings: A coupled tensor factorization method," in *Proc. 37th IEEE Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2015, pp. 6983–6986.

[11] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2001, pp. 556–562.

[12] M. Kim, J. Yoo, K. Kang, and S. Choi, "Nonnegative matrix partial cofactorization for spectral and temporal drum source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1192–1204, Oct. 2011.

[13] E. Acar, T. G. Kolda, and D. M. Dunlavy, "All-at-once optimization for coupled matrix and tensor factorizations," in *Proc. Workshop on Mining and Learning with Graphs*, arXiv:1105.3422, 2011.

[14] N. Seichepine, S. Essid, C. Févotte, and O. Cappé, "Soft nonnegative matrix co-factorization with application to multimodal speaker diarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 3537–3541.

[15] N. Seichepine, S. Essid, C. Févotte, and O. Cappé, "Soft nonnegative matrix co-factorization," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5940–5949, Nov. 2014.

[16] D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J.-L. Schwartz, and C. Jutten, "A study of lip movements during spontaneous dialog and its application to voice activity detection," *J. Acoust. Soc. Amer.*, vol. 125, no. 2, pp. 1184–1196, 2009.

[17] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura–Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.

[18] C. Févotte, "Majorization-minimization algorithm for smooth Itakura–Saito nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 1980–1983.

Farnaz Sedighin received the B.S. and M.S. degrees in electrical engineering from Isfahan University of Technology, Isfahan, Iran, in 2007 and 2010, respectively. She is currently working toward the Ph.D. degree at Sharif University of Technology, Tehran, Iran. Her main research areas include multimodal signal processing and audio–visual source separation.

Massoud Babaie-Zadeh (SM'09) received the B.S. degree in electrical engineering from Isfahan University of Technology, Isfahan, Iran, in 1994, and the M.S. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 1996, and the Ph.D. degree in signal processing from Institute National Polytechnique of Grenoble, Grenoble, France, in 2002. He received the best Ph.D. Thesis Award of INPG for his Ph.D. dissertation. Since 2003, he has been a faculty member of the Electrical Engineering Department, Sharif University of Technology, Tehran, Iran, in which, he is currently a Full Professor. His main research areas are blind source separation and sparsity-aware signal processing.

Bertrand Rivet was graduated from the École Normale Supérieure de Cachan, France. He received the Agrégation de Physique Appliquée in 2002, the Master's degree from the University of Paris-XI, France, in 2003 and the Ph.D. degree from Grenoble Institute of Technology (GIT), France, in 2006.

He is currently an Associate Professor in signal processing with PHELMA and a member of GIPSA-lab, GIT, France. His research concerns linear and nonlinear source separation/extraction and sensors network based on multimodal recordings applied on biomedical signal processing, audiovisual speech processing and chemical data.

Christian Jutten (AM'92–M'03–SM'06–F'08) received Ph.D. and Doctor és Sciences degrees in signal processing from Grenoble Institute of Technology (GIT), Grenoble, France, in 1981 and 1987, respectively. Since 1989, he has been a Full Professor at University Grenoble-Alpes. For 35 years, his research interests have been machine learning and source separation, including theory (separability, source separation in nonlinear mixtures, sparsity, multimodality) and applications (brain and hyperspectral imaging, chemical sensor array, speech). He is author or coauthor of more than 95 papers in international journals, 4 books, 27 keynote plenary talks and about 215 communications in international conferences.

He has been a member of a few IEEE Technical Committees, and currently in "SP Theory and Methods" of the IEEE Signal Processing Society. He received best paper awards of EURASIP in 1992 and of IEEE GRSS in 2012, and Medal Blondel in 1997 from the French Electrical Engineering Society for his contributions in source separation and independent component analysis. He was elevated as IEEE fellow (2008) and EURASIP fellow (2013). He is a senior member of Institut Universitaire de France since 2008, with renewal in 2013 for 5 years. He received a 2012 ERC Advanced Grant for a project on challenges in extraction and separation of sources. In 2016, he received one Grand Prix of the French Académie des Sciences.