# Two Multimodal Approaches for Single Microphone Source Separation

Farnaz Sedighin*, Massoud Babaie-Zadeh*, Bertrand Rivet† and Christian Jutten†
* School of Electrical Engineering, Sharif University of Technology, Tehran, Iran
Email: f_sedighin@ee.sharif.edu, mbzadeh@yahoo.com
† GIPSA Lab, Univ. Grenoble Alpes, Grenoble, France
Email: bertrand.rivet@gipsa-lab.grenoble-inp.fr, christian.jutten@gipsa-lab.grenoble-inp.fr

***Abstract*—In this paper, the problem of single microphone source separation via Nonnegative Matrix Factorization (NMF) by exploiting video information is addressed. Respective audio and video modalities coming from a single human speech usually have similar time changes. It means that changes in one of them usually corresponds to changes in the other one. So it is expected that activation coefficient matrices of their NMF decomposition are similar. Based on this similarity, in this paper the activation coefficient matrix of the video modality is used as an initialization for audio source separation via NMF. In addition, the mentioned similarity is used for post-processing and for clustering the rows of the activation coefficient matrix which were resulted from randomly initialized NMF. Simulation results confirm the effectiveness of the proposed multimodal approaches in single microphone source separation.**

***Index Terms*—Single microphone source separation, Nonnegative matrix factorization, Multimodal source separation.**

## I. Introduction

Single microphone speech separation is a challenging task in source separation. In this problem, the observed signal is a superposition of a set of original speech signals, that is

$$x(t) = \sum_{i=1}^{L} s_i(t), \tag{1}$$

where $x(t)$ is the observed signal, $s_i(t)$'s are original sources and $L$ is the number of sources. The goal is to separate the original sources from $x(t)$.

Different approaches have already been proposed for single microphone source separation. In [1] single microphone source separation is achieved by learning a set of time domain basis functions. Single channel Independent Component Analysis (ICA), wavelet ICA, Ensemble Empirical Mode Decomposition (EEMD) ICA and using deep learning are some of other methods [2], [3].

One of the approaches used for single microphone source separation is Nonnegative Matrix Factorization (NMF) [4], [5]. NMF decomposes a nonnegative data matrix as the product of two matrices with nonnegative components as [6], [7]

$$\mathbf{V} \simeq \mathbf{WH}, \tag{2}$$

where $\mathbf{V} \in \Re^{F\times N}$, $\mathbf{W} \in \Re^{F\times K}$ and $\mathbf{H} \in \Re^{K\times N}$. $\mathbf{V}$ is called observation matrix, $\mathbf{W}$ is called basis dictionary matrix and $\mathbf{H}$ is called activation coefficient matrix [4], [5]. We will explain in Section II how the model (1) leads to (2) in the time-frequency domain, using Short Time Fourier Transform (STFT).

For achieving good results in source separation via NMF, $\mathbf{W}$ should be clustered so that basis dictionary vectors corresponding to the same source are clustered together. So some algorithms for single microphone source separation via NMF such as [4] use a prior training for estimating $\mathbf{W}$. In these algorithms, $\mathbf{W}$ is learned from training data and kept fixed during the updating process, so during the estimation process only $\mathbf{H}$ is updated. Manual clustering of basis vectors after NMF decomposition is also proposed in [4].

Initialization of $\mathbf{W}$ and $\mathbf{H}$ highly affects the quality of NMF decomposition. Good initialization results in a good decomposition. Many NMF based algorithms are initialized randomly, so they have to be run several times and the best final result is picked up [8]. However, such a procedure can be expensive especially when the algorithm has a high computational load. So different initialization algorithms are proposed. In [8] an initialization algorithm based on Singular Value Decomposition (SVD) named as NonNegative Double SVD (NNDSVD) is proposed for initializing $\mathbf{W}$ and $\mathbf{H}$. In [9] another initialization approach based on hierarchical clustering of attributes through the similarity measure is presented. Other initialization algorithms can be found in [10]. For single microphone source separation, user guided audio source separation is proposed in [5], [11]. In these algorithms, an end user is asked to manually annotate the activity of each source. This information is then used for initializing the activation coefficients matrix.

Recently, separating audio signals using multimodal nature of speech is studied in different source separation algorithms such as [12], [13]. Different aspects of a multimodal phenomenon are measured by different instruments [14]. Each of these measurements is called a modality. A human speech is an example of a multimodal (bimodal) phenomenon which consists of audio and video modalities of the speaker. Bimodal nature of speech is also exploited in NMF for various applications such as speaker diarization [15].

In this paper, we use the bimodal nature of speech along

This work has been partly supported by the European project ERC-2012-AdG-320684-CHESS, and also by the Center for International Scientific Studies and Collaboration (CISSC).

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

with NMF to enhance the quality of single microphone source separation. Basic source separation approach of this paper is similar to [4] but no prior training or manual clustering of basis vectors is needed. Special initialization is used in this paper to force the algorithm to converge to a proper minimum. Moreover, instead of manually annotating the activity of each source, which is used in [5], the similarity of the activation coefficient matrices of the respective modalities is used. In addition, this similarity is used for post-processing of the NMF based single microphone source separation which is initialized randomly. This post-processing is done by clustering the rows of the activation coefficient matrix after NMF decomposition.

The paper is organized as follows. In Section II, NMF model and its usage for single microphone source separation is reviewed. The main ideas are presented in Section III and finally experimental results are presented in Section IV.

## II. Preliminaries

### A. NMF Model

The goal of NMF is to factorize a matrix with nonnegative entries as the product of the two matrices with nonnegative entries, as in (2). It can be achieved by solving the optimization problem [7]

$$\min_{H\geq 0, W\geq 0} D(\mathbf{V}\|\mathbf{WH}), \tag{3}$$

where $D$ is a measure of the difference between $\mathbf{V}$ and $\mathbf{WH}$. Different functions have already been used as the above measure. One of the popular functions is Itakura-Saito (IS) divergence, defined as [5], [11]

$$D_{\text{IS}}(\mathbf{V}\|\tilde{\mathbf{V}}) = \sum_{ij} \frac{v(i,j)}{\tilde{v}(i,j)} - \log \frac{v(i,j)}{\tilde{v}(i,j)} - 1,$$

where $\tilde{\mathbf{V}} \triangleq \mathbf{WH}$ and $v(i,j)$ and $\tilde{v}(i,j)$ are the $(i,j)$-th entries of $\mathbf{V}$ and $\tilde{\mathbf{V}}$ respectively. $\mathbf{H}$ and $\mathbf{W}$ are updated during an optimization process. For example, a multiplicative update rule for optimization of the above cost function is presented as follows [16], [17],

$$h(a,\mu) \leftarrow h(a,\mu)\sqrt{\frac{\sum_i w(i,a)v(i,\mu)/\tilde{v}^2(i,\mu)}{\sum_k w(k,a)/\tilde{v}(k,\mu)}}, \tag{4}$$

$$w(i,a) \leftarrow w(i,a)\sqrt{\frac{\sum_\mu h(a,\mu)v(i,\mu)/\tilde{v}^2(i,\mu)}{\sum_\nu h(a,\nu)/\tilde{v}(i,\nu)}}, \tag{5}$$

where $w(i,a)$ and $h(a,\mu)$ are the components of $\mathbf{W}$ and $\mathbf{H}$ respectively.

### B. Single Microphone Source Separation Based on NMF

The use of NMF for single microphone source separation is studied in different papers such as [4], [5], [18]. STFT matrix of a signal is a matrix whose $n$-th column is the Fourier transform of the $n$-th frame of that signal. The $n$-th frame of a time domain signal such as $y(k)$ is calculated as

$$y_n(k) = y(k+nM')\mathcal{W}(k) \qquad k = 0,1,...,M-1, \tag{6}$$

where $y_n(k)$ is the $n$-th frame of $y$ and $\mathcal{W}$ is a finite-length window with length $M$, and $M'$ is the amount of the shift of the window. Since STFT is a linear transform, for a mixture of $L$ sources

$$x(f,n) = \sum_{i=1}^{L} s_i(f,n), \tag{7}$$

where $x(f,n)$ is the $(f,n)$-th component of the STFT matrix of the mixture and $s_i(f,n)$ is the $(f,n)$-th component of the STFT matrix of the $i$-th source signal. The power spectrum matrix $\mathbf{V}$ is a matrix whose components are equal to

$$v(f,n) \triangleq |x(f,n)|^2. \tag{8}$$

$\mathbf{V}$ is a $F \times N$ matrix where $F$ is the number of frequency bins and $N$ is the number of time frames. The use of NMF for single microphone source separation is based on the assumption that the power spectrum matrix of the $i$-th source is factorized to $\mathbf{W}_i\mathbf{H}_i$ where $\mathbf{W}_i$ and $\mathbf{H}_i$ are the basis dictionary and activation coefficient matrices of the NMF decomposition of the power spectrum of the $i$-th source, respectively [4]. So separation of sources is achieved by the following NMF decomposition [4]

$$\mathbf{V} \simeq \mathbf{WH} = \sum_{i=1}^{L} \mathbf{W}_i\mathbf{H}_i, \tag{9}$$

where $\mathbf{W} = [\mathbf{W}_1, ..., \mathbf{W}_L]$ and $\mathbf{H} = [\mathbf{H}_1^T, ..., \mathbf{H}_L^T]^T$. Finally, by using Wiener filtering, the STFT matrix of the $i$-th separated signal is estimated as the matrix whose $(f,n)$-th component is given by

$$\hat{s}_i(f,n) = \frac{p_i(f,n)}{(\sum_{i=1}^{L} p_i(f,n))} x(f,n), \tag{10}$$

where $p_i(f,n)$ is the $(f,n)$-th component of $\mathbf{P}_i = \mathbf{W}_i\mathbf{H}_i$. In the above single microphone source separation approach, *the main challenging task is grouping* $\mathbf{W}$ *to* $\mathbf{W}_i$*'s, so that each* $\mathbf{W}_i$ *corresponds to the* $i$*-th source.*

For achieving good results in single microphone source separation, in [4] $\mathbf{W}_i$ is learned from training data for each individual source and then only $\mathbf{H}_i$ is updated during the update procedure. However, training data is not always available so there are other algorithms like [5], [11] that use manual annotation of activity of sources for initializing $\mathbf{H}$. But, when the parameters of the NMF based single microphone source separation approach $(\mathbf{W}, \mathbf{H})$ are initialized randomly and no training data is available, clustering of basis vector after decomposition is needed. So in [4], this clustering of basis vectors has been done manually.

## III. Main Ideas

Bimodal nature of speech is exploited in different speech source separation algorithms [12], [13]. In this paper, we use the lip surface of the speaker as the video modality. Since respective audio and video modalities have a similar physical origin (human speech), changes in the audio modality usually corresponds to changes in the lip surface of the speaker.

Due to this similarity, it is expected that the activation coefficient matrices of the NMF factorization of the power spectra of respective audio and video modalities ($\mathbf{H}^a$ and $\mathbf{H}^v$, respectively) have similar shapes. More specifically, it is expected that respective modalities have zeros in nearly the same components of their activation coefficients matrices. It means that their inactive periods are expected to be nearly the same. Based on this similarity two different ideas are proposed in the rest of the paper: one for initialization of the NMF based separation approach of Section II-B and one for clustering of basis vectors of NMF approach of Section II-B when the parameters $(\mathbf{W}, \mathbf{H})$ are initialized randomly and no training data is available. The ideas are compared with each other.

### A. *First Idea: Multimodal Initialization Approach*

Since the update rules (4), (5) are multiplicative, zero components do not change during the update process. So the zero components of the finalized and the initialization matrices are mostly the same. Having in mind the similarity of activation coefficient matrices of respective modalities, in this paper we propose that the activation coefficient matrix of the video modality ($\mathbf{H}^v$) is used as an initialization for the single microphone audio source separation. Note that in this paper the video modality is a one dimensional signal consisting of the lip surface of the speaker extracting from its three dimensional video. So computing its STFT matrix, NMF decomposition of its power spectrum matrix and consequently computing its activation coefficient matrix is similar to what is done in the previous section. In addition, video modalities (lip surface signals) can have constant nonzero values during their inactive periods. In this situation, derivative of the lip surface signals is used. So, the lip surface signals are first differentiated and then their activation coefficient matrices are extracted. So $\mathbf{H}^v$ can be used instead of manual annotation of the activity of audio signals which is used in [5], [11]. Therefore we initialize the single microphone audio source separation algorithm of [4], with the following initialization matrix

$$\mathbf{H}^a_{\text{init}} = [\mathbf{H}_1^{vT}, ..., \mathbf{H}_i^{vT}, ..., \mathbf{H}_L^{v\,T}]^T,$$

where $\mathbf{H}_i^v$ is the activation coefficients matrix of the *i*-th video modality. Each row of $\mathbf{H}_i^v$ is normalized to have unity summation. It should be noted that rows of $\mathbf{H}_i^v$'s have similar zero patterns but they are not exactly the same. The number of the rows of $\mathbf{H}_i^v$ is set to a predetermined positive integer $\kappa$. $\mathbf{H}^a_{\text{init}}$ is then used as an initialization for the activation coefficient matrix in NMF decomposition of the power spectrum matrix of the audio mixture. It is clear that for the NMF decomposition of the mixture of $L$ signals, $K$ (the number of the rows of $\mathbf{H}^a_{\text{init}}$) is set to $L \times \kappa$. Zero components do not change during the update procedure, so the finalized $\mathbf{H}$ after the update procedure has mostly the same zero components as its initialization matrix ($\mathbf{H}^a_{\text{init}}$). Since the $(\kappa(i-1)+1 : \kappa i)$-th rows of the activation coefficient matrix have been initialized by $\mathbf{H}_i^v$, it is expected that after the update procedure, the $(\kappa(i-1)+1 : \kappa i)$-th rows of the finalized $\mathbf{H}$ and the $(\kappa(i-1)+1 : \kappa i)$-th columns of the finalized $\mathbf{W}$ will correspond to $s_i$. The original signals are then reconstructed using (10).

### B. *Second Idea: Multimodal Clustering Approach*

The second idea is clustering the basis vectors resulted from NMF based single microphone source separation when the parameters $(\mathbf{W}, \mathbf{H})$ are initialized randomly. As mentioned before, in [4] manual clustering of basis vectors is done after NMF decomposition. In this paper, it is proposed to use the similarity between activation coefficient matrices of modalities for clustering the rows of activation coefficient matrix and consequently the basis vectors. This should be noted again that this clustering is used after a randomly initialized NMF algorithm. The number of clusters is equal to the number of original sources ($L$). Center of the $i$-th cluster is the average of the rows of $\mathbf{H}_i^v$, *i.e.*:

$$C_i = \frac{\sum_{j=1}^{\kappa} \mathbf{H}_i^v(j,:)}{\kappa}, \tag{11}$$

where $C_i$ is a row vector of size $1 \times N$ ($N$ is the number of the columns of $\mathbf{H}_i^v$) and is the center of the $i$-th cluster. $\mathbf{H}_i^v(j,:)$ is the $j$-th row of $\mathbf{H}_i^v$. Since activation coefficient matrices of respective audio and video modalities have nearly the same form, it is expected that the matrix $[C_i^T, \mathbf{H}(j,:)^T]^T_{2\times N}$ be nearly a rank one matrix. Matrix rank can be measured using the following criterion [19]

$$\rho_{ij}([C_i^T, \mathbf{H}(j,:)^T]^T) = \frac{\sigma_1^2}{\sum_{m=2}^{p} \sigma_m^2}, \tag{12}$$

where $\mathbf{H}(j,:)$ is the $j$-th row of the activation coefficient matrix of the mixture, $\sigma_i$ is the $i$-th singular value of the SVD of $[C_i^T, \mathbf{H}(j,:)^T]^T$ and $p$ is the number of singular values. $\sigma_1$ is the largest singular value. Then $\mathbf{H}(j,:)$ is clustered to the $m$-th cluster if

$$\rho_{mj}([C_m^T, \mathbf{H}(j,:)^T]^T) > \rho_{ij}([C_i^T, \mathbf{H}(j,:)^T]^T) \qquad \forall i \neq m.$$

The rows of $\mathbf{H}$ belong to the $m$-th cluster are clustered into $\mathbf{H}_{\text{c}}^m = [\mathbf{H}(m_1,:)^T, ..., \mathbf{H}(m_\kappa,:)^T]^T$ , where $m_1, ..., m_\kappa$ are the rows of $\mathbf{H}$ belong to the $m$-th cluster. So the clustered activation coefficient matrix is calculated as

$$\mathbf{H}_{\text{c}} = [\mathbf{H}_{\text{c}}^{1\,T}, \mathbf{H}_{\text{c}}^{2\,T}, ..., \mathbf{H}_{\text{c}}^{L\,T}]^T. \tag{13}$$

Then $\mathbf{W}$ is also clustered to $\mathbf{W}_{\text{c}}$ such that it corresponds to the above clustering in $\mathbf{H}$. Separated sources are then reconstructed using $\mathbf{H}_{\text{c}}$ and $\mathbf{W}_{\text{c}}$.

## IV. Numerical experiments

In this Section, the effect of the proposed multimodal initialization and multimodal clustering approaches for source separation is investigated. Signals used for simulation are pairs of audio and video modalities of different speeches [20]. It should be noted that in this paper, the one dimensional signal of the lip surface of the speaker is used as the video modality. The audio signals are sampled at 16 kHz and the lip surface signals are sampled at 50 Hz. So the lip surface signals are upsampled by factor 320 using "`interp.m`" function of

TABLE I
SNR (IN dB) FOR THE SEPARATED SIGNALS USING THE PROPOSED, NNDSVD AND RANDOM INITIALIZATIONS FOR10 TRIALS.

| # | $\hat{\mathbf{S}}_{1\text{MM}}$ | $\hat{\mathbf{S}}_{2\text{MM}}$ | $\hat{\mathbf{S}}_{1\text{SVD}}$ | $\hat{\mathbf{S}}_{2\text{SVD}}$ | $\hat{\mathbf{S}}_1$ | $\hat{\mathbf{S}}_2$ |
|---|---|---|---|---|---|---|
| 1 | 7.67 | 9.70 | -0.91 | 1.77 | -4.08 | 0.72 |
| 2 | 5.30 | 6.72 | -0.77 | 0.99 | -0.68 | 0.63 |
| 3 | 2.90 | 4.18 | -0.56 | 1.62 | -1.09 | 1.69 |
| 4 | 6.48 | 3.93 | 1.87 | -1.80 | 1.31 | -0.13 |
| 5 | 4.06 | 2.90 | 2.54 | 0.55 | 2.28 | 1.14 |
| 6 | 4.02 | 3.38 | 2.35 | -0.69 | 1.85 | 1.03 |
| 7 | 7.02 | 2.77 | 3.62 | -2.82 | 2.07 | -3.32 |
| 8 | 5.61 | 1.33 | 3.89 | -2.47 | 2.21 | -1.33 |
| 9 | 3.94 | 3.12 | 3.09 | 1.00 | -0.22 | 0.64 |
| 10 | 8.39 | 3.07 | 6.40 | 0.27 | 1.44 | -0.74 |
| **avg** | 5.54 | 4.11 | 2.15 | -0.15 | 0.5 | -0.13 |

TABLE II
SNR (IN dB) FOR THE SEPARATED SIGNALS USING THE PROPOSED INITIALIZATION, PROPOSED CLUSTERING ALGORITHM AND RANDOM INITIALIZATION FOR 10 TRIALS.

| # | $\hat{\mathbf{S}}_{1\text{MM}}$ | $\hat{\mathbf{S}}_{2\text{MM}}$ | $\hat{\mathbf{S}}_{1\text{c}}$ | $\hat{\mathbf{S}}_{2\text{c}}$ | $\hat{\mathbf{S}}_1$ | $\hat{\mathbf{S}}_2$ |
|---|---|---|---|---|---|---|
| 1 | 7.05 | 7.67 | -0.62 | 4.08 | -2.60 | 0.20 |
| 2 | 6.08 | 7.60 | -0.60 | 3.59 | -1.70 | 0.05 |
| 3 | 4.02 | 5.58 | -0.04 | 3.30 | -1.30 | 0.66 |
| 4 | 6.02 | 3.33 | 1.04 | -0.48 | 1.22 | 0.59 |
| 5 | 2.79 | 2.69 | 3.17 | 1.17 | 2.28 | 1.07 |
| 6 | 3.77 | 2.97 | 1.08 | 0.94 | 0.72 | -1.14 |
| 7 | 6.35 | 2.15 | 2.62 | -1.88 | 0.33 | -4.02 |
| 8 | 5.37 | 1.42 | 4.89 | 1.15 | 2.20 | -2.00 |
| 9 | 4.67 | 2.36 | 3.71 | 1.42 | 1.03 | 2.24 |
| 10 | 8.11 | 1.88 | 7.39 | 0.09 | 6.75 | -0.33 |
| **avg** | 5.42 | 3.76 | 2.26 | 1.33 | 0.89 | -0.26 |

MATLAB. Mixtures are produced by mixing two real audio signals ($L = 2$). There is a video modality (lip surface signal) correspond to each of the audio signals. Each of the respective pairs of audio and lip surface signals are modalities of a human speech. It is worth noting that there is no mixture of video modalities, $i.e.$, each video modality corresponds to a single speech. The duration of the original audio signals is 32 sec. NMF decomposition of lip surface modalities is initialized randomly. STFT is computed using 0.0625 sec length window (1000 samples). Quality of the separated signals is measured using the following SNR

$$\text{SNR} = 10\log_{10}\Big(\frac{\sum_{i,j}\mathbf{S}(i,j)^2}{\sum_{i,j}(\mathbf{S}(i,j)-\hat{\mathbf{S}}(i,j))^2}\Big), \tag{14}$$

where $\mathbf{S}$ and $\hat{\mathbf{S}}$ are the magnitude spectra of the original and the estimated signals respectively.

In the first simulation, the effect of the proposed initialization approach (first idea) comparing to NNDSVD of [8] and random initialization approaches is investigated. It should be noted that for a better comparison of the initialization algorithms, in the NNDSVD approach, only $\mathbf{H}$ is initialized using the NNDSVD algorithm. $\mathbf{W}$ is initialized randomly for all of the algorithms. 10 mixtures are separated using NMF with the proposed, NNDSVD and random initialization approaches. $\kappa$ is set to 10. Source separation results are presented in Table I. The first two columns of the table ($\hat{\mathbf{S}}_{1\text{MM}}$ and $\hat{\mathbf{S}}_{2\text{MM}}$) correspond to the SNRs of the two separated signals using the proposed initialization. $\hat{\mathbf{S}}_{1\text{SVD}}$ and $\hat{\mathbf{S}}_{2\text{SVD}}$ correspond to the NNDSVD initialized separation algorithm and the two last columns correspond to the randomly initialized separation algorithm.

It is inferred from the results that NMF separation algorithm with the proposed multimodal initialization has a better quality compared to NNDSVD and randomly initialized NMF algorithms. As mentioned before, this better quality is due to the proposed initialization which prevents disordering of the rows of the matrix $\mathbf{H}$.

In the second simulation, the proposed clustering algorithm (second idea) is investigated. In Table II, the quality of the classical NMF (randomly initialized NMF) source separation algorithm with the proposed clustering algorithm is compared to the classical NMF without clustering of basis vectors. In addition, the proposed initialization approach and the proposed clustering algorithms are also compared with each other in Table II. In this simulation $\kappa$ is set to 5.

In Table II, $\hat{\mathbf{S}}_{1\text{c}}$ and $\hat{\mathbf{S}}_{2\text{c}}$ are the separated signals estimated after the proposed clustering algorithm. $\hat{\mathbf{S}}_{1\text{MM}}, \hat{\mathbf{S}}_{2\text{MM}}$, $\hat{\mathbf{S}}_1$ and $\hat{\mathbf{S}}_2$ are defined before. It is clear from the results that the proposed clustering algorithm increases the quality of source separation comparing to the randomly initialized NMF source separation algorithm without clustering. But the quality of source separation algorithm with the proposed initialization is greater than the randomly initialized algorithms with or without clustering.

In the third simulation, the situation when the video information (lip surface signal) is only available for one of the sources, say $s_1$, is investigated. This occurs when we are only interested to extract one of the sources from the mixture. So only first $\kappa$ rows of $\mathbf{H}$ are initialized by using the video information and the rest of the rows are initialized randomly. Separation results for the mentioned, NNDSVD and randomly initialized NMF based single microphone source separation approach are presented in Table III. The first two columns of the Table III ($\tilde{\mathbf{S}}_{1\text{MM}}$ and $\tilde{\mathbf{S}}_{2\text{MM}}$) correspond to the mentioned initialization. $\tilde{\mathbf{S}}_{i\text{MM}}$ is used instead of $\hat{\mathbf{S}}_{i\text{MM}}$ to indicate that only first $\kappa$ rows of $\mathbf{H}$ are initialized by using video (lip surface signal) information.

It can be inferred from the results that using the video information of only one of the sources increases the quality of separation comparing to NNDSVD and random initializations. However, separation quality using the video information of only one of the sources is less than the separation quality when the video information of all of the sources is used (the difference between $\hat{\mathbf{S}}_{1\text{SVD}}$ ($\hat{\mathbf{S}}_{2\text{SVD}}$) in Table I and $\hat{\mathbf{S}}_{1\text{SVD}}$ ($\hat{\mathbf{S}}_{2\text{SVD}}$) in Table III and the difference between $\hat{\mathbf{S}}_{1\text{MM}}$ ($\hat{\mathbf{S}}_{2\text{MM}}$) in Table I and $\hat{\mathbf{S}}_{1\text{MM}}$ ($\hat{\mathbf{S}}_{2\text{MM}}$) in Table II is due to the random initialization of $\mathbf{W}$). So separation quality for the mentioned initializations

TABLE III
SNR (IN dB) FOR THE SEPARATED SIGNALS WHEN THE VIDEO INFORMATION IS AVAILABLE ONLY FOR THE FIRST SOURCE COMPARING TO NNDSVD AND RANDOM INITIALIZATIONS FOR 10 TRIALS.

| # | $\tilde{\mathbf{S}}_{1\text{MM}}$ | $\tilde{\mathbf{S}}_{2\text{MM}}$ | $\hat{\mathbf{S}}_{1\text{SVD}}$ | $\hat{\mathbf{S}}_{2\text{SVD}}$ | $\hat{\mathbf{S}}_1$ | $\hat{\mathbf{S}}_2$ |
|---|---|---|---|---|---|---|
| 1 | 5.80 | 8.69 | -0.72 | 1.97 | -1.94 | 1.21 |
| 2 | 3.50 | 6.14 | -1.29 | 0.56 | 1.91 | 4.25 |
| 3 | 1.99 | 3.91 | 0.02 | 1.93 | 0.30 | 2.75 |
| 4 | 4.03 | 3.10 | 1.85 | -0.38 | 1.51 | -0.54 |
| 5 | 1.87 | 2.29 | 2.06 | 0.91 | 0.88 | 1.58 |
| 6 | 2.15 | 1.51 | 1.85 | -0.31 | -0.01 | -1.83 |
| 7 | 3.26 | -0.99 | 4.11 | -1.42 | 2.22 | -2.40 |
| 8 | 3.05 | -0.89 | 3.50 | -0.72 | 1.80 | -1.62 |
| 9 | 3.26 | 1.18 | 5.70 | 0.31 | 1.82 | -1.91 |
| 10 | 0.89 | 3.51 | 0.29 | 0.46 | -0.47 | 0.05 |
| **avg** | 2.98 | 2.84 | 1.73 | 0.33 | 0.8 | 0.15 |

can be sorted as: multimodal initialization for all of the sources $>$ multimodal initialization for one of the sources$>$ NNDSVD initialization $\geq$ random initialization.

## V. CONCLUSION

In this paper, single microphone source separation via NMF was addressed. We proposed to use the similarity of activation coefficient matrices of audio and video modalities coming from a single speech to initialize an already known approach for single microphone speech separation based on NMF (Section II-B). In addition, this similarity was used for clustering of basis vectors when the parameters of the mentioned approach are initialized randomly without using any training data. Then the effectiveness of the proposed algorithms was verified by some simulations. The experiments show the crucial aspect of initialization for NMF. Moreover, the proposed multimodal approach is a good choice for initializing the NMF based single microphone source separation approach even if only one video modality is known. So based on the proposed algorithms, single microphone source separation can be achieved without using any training data, manual clustering of basis vectors or manual annotation the activity of sources.

## REFERENCES

[1] G.-J. Jang, T.-W. Lee, and Y.-H. Oh, "Single-channel signal separation using time-domain basis functions," *IEEE Signal Processing Letters*, vol. 10, no. 6, pp. 168–171, 2003.

[2] B. Mijović, M. De Vos, I. Gligorijević, J. Taelman, and S. Van Huffel, "Source separation from single-channel recordings by combining empirical-mode decomposition and independent component analysis," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 9, pp. 2188–2196, 2010.

[3] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1562–1566.

[4] B. Wang and M. D. Plumbley, "Investigating single-channel audio source separation methods based on non-negative matrix factorization," in *Proc. ICA Research Network International Workshop*, 2006, pp. 17–20.

[5] N. Q. Duong, A. Ozerov, and L. Chevallier, "Temporal annotation-based audio source separation using weighted nonnegative matrix factorization," in *IEEE Fourth International Conference on Consumer Electronics Berlin (ICCE-Berlin)*, 2014, pp. 220–224.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[7] ——, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[8] C. Boutsidis and E. Gallopoulos, "Svd based initialization: A head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350–1362, 2008.

[9] Y.-D. Kim and S. Choi, "A method of initialization for nonnegative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2007, pp. II–537.

[10] A. N. Langville, C. D. Meyer, R. Albright, J. Cox, and D. Duling, "Initializations for the nonnegative matrix factorization," in *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, 2006, pp. 23–26.

[11] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 257–260.

[12] 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 Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 96–108, 2007.

[13] ——, "Visual voice activity detection as a help for speech source separation from convolutive mixtures," *Speech Communication*, vol. 49, no. 7, pp. 667–677, 2007.

[14] D. Lahat, T. Adali, and C. Jutten, "Challenges in multimodal data fusion," in *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 101–105.

[15] N. Seichepine, S. Essid, C. Févotte, and O. Cappé, "Soft nonnegative matrix co-factorizationwith application to multimodal speaker diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3537–3541.

[16] C. Févotte, "Majorization-minimization algorithm for smooth itakura-saito nonnegative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 1980–1983.

[17] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[18] D. El Badawy, N. Q. Duong, and A. Ozerov, "On-the-fly audio source separation," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.

[19] J. Bhattacharya, P. P. Kanjilal, and V. Muralidhar, "Analysis and characterization of photo-plethysmographic signal," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 1, pp. 5–11, 2001.

[20] 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," *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 1184–1196, 2009.