SINGLE SENSOR AUDIOVISUAL SPEECH SOURCE SEPARATION

Pierre Narvor, Bertrand Rivet, Christian Jutten

Univ. Grenoble Alpes, Gipsa-Lab, F-38000 Grenoble, France
CNRS, GIPSA-Lab, F-38000 Grenoble France

ABSTRACT

The Kernel Additive Modeling (KAM) is a recent promising framework for the separation of underdetermined convolutive mixture of audio signal. The principle of this method is to estimate the short term Power Spectral Densities (PSD) of the sources directly from the mixture by taking advantage of redundant features in the PSD of the source, such as periodicity or smoothness. The separation itself is then performed with a generalized Wiener filter. This preliminary study aims to evaluate the improvement of using the video of the speaker's face to directly detect such redundancies in the speech that could be used in the KAM framework to perform the extraction of the speech signal.

Index Terms— Multimodality, Convolutive Informed Source Separation, Audiovisual, Wiener Filtering

1. INTRODUCTION

The problem of Convolutive Blind Source Separation (CBSS) of audio signals is still a challenging task [1]. Many methods have been developed in the last decades using as many or more microphone than the sources, e.g. independent component analysis (ICA) [2] or independent vector analysis (IVA) [3]. On the other hand, underdetermined audio source separation is mainly based on Wiener filters [4] estimating e.g., in a Bayesian framework [5] or by non-negative matrix factorization [6].

Recently, a new framework has given promising results when applied to underdetermined convolutive mixtures [7]. The Power Spectral Densities (PSD) of the sources used by the Wiener filter are estimated directly from the mixture. The source separation problem is therefore transformed into a source parameter estimation problem. The Kernel Additive Modelling (KAM) approach for source separation presented in [8] is based on this framework. This study partly focuses on the extraction of lead voice from stereo recordings of music, in which features like smoothness or periodicity are strongly present in the PSD of the instrumental signals. The main goal of this approach is to take advantage of the so called "local dynamics" of the PSD of audio signals to estimate their parameters. The source features allow an efficient modelling of the different PSD structures via a source specific proximity kernel based on some redundancy. For instance, in the time-frequency domain, such redundancy can be based on local smoothness of the PSD (e.g., speech) or due to its periodicity (e.g., percussions in music).

When dealing with speech signals, it is well-known that there exist strong links between the audio signal and the face of the speaker (e.g., the movements of the lips' speaker) [9, 10]. Such redundancies have already be applied to a source extraction problem [11]: based on spectral subtraction, on ICA or IVA, on audiovisual dictionary decomposition or time-frequency masking. However, the relationship between the audio and video signals is complex and does not allow an easy translation from the lip movements to the audio speech signal. This contribution is a preliminary study in which we aim at assessing that if using the video of the face of the speaker as an extra modality to detect more redundancy in the speech PSD could effectively help the separation by extending the KAM framework into a multimodal one.

The remaining of this paper is organized as follows. Section 2 described the proposed multimodal KAM algorithm. The results are presented in Section 3 before conclusions and perspectives in Section 4.

2. METHOD

In this section, the modelling of the tackled problem is defined in subsection 2.1, then the separation principle of KAM presented in [8] is recalled in subsection 2.2, and our contribution is presented in subsection 2.3.

2.1. Modelling of the source separation problem

In this paper, the considered recorded signal $x(t)$ is assumed to be a linear mixture of latent sources $s_i(t)$:

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

where I is the number of sources. The extraction of the ith source $s_i(t)$ is then achieved in the time-frequency domain

This work have been partly supported by the ERC project CHESS: 2012-ERC-AdG-320684

by a time-varying Wiener filter. Let $X(t,f)$ denotes the short term Fourier transform (STFT) of $x(t)$. The estimation of $s_i(t)$ is thus provided by the inverse short term Fourier transform of

$$\hat{S}_i(t,f) = H_i(t,f)X(t,f), \tag{2}$$

with $H_i(t,f)$ the Wiener filter defined by

$$H_i(t,f) = \frac{P_i(t,f)}{\sum_{j=1}^{I} P_j(t,f)}, \tag{3}$$

where $P_j(t,f)$ is the PSD of the jth source $s_j(t)$. In practice, the PSD in (3) are substituted by their estimates that must be computed from $X(t,f)$ using some prior knowledge on the sources properties.

2.2. Kernel Additive Modeling

2.2.1. Modelization and estimation of the PSD of the sources

The KAM framework [8] provides a convenient model to estimate the PSD of the sources $s_j(t)$ from the PSD of the mixture $x(t)$. For each frequency bin f and time t, $S_j(t,f)$ is assumed independent of each other, and distributed with respect to a centered Gaussian distribution:

$$\forall (t,f), \quad S_j(t,f) \sim \mathcal{N}\big(0, P_j(t,f)\big), \tag{4}$$

where $P_j(t,f) \geq 0$ is the PSD of the source j at STFT bin (t,f). Being the sum of I independent Gaussian random variables, $X(t,f)$ also follows a Gaussian distribution defined as:

$$\forall (t,f), \quad X(t,f) \sim \mathcal{N}\left(0, \sum_{i=1}^{I} P_i(t,f)\right). \tag{5}$$

To estimate the PSD $P_j(t,f)$, the KAM assumes that at the time-frequency (TF) coordinates (t,f), a specific set $\mathcal{I}_j(t,f)$ of pair (t',f') can be defined, for which the PSD $P_j(t',f')$ has a value close to $P_j(t,f)$:

$$\forall (t',f') \in \mathcal{I}_j(t,f), \quad P_j(t',f') \approx P_j(t,f). \tag{6}$$

The set $\mathcal{I}_j(t,f)$ actually defines a proximity kernel in the form of a binary mask. The shape of this binary mask is chosen accordingly to the structure of the PSD of the considered source. For example, if the PSD is known to be periodic of period T_0 with respect to time, the proximity kernel can be defined as a periodic set of time-frequency coordinates: $\mathcal{I}_j(t,f) = \{(t+kT_0, f) \,|\, k \in \mathcal{K} \subset \mathbb{Z}\}$. The estimation can then be achieved by minimizing the absolute deviation:

$$\hat{P}_j(t,f) = \underset{P}{\operatorname{argmin}} \sum_{(t',f') \in \mathcal{I}_j(t,f)} |Z_j(t',f') - P|, \tag{7}$$

where $Z_j(t',f')$ is the observed value of $P_j(t',f')$. The minimization is achieved by:

$$\hat{P}_j(t,f) = \operatorname{median}\big(Z_j(t',f') \,|\, (t',f') \in \mathcal{I}_j(t,f)\big). \tag{8}$$

2.2.2. Separation algorithm

The extraction is performed by an iterative algorithm based on some chosen proximity kernels for each source defined from their properties.

Let $\hat{S}_i^{(k)}(t,f)$ be the estimate of the SFTF of the ith source at kth iteration . Each iteration is decomposed in three steps:

- compute the observed source PSD $Z_j(t,f)$ as

$$\forall i, \quad Z_i^{(k)}(t',f') = \big|\hat{S}_i^{(k-1)}(t,f)\big|^2, \tag{9}$$

- estimate the PSD of each source $\hat{P}_i^{(k)}(t,f)$ by (8),
- estimate the STFT of each source $\hat{S}_i^{(k)}(t,f)$ by (2) with the Wiener filter (3) using the estimated PSD $\hat{P}_i^{(k)}(t,f)$.

Without any prior knowledge on the sources other than their kernels, their STFT are initialized as $\hat{S}_i(t,f) = X(t,f)/I$. This process is then repeated a fixed number of iterations, typically 5.

2.3. Multimodal KAM

The KAM framework has been shown very efficient on the separation of sources of different PSD structures, as with the extraction of lead vocals from a stereo recording of music [8]. However, its main drawback is that it cannot separate two sources which follow the same kind of time-frequency structure, as for example two speech sources. The main idea of this paper is to discriminate a target speech source form the others by using redundancy in the contents of the speech that would be detected with the video of the speaker's face using a visual speech redundancy detector (VSRD).

2.3.1. Principle of the method

The proposed multimodal KAM (MM-KAM) is an extension of the KAM framework based on a simple speech redundancy detector using video as detailed in subsection 2.3.2. Here the proximity kernels are not only defined on the assumed structure of the speech signal but also on repetition of words detected by the video. Let $\mathcal{T}(t)$ be the set of time frames that have been detected by the VSRD as a repetition of the frame at time t. In other words, at a given time t, the set $\mathcal{T}(t)$ contains all time indexes of frames similar to the one at time t. Then the multimodal proximity kernel to use for the separation of the repeated words is defined as:

$$\mathcal{I}_r(t,f) = \bigcup_{t_0 \in \mathcal{T}(t)} \mathcal{I}_0(t_0, f), \tag{10}$$

where $\mathcal{I}_0$ is a proximity kernel well suited to speech signals.

The separation algorithm is then performed with a similar algorithm than the one described in Section 2.2.2. The PSD

of the redundancies are supposed to be equal to each other and are thus estimated as:

$$\hat{P}_r(t,f) = \text{median}\big(Z_j(t',f') \,|\, (t',f') \in \mathcal{I}_r(t,f)\big), \quad (11)$$

where $\mathcal{I}_r(t,f)$ is defined by (10).

2.3.2. Visual speech redundancy detector

The VSRD is based on the optical flow of the video previously centered on the speaker's face. This feature is be a good candidate to detect redundancies in the video signal since it has been recently shown to be efficient to discriminate visemes [12], even between different speakers. It is also immune to conditions of illumination of the face and does not require the same level of precision of segmentation than lip segmentation based recognition. However, the indeterminacies between lip movements and speech signal shall prevent to make a perfect speech redundancy detector. To increase the accuracy, it is possible to compare longer speech segments. This gives more information to compare two segments but at the expense of the redundancy density. For these reasons, the figures presented in Section 3 where computed using 750ms long speech segment, which is around the time scale of a word. Finally, a video segment is the concatenation of 37 frames of optical flow. The similarity measure is the normalized dot product. The occurrences are then detected with the local maxima of a normalized correlation between a reference speech segment and the entire optical flow data.

2.3.3. Limitations of the MM-KAM

The MM-KAM method suffers from several drawbacks intrinsic to lip reading. The first one is the non-bijective relationship between the movements of the lip and the contents of the speech. It is not guaranteed that a match in the VSRD correspond to speech segments with the same content. This is a major problem that is not tackled in this paper. However, even with a proper detection of redundancies in the speech, there are other potential limitations to the use of this method. The first one is the density of the redundancies. If the density of redundancies is not high enough, the VSRD will not find enough matches for the MM-KAM to be efficient. Also, even if a proper repetition is detected, the speech segments can differ in rate of locution, tone of the voice, and a delay caused by the frame rate of the video which is usually low compared to the speech signal rate of variations. The difference in rate of locution is not considered in this study because the VRSD described in this paper should consider that the two segments are not repetitions. A possible improvement of the method would be to handle the difference in rates of locution directly with the VSRD by a time warping [13].

3. RESULTS

In this section, the behaviour of the proposed MM-KAM is assessed on three experiments (subsection 3.2) after the definition of the performance index (subsection 3.1).

3.1. Performance Measure

The extraction performance was evaluated using the Signal to Distortion Ratio (SDR) [14]. The SDR represents a signal to noise ratio where the signal is the best estimate of a source that could be obtain by the extraction algorithm if all of the separation parameters, in this case the sources PSDs, have been perfectly estimated. In a simulation case, these parameters are given by an oracle. The noise in the SDR definition is the difference between the estimated source $\hat{y}$ and the oracle source y. The SDR is then defined as:

$$\text{SDR} \;=\; 10\log_{10}\frac{\|\, y \,\|^2}{\|\, y-\hat{y} \,\|^2} \quad (12)$$

The SDR is expressed in decibel (dB), and the higher it is, the better is the extraction quality.

3.2. Experiements

The results are averaged over 76 trials. For each trial, a random mixture is generated containing (i) a target speech signal with known repetitions, (ii) another speech signal, and (iii) a music recording containing several instruments and a lead voice. The SDR of the desired speech signal is -0.5 dB at the initialization. The audio signals where sampled at 16kHz and STFT of the sources where computed on 50ms long windows with an overlap of 90%.

For the target speech signal, the kernel $\mathcal{I}_0$ is a cross-like kernel with a bandwidth of 20Hz and a duration of 25ms such as the one used in [8]. Another cross-like kernel is used to define another potential speech source, or another source with a similar structure, like a piano. The bandwidth of the kernel is 40Hz and its duration is 45ms. A last kernel is defined to represent a nearly stationary source, like a synthesizer in a music recording or a background constant white noise. This kernel is a rectangle of bandwidth 20Hz and duration 750ms.

3.2.1. Number of occurrences of a speech segment

This experiment investigates the impact of using redundancy in the proposed MM-KAM compared to the KAM that does not use them. As reported in Tab. 1, a dramatic increase in SDR is obtained when using at least two occurrences instead of only one. Indeed, the mixture consists of several speech-like sources and without using at least one repetition, there

Table 1: Maximum obtained SDR in the ideal case with respect to the number of occurrences

Number of occurences	1	2	3	4	5
SDR [dB]	-1.3	7.66	9.85	11.5	11.9

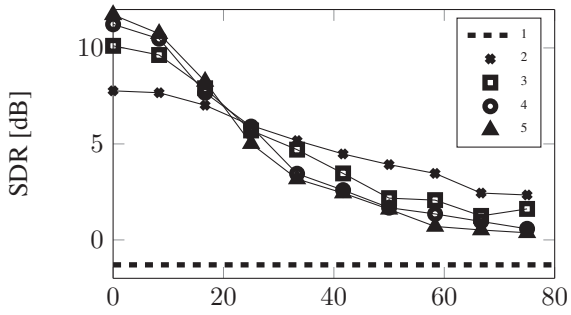

Fig. 1: Evolution of SDR with respect to the half width of the distribution of the delay error between repetitions. Caption reports the number of occurrences used in the MM-KAM.

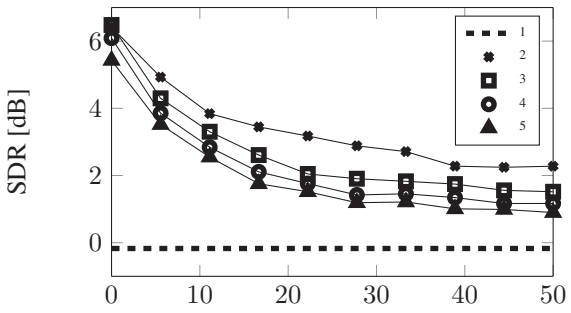

Fig. 2: Evolution of SDR with respect to the half width of the distribution of f0 relative error between repetitions. Caption reports the number of occurrences used in the MM-KAM.

is no information to discriminate these sources while the proposed MM-KAM can. It is worth noting that the SDR of the target source decreases from -0.5dB to -1.3dB when the redundancies are not taken into account (i.e. KAM). This could seems strange at first but since only three kernels have been defined for this separation so only three sources are estimated. A lesser SDR at the end only means that the estimated source remains a mixture of several sources and that an other source is more predominant than the target one. Finally, the quality of separation increases with the number of used occurrences, but the gain in SDR is small when using more than 4 occurrences.

3.2.2. Synchrony issue

To evaluate the error induced by synchrony issues between the audio and video, the separation is performed by the MM-KAM after adding a shift from their actual positions of the redundancy kernels. The shifts are randomly chosen by sampling a centered uniform distribution. Ten widths of uniform distribution were chosen from 0ms to 150ms. As expected, it can be seen on Fig. 1 that the higher is the error in delays between modalities, the lesser the SDR is. Moreover, when less occurrences are used, the method is less sensitive to the delay: the decreases of SDR with a maximum delay of 75ms are 5.4dB and 11dB when using 2 or 5 occurrences, respectively, compared to the case with a perfect synchrony (i.e. no delay). Indeed, a misalignment in the definition of the MM-kernel (10) leads to a less accurate estimation of the PSD (11). Nevertheless, the proposed method is still efficient when the delay between occurrences is above 20ms, which is the time between frames in a video sampled at 50Hz. However a loss of SDR of around 3dB is still observed.

3.2.3. Diversity in fundamental frequency

To objectively measure the loss of SDR induced by a variation on tone of the voice, it is needed to control the fundamental frequency of the vowels. A shift in fundamental frequency is performed using a simple phase vocoder. This treatment is performed in the frequency domain. To dilate or shrink the fundamental frequency, some time frequency bin are added to or deleted from the spectrum before resampling it back to its original size. The formants are kept at the same location with the help of an estimation of the envelope of the spectrum based on a linear prediction coding (LPC) procedure. It is worth noting that this simple vocoder induces some distortions on the signal: it explains that even with no f0 shift the results are lower than in the ideal case presented in previous section. The f0 where shifted from a value sampled from a uniform centered distribution. Ten width of distribution where chosen from 0% to 100% of the original f0 value. Fig. 2 shows that the larger is variations of f0, the worst the performance are. It is worth noting that the lost quality is greater when more occurrences are used. Indeed, a misalignment of f0 leads to misaligned observed source PSD (9) and thus to a less accurate PSD estimated by (11). However, there are still an improvement compared to the KAM.

4. CONCLUSIONS AND PERSPECTIVE

In this study, a multimodal extension of the KAM framework has been proposed as MM-KAM. Based on the redundancy of speech between the audio signals and the video of the speaker's face, a specific multimodal kernel of "local smoothness" has been defined from the VSRD. The numerical experiments show the advantage of the proposed MM-KAM compared to an audio only KAM to extract a specific speech, especially if several speech signals are mixed. However, one of the main drawbacks of this method is that long speech segments around the length of a word have to be used to detect accurately speech redundancies. This drastically reduces the number of used redundancies and therefore the efficiency of the method with natural speech.

Future work will focus on bringing down the needed length of the audiovisual segment with more refined techniques based on both the video and the sound mixture to detect the redundancies. The field of Automated Lip Reading shall be of a great help to achieve this objective.

5. REFERENCES

[1] Pierre Comon and Christian Jutten, Eds., *Handbook of Blind Source Separation Independent Component Analysis and Applications*, Academic Press, 2010.

[2] Pierre Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, April 1994.

[3] T. Adali, M. Anderson, and Geng-Shen Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *Signal Processing Magazine, IEEE*, vol. 31, no. 3, pp. 18–33, May 2014.

[4] Steven M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[5] L. Benaroya, L. McDonagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for Wiener based source separation with a single sensor," in *Proc. ICASSP*, Hong-Kong 2003, pp. 613–616.

[6] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, March 2010.

[7] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[8] Antoine Liutkus, Derry Fitzgerald, Zafar Rafii, Bryan Pardo, and Laurent Daudet, "Kernel additive models for source separation," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4298–4310, 2014.

[9] W.H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of Aucoustical Society of America*, vol. 26, pp. 212–215, 1954.

[10] Norman P. Erber, "Interaction of audition and vision in the recognition of oral speech stimuli," *J. Speech and Hearing Research*, vol. 12, pp. 423–425, 1969.

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

[12] Ayaz A Shaikh, Dinesh K Kumar, and Jayavardhana Gubbi, "Visual speech recognition using optical flow and support vector machines," *International Journal of Computational Intelligence and Applications*, vol. 10, no. 02, pp. 167–187, 2011.

[13] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, Feb 1978.

[14] Emmanuel Vincent, Rémi Gribonval, and Mark D Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007.