Contributions in Audio Modeling and Multimodal Data Analysis via Machine Learning

Ngoc Q. K. Duong

ANNÉE 2020

UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

Habilitation à Diriger des Recherches

Mention : Traitement du Signal et l'Apprentissage Automatique

présentée par

Rémi, Quang-Khanh-Ngoc DUONG

préparée et soutenue chez InterDigital
Cesson-Sévigné, France, le 8 Octobre 2020.

Contributions in Audio Modeling and Multimodal Data Analysis via Machine Learning

devant le jury composé de :

Mark Plumbley
Professeur, University of Surrey, UK / rapporteur

Laurent Girin
Professeur, Institut Polytechnique de Grenoble / rapporteur

Caroline Chaux
Chargée de Recherche, CNRS / rapporteur

Patrick Le Callet
Professeur, Polytech Nantes/Université de Nantes / examinateur

Guillaume Gravier
Directeur de Recherche, CNRS/ examinateur

Acknowledgements

I would like to acknowledge the support from many people who have helped me along the way to this milestone. First of all, I would like to thank all the jury members: Mark Plumbley, Laurent Girin, Caroline Chaux, Patrick Le Callet, and Guillaume Gravier for the evaluation of this thesis.

My special thank goes to Alexey Ozerov, Patrick Pérez, Claire-Hélène Demarty, Slim Essid, Gaël Richard, Bogdan Ionescu, Mats Sjöberg, and many other colleagues whose names I could not mention one by one here. They have kindly collaborated with me along the way to this milestone. I would like to thank Emmanuel Vincent and Rémi Gribonval for supervising my PhD work, which opened a new door to my professional career. I would also like to thank Seung-Hyon Nam for supervising my Master's thesis, which initiated me to the signal processing area.

I am also very grateful to Romain Cohendet, Sanjeel Parekh, Hien-Thanh Duong, Dalia El Badawy, and other thirteen Master's intern students whom I co-supervised with colleagues. Without them the work described in this manuscript would not be possible.

I thank all my friends in Vietnam, France, and other countries without mentioning their names as there are many. You are always with me during the happy or sad moments.

Last but not least, my special love goes to my parents, my sisters, my relatives in Vietnam, my wife, and my children Kévin and Émilie.

Abstract

This HDR manuscript summarizes our work concerning the applications of machine learning techniques to solve various problems in audio and multimodal data. First, we will present nonnegative matrix factorization (NMF) modeling of audio spectrograms to address audio source separation problem, both in single-channel and multichannel settings. Second, we will focus on the multiple instance learning (MIL)-based audio-visual representation learning approach, which allows to tackle several tasks such as event/object classification, audio event detection, and visual object localization. Third, we will present contributions in the multimodal multimedia interestingness and memorability, including novel dataset constructions, analysis, and computational models. This summary is based on major contributions published in three journal papers in the IEEE/ACM Transactions on Audio, Speech, and Language Processing, and a paper presented at the International Conference on Computer Vision (ICCV). Finally, we will briefly mention our other works in different applications concerning audio synchronization, audio zoom, audio classification, audio style transfer, speech inpainting, and image inpainting.

Contents

Introduction

This HDR thesis is an extensive summary of a major part of the work done since my PhD defense in 2011. Following the PhD focusing on audio signal processing, I first worked on audio related topics such as non-negative matrix factorization (NMF)-based audio source separation [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11], audio synchronization using fingerprinting techniques [12, 13], and audio zoom using beamforming techniques [14]. Then in the deep learning era, thanks to the collaboration with a number of colleagues and PhD/Master's students, I have extended my research interest to the application of machine learning techniques in audio/image manipulation and multimodal data analysis. In the first area, I have considered multiple problems such as audio style transfer [15], speech inpainting [16], and image inpainting [17]. In the second area, I have investigated other challenges such as audio-visual source separation [18, 19], audio-visual representation learning applied to event/object classification and localization [20, 21, 22], image/video interestingness [23, 24, 25, 26], and image/video memorability [27, 28, 29] for visual content assessment. Especially, to push forward for research in such high level concepts of how media content can be interesting and memorable to viewers, I co-founded two series of international challenges in the MediaEval benchmarch[1]: Predicting Media Interestingness Task running in 2016 [30], 2017 [31], and Predicting Media Memorability Task running in 2018 [32], 2019 [33]. These tasks have greatly interested the multimedia research community as shown by a large number of international participants.

As most of my work applies machine learning techniques (whether it is a conventional model such as NMF or the emerging deep learning approach) to analyse audio and multimodal data, I entitle the thesis as "Contributions in Audio Modeling and Multimodal Data Analysis via Machine Learning" and decide to present in this document mostly about the work described in four major publications as follows:

- Paper 1 [8]: a novel user-guided audio source separation framework based on NMF

[1] http://www.multimediaeval.org/

with group sparsity constraints is introduced for single-channel setting. The NMF-based generic source spectral models (GSSM) that govern the separation process are learned on-the-fly from audio examples retrieved online.

- Paper 2 [11]: a combination of the GSSM introduced in the paper 1 with the full-rank spatial covariance model within a unified Gaussian modeling framework is proposed to address multichannel mixtures. In particular, a new source variance separation criterion is considered in order to better constrain the intermediate source variances estimated in each EM iteration.

- Paper 3 [22]: a novel multimodal framework that instantiates multiple instance learning (MIL) is proposed for audio-visual (AV) representation learning. The learnt representations are shown to be useful for performing several tasks such as event/object classification, audio event detection, audio source separation and visual object localization. Especially, the proposed framework has capacity to learn from unsynchronized audio-visual events.

- Paper 4 [29]: this work focuses on understanding the intrinsic memorability of visual content. For this purpose, a large-scale dataset (VideoMem10k) composed of 10,000 videos with both short-term and long-term memorability scores is introduced to the public. Various deep neural network-based models for the prediction of video memorability are investigated, and our model with attention mechanism provides insights of what makes a content memorable.

The remainder of this thesis is structured as follows. Chapter 1 presents the contributions in audio source separation mainly described in the papers [8, 11]. Chapter 2 focuses on the contributions in the application of machine learning for audio-visual scene analysis described in the paper [22]. Chapter 3 is dedicated to my recent works in multimodal multimedia interestingness and memorability, which were published in a number of papers [23, 26, 34, 27, 28, 29], and especially the paper [29]. Chapter 4 briefly summarizes my other works on different applications: audio synchronization [35, 13], audio zoom for smartphones [14], audio classification [36, 37, 38], audio style transfer [15], speech inpainting [16], and image inpainting [17]. Finally, Chapter 5 is devoted to the conclusion and some future research perspectives. The four major papers [8, 11, 22, 29] together with my Curriculum Vitae are annexed at the end of the thesis.

I must acknowledge that I did not do all the work mentioned in this thesis alone, but with many collaborators including colleagues and the students whom I co-supervised. I am very grateful to all these people, and without them this work would not be possible.

Thus, from now on in this manuscript, unless I intend to express my personal opinion, I will use "we" and "our" while speaking about the work.

Chapter 1

Audio source separation

Supervision: Hien-Thanh Duong (PhD student), Dalia El Badawyd (MSc intern), Luc Le Magarou (MSc intern)
Main collaborator: Alexey Ozerov (Technicolor).

This chapter summarizes our work on audio source separation, both in single-channel [8] and multichannel setting [11]. Some works target consumer application such as on-the-fly source separation [3, 6, 8], while others focus more on professional scenarios considered at Technicolor such as text-informed source separation [1, 1, 2] and interactive user-guided source separation [4, 5]. Two types of information are generally exploited for the task: spectral cues and spatial cues. In our work, the former model is based on NMF, and the latter (when applied) is based on the full-rank spatial covariance model. The organization of the chapter is as follows. We begin by the problem formulation and motivation in Section 1.1. We then present the background of NMF model for supervised source separation in Section 1.2. Our contributions in single-channel setting and multichannel setting exploiting the generic source spectral model (GSSM) are presented in Section 1.3 and Section 1.4, respectively. Section 1.5 briefly summarizes other works on text-informed and interactive source separation. Finally we conclude in Section 1.6.

1.1 Motivation and problem formulation

Audio plays a central role in both human perception of surrounding environments and machine listening tasks. Real-world audio data has a complex structure due to the superposition of different sound sources. For instance, speech recordings often include concurrent speakers, music background, or environmental noise. Such noisy mixtures

challenge both human and machine to localize, separate, and understand a target sound source. Thus audio source separation, which aims at extracting individual sound sources from an observed noisy mixture signal, has been an active research topic in audio community for several decades. It is a desired processing step within many real-world applications such as automatic speech recognition, hearing aids, sound post-production, robotics, *etc* [BMC05].

Several settings have been considered in the literature. When the number of sources J is smaller than or equal to the number of observed channel I, the problem is over-determined or determined, and techniques based on independent component analysis (ICA) have been actively used during 1990s [HO00]. When $I < J$, the problem is ill-posed, and is known as *under-determined* case. In the extreme *single-channel* case when $I = 1$, the problem is highly ill-posed and, without training data to learn the source spectral patterns, additional assumptions about the sources such as *temporal continuity* or *sparsity* must be made in order to solve such an inverse problem [Vir07]. Another axis of research known as *informed* audio source separation [LDDR13, EPMP14], where the separation process is guided by some auxiliary information, has also attracted a lot of research interest since classical *blind* approaches often do not lead to satisfactory performances in many practical applications. Recently, with the advances of deep neural network (DNN), various powerful DNN-based approaches have been proposed [HKHS15, HCRW16, LAPGH20, KWS+20] which offer very promising results. However, they usually require a large amount of labeled data for training and the training is usually computationally expensive. As most of our works was done before the DNN-based source separation era, we will not discuss more about such approaches in this chapter.

Let us denote by $\mathbf{s}_j(t)$ the contribution of j-th source at the microphone array and let $\mathbf{x}(t)$ denote the observed mixture. The mixing model is written as:

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{s}_j(t) \tag{1.1}$$

The goal of source separation is to recover $\mathbf{s}_j(t)$ given $\mathbf{x}(t)$. In reverberant mixing conditions, $\mathbf{s}_j(t)$ results from the convolution of the original source with a mixing filter characterizing the acoustic environment. This convolution in the time domain is often approximated by a simple multiplication in the time-frequency (T-F) domain by means of the short-term Fourier transform (STFT). Besides, as audio sources are often sparse and non-overlapped in the T-F domain [AJY00], most audio source separation algorithms operate in such T-F domain. In our works, we considered non-negative matrix

factorization (NMF) [LS01, FBD09] for the source spectrogram model and the local Gaussian model [39] for multichannel reverberant mixing conditions when applicable.

1.2 Background of NMF-based supervised audio source separation

Let us denote by $\mathbf{X} \in \mathbb{C}^{F \times N}$ and $\mathbf{S}_j \in \mathbb{C}^{F \times N}$ the STFT coefficients of the $\mathbf{x}(t)$ and $\mathbf{s}_j(t)$, respectively, where F is the number of frequency bins and N is the number of time frames. The mixing model (1.1) is written in the T-F domain as

$$\mathbf{X} = \sum_{j=1}^{J} \mathbf{S}_j. \tag{1.2}$$

Let $\mathbf{V} = |\mathbf{X}|^{\cdot 2}$ be the power spectrogram of the mixture, where $\mathbf{X}^{\cdot p}$ is the matrix with entries $[\mathbf{X}]_{il}^{p}$, $^{\cdot p}$ denotes an element-wise operation. In NMF, it is decomposed into two smaller non-negative matrices $\mathbf{W} \in \mathbb{R}^{F \times K}$ and $\mathbf{H} \in \mathbb{R}^{K \times N}$ such that $\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH}$. The factorization is usually done by solving the following optimization problem [LS01]

$$\mathbf{W}^*, \mathbf{H}^* = \underset{\mathbf{H} \geq 0, \mathbf{W} \geq 0}{\arg\min}\ D(\mathbf{V} \| \mathbf{WH}), \tag{1.3}$$

where

$$D(\mathbf{V} \| \hat{\mathbf{V}}) = \sum_{f,n=1}^{F,N} d(\mathbf{V}_{fn} \| \hat{\mathbf{V}}_{fn}) \tag{1.4}$$

and $d(\cdot\|\cdot)$ is a scalar divergence measure. With power spectrogram matrix, Itakura-Saito (IS) divergence is often used thank to its scale invariance property and is defined as [FBD09] $d_{IS}(x\|y) = \frac{x}{y} - \log\left(\frac{x}{y}\right) - 1$. Note that, one can also use magnitude spectrogram (when $p = 1$) and other distance measures such as Euclidean and Kullback-Leibler divergence. The parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ are usually initialized with random non-negative values and are iteratively updated via multiplicative update (MU) rules [LS01, FBD09]. With IS divergence used in our work, the MU rules are as follow:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T \left((\mathbf{WH})^{\cdot -2} \odot \mathbf{V} \right)}{\mathbf{W}^T (\mathbf{WH})^{\cdot -1}} \tag{1.5}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left((\mathbf{WH})^{\cdot -2} \odot \mathbf{V} \right) \mathbf{H}^T}{(\mathbf{WH})^{\cdot -1} \mathbf{H}^T} \tag{1.6}$$

where $\odot$ denotes the Hadamard entry-wise product.

In supervised setting, we assume that some training examples are available for each source. Thus a spectral model for each source j, denoted by $\mathbf{W}_{(j)}$, can be first learned from such training examples by optimizing criterion (1.3) during training phase. Then the spectral model for all sources $\mathbf{W}$ is obtained by concatenating the individual source models as:

$$\mathbf{W} = [\mathbf{W}_{(1)}, \ldots, \mathbf{W}_{(J)}]. \tag{1.7}$$

In the separation step, the time activation matrix $\mathbf{H}$ is estimated via the MU rules as (1.5), while $\mathbf{W}$ is kept fixed. Note that the activation matrix is also partitioned into horizontal blocks as

$$\mathbf{H} = [\mathbf{H}_{(1)}^T, \ldots, \mathbf{H}_{(J)}^T]^T, \tag{1.8}$$

where $\mathbf{H}_{(j)}$ denotes the block characterizing the time activations for the j-th source.

Once the parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ are obtained, Wiener filtering is applied to compute the source STFT coefficients as

$$\hat{\mathbf{S}}_j = \frac{\mathbf{W}_{(j)}\mathbf{H}_{(j)}}{\mathbf{WH}} \odot \mathbf{X}, \tag{1.9}$$

Finally, the inverse STFT is used to produce the time domain source estimates.

1.3 Single-channel audio source separation exploiting the generic source spectral model (GSSM)

1.3.1 Motivation, challenges, and contributions

So far source separation has been considered as a difficult task, and mostly performed by audio signal processing experts. In order to make audio source separation simple and accessible by non expert people, we introduced a friendly user-guided framework named *on-the-fly* audio source separation inspired by on-the-fly visual search methods [PVZ12, CZ12] from the computer vision research. In this framework, a user only needs to provide some search keywords. Such keywords describe the sources in the mixture so that the corresponding audio examples can be retrieved on-the-fly from the internet. These examples are then used to learn the generic source spectral models (GSSM) via non-negative matrix factorization to guide the separation process. The workflow of the proposed approach is shown in Figure 1.1.

Although the on-the-fly approach simplifies the user interactions as they are now

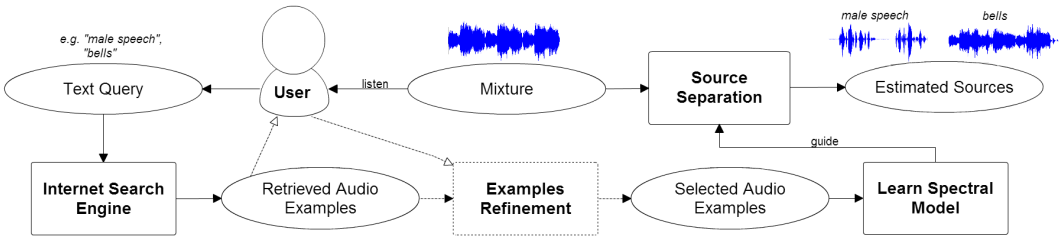


Figure 1.1: General workflow of the proposed on-the-fly framework. A user listens to the mixture and types some keywords describing the sources. These keywords are then used to retrieve examples to learn spectral models for the described sources. Optionally, the user may listen to the retrieved examples and discard irrelevant ones (figure is from [8]).

carried out at a higher semantic level, there are several challenges that need to be addressed as follows:

- (C1) *Irrelevant examples*: Some retrieved examples may contain sounds with entirely different spectral characteristics than those of the source in the mixture, *e.g.*, searching for “bird chirps" and obtaining some “chirp signal" examples too. Those examples should be automatically eliminated by the optimization algorithm.

- (C1) *Noisy examples*: Some retrieved examples are actually mixtures of relevant and irrelevant sounds, *e.g.*, “speech" with a music in the background. Those examples may still be useful but need carefully handling by the algorithm.

- (C1) *Missing examples*: This may happen when the user describes only the sources of interest and ignores the remaining sources or when the search engines do not return results for some of the provided keywords. We refer to this challenge as the semi-supervised case where all non-described sources that possibly appear in the mixture should be grouped as one background source.

The on-the-fly paradigm was published in two conference papers [3, 6] and a journal paper [8]. The main contributions are summarized as follows:

- We introduced a general framework for on-the-fly audio source separation which greatly simplifies the user interaction.

- We proposed several *group sparsity* constraints for the task and showed their benefit in both supervised and the semi-supervised cases where training examples for some sources are missing.

- We derived several algorithms for parameter estimation when different group sparsity constraints are used.
- We performed a range of evaluations, including both supervised and semi-supervised scenarios, and a user-test to validate the benefit of the proposed framework.

1.3.2 GSSM construction and model fitting

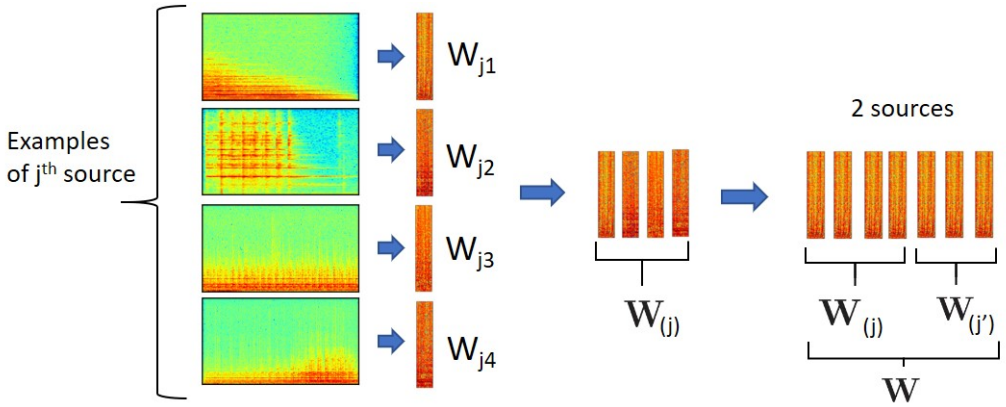


Figure 1.2: Example of an GSSM construction.

In order to address all the mentioned challenges concerning the on-the-fly framework, we considered a so-called generic source spectral models (GSSM) learned in advance from training examples, with sparsity constraints on the activation matrix in order to enforce the selection of only a few representative spectral patterns during the model fitting. The idea of GSSM was first used as "universal background model" for speaker verification in [RQD], and was later introduced in [SM13] as "universal speech model" for the separation of speech and noise.

Let us denote by $\mathbf{V}_{jp}$ the spectrogram of the p-th training example corresponding to the j-th source. First, $\mathbf{V}_{jp}$ is used to learn the NMF spectral model, denoted by $\mathbf{W}_{jp}$, by optimizing the criterion (similar to (1.3)):

$$\mathbf{H}^*_{jp}, \mathbf{W}^*_{jp} = \underset{\mathbf{H}_{jp}\geq 0, \mathbf{W}_{jp}\geq 0}{\arg\min} \; D(\mathbf{V}_{jp} \| \mathbf{W}_{jp}\mathbf{H}_{jp}), \tag{1.10}$$

where $\mathbf{H}_{jp}$ is the corresponding time activation matrix. Given $\mathbf{W}_{jp}$ for all examples, the GSSM for the j-th source is constructed as

$$\mathbf{W}_{(j)} = [\mathbf{W}_{j1}, \ldots, \mathbf{W}_{jP_j}] \tag{1.11}$$

where P_j is the number of retrieved examples for the j-th source.

Model fitting for supervised source separation

In the supervised setting, we assume having GSSM for all the sources in the mixture as the users describe all of them. $\mathbf{W}_{(j)}$ constructed in (1.11) is actually a large matrix when the number of examples increases, and it is often redundant since different examples may share similar spectral patterns. Therefore, in the NMF decomposition of the mixture, the need for a sparsity constraint arises to fit only a subset of each $\mathbf{W}_{(j)}$ to the source in the mixture. In other words, the mixture is decomposed in a supervised manner, given $\mathbf{W}$ constructed from $\mathbf{W}_{(j)}$ as in (1.7) and fixed, by solving the following optimization problem

$$\mathbf{H}^* = \underset{\mathbf{H}\geq 0}{\arg\min}\, D(\mathbf{V}\|\mathbf{W}\mathbf{H}) + \Psi(\mathbf{H}) \tag{1.12}$$

where $\Psi(\mathbf{H})$ denotes a penalty function imposing sparsity on the activation matrix $\mathbf{H}$.

Model fitting for semi-supervised source separation

We refer to a semi-supervised setting when not all of the source models can be learned in advance. In our considered on-the-fly approach, this occurs either when the user only describes the sources of interest and not all of them or when the search engine fails to retrieve examples for a given query. We can model all the "missing" sources as one background source whose spectrogram can be approximated by $\mathbf{W}_\mathrm{b}\mathbf{H}_\mathrm{b}$, where $\mathbf{W}_\mathrm{b}$ and $\mathbf{H}_\mathrm{b}$ are the corresponding spectral model and activation matrices, respectively. All the other sources, for which some examples are available, are modeled as in the supervised case by $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$. The parameter $\boldsymbol{\theta}_\mathrm{b} = \{\mathbf{W}_\mathrm{b}, \mathbf{H}_\mathrm{b}\}$ can be randomly initialized with a small number of components. All unknown parameters are then estimated altogether by optimizing the following criterion

$$\mathbf{H}^*, \mathbf{W}_\mathrm{b}^*, \mathbf{H}_\mathrm{b}^* = \underset{\mathbf{H}\geq 0, \mathbf{W}_\mathrm{b}\geq 0, \mathbf{H}_\mathrm{b}\geq 0}{\arg\min}\, D(\mathbf{V}\|\mathbf{W}\mathbf{H} + \mathbf{W}_\mathrm{b}\mathbf{H}_\mathrm{b}) + \Psi(\mathbf{H}). \tag{1.13}$$

Note that, unlike as in (1.12), in this setting $\mathbf{W}_\mathrm{b}$ is updated as well and there is no group sparsity-inducing penalty on $\mathbf{H}_\mathrm{b}$. The reason is that, as opposed to $\mathbf{W}$, $\mathbf{W}_\mathrm{b}$ is neither an overcomplete dictionary nor has an underlying structure that can be exploited for regularization.

1.3.3 Group sparsity constraints

The general group sparsity-inducing penalty is defined as

$$\Psi_\mathrm{gr}(\mathbf{H}) = \sum\nolimits_{j=1}^{J} \lambda_j \sum\nolimits_{g=1}^{G_j} \log\left(\epsilon + \|\mathbf{H}_{(j,g)}\|_1\right), \tag{1.14}$$

where $\mathbf{H}_{(j,g)}$ $(g = 1, \ldots, G_j)$ are the groups within the activation sub-matrix $\mathbf{H}_{(j)}$ corresponding to the j-th GSSM (see equation (1.8) for the definition of $\mathbf{H}_{(j)}$), G_j the total number of groups for the j-th source, $\|\cdot\|_1$ denotes the ℓ_1 matrix norm, $\epsilon > 0$ and $\lambda_j \geq 0$ are trade-off parameters determining the contribution of the penalty for each source. Note that in the remainder of the paper, $\mathbf{H}_{(j,g)}$ should not be confused with $\mathbf{H}_{jp}$ in (1.10). In [3, 8], we investigated two options for defining the groups $\mathbf{H}_{(j,g)}$ and derive the corresponding MU rules for the parameter estimation in both supervised and semi-supervised settings as follows.

Block sparsity-inducing penalty

As in [SM13], we considered the groups to be sub-matrices of $\mathbf{H}_{(j)}$ corresponding to the spectral models $\mathbf{W}_{jp}$ trained using the p-th example (see (1.10) for the definition of $\mathbf{W}_{jp}$). In that case the indices g and p coincide and $G_j = P_j$. This *block sparsity*-inducing strategy allows filtering out irrelevant spectral models $\mathbf{W}_{jl}$, thus dealing with irrelevant retrieved examples (challenge C_1). An illustration for the estimated activation matrix $\mathbf{H}$ for that case is shown in Figure 1.3-middle where blocks corresponding to irrelevant examples for each source are set to zero.

Component sparsity-inducing penalty

As an alternative solution to fitting the universal model, we proposed the groups to be rows of $\mathbf{H}_{(j)}$ corresponding to different spectral components (in that case the number of groups G_j is simply equal to the number of rows in $\mathbf{H}_{(j)}$). This so-called *component sparsity*-inducing strategy allows filtering out irrelevant spectral components, thus dealing with noisy retrieved examples (challenge C_2). Figure 1.3-right shows an estimated activation matrix $\mathbf{H}$ where rows corresponding to irrelevant spectral components for each source are set to zero.

1.3.4 Relative group sparsity constraints

With the group sparsity penalty, we observed that, in some practical cases, the group of different sources are fit together using the same source model, instead of separately using their designated models. This makes the separation impossible. We called this as "source vanishing" phenomenon. This issue is even worse in the semi-supervised case where the entire mixture is fit by the estimated background model only. This is due to the fact that $\mathbf{W}_{\mathrm{b}}$ and $\mathbf{H}_{\mathrm{b}}$ are now fully unconstrained in (1.13), whereas $\mathbf{W}$ is fixed and $\mathbf{H}$ is constrained by the group sparsity-inducing penalty. To solve this problem, we

Algorithm 1.1 MU rules for NMF with group sparsity in the supervised case (without formulas in red). When relative group sparsity is applied, formulas in red are added.

Require: $\mathbf{V}$, $\mathbf{W}$, $\boldsymbol{\lambda}$, η
Ensure: $\mathbf{H}$
  Initialize $\mathbf{H}$ randomly
  $\hat{\mathbf{V}} = \mathbf{WH}$
  repeat
    for $j = 1, \ldots, J, g = 1, \ldots, G_j$ **do**
      $\mathbf{P}_{(j,g)} \leftarrow \frac{\lambda_j}{\epsilon + \|\mathbf{H}_{(j,g)}\|_1}$
      $\mathbf{Q}_{(j,g)} \leftarrow \frac{\lambda_j G_j \gamma_j}{\|\mathbf{H}_{(j)}\|_1}$
    end for
    $\mathbf{P} = [\mathbf{P}_{(1,1)}^T, \ldots, \mathbf{P}_{(1,G_1)}^T, \ldots, \mathbf{P}_{(J,1)}^T, \ldots, \mathbf{P}_{(J,G_J)}^T]^T$
    $\mathbf{Q} = [\mathbf{Q}_{(1,1)}^T, \ldots, \mathbf{Q}_{(1,G_1)}^T, \ldots, \mathbf{Q}_{(J,1)}^T, \ldots, \mathbf{Q}_{(J,G_J)}^T]^T$
    $\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^T(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2}) + \mathbf{Q}}{\mathbf{W}^T(\hat{\mathbf{V}}^{\cdot -1}) + \mathbf{P}} \right)^{\cdot \eta}$
    $\hat{\mathbf{V}} \leftarrow \mathbf{WH}$
  until convergence

Algorithm 1.2 MU rules for NMF with group sparsity in the semi-supervised case (without formulas in red). When relative group sparsity is applied, formulas in red are added.

Require: $\mathbf{V}$, $\mathbf{W}$, $\boldsymbol{\lambda}$, η
Ensure: $\mathbf{H}$
  Initialize $\mathbf{H}$, $\mathbf{H}_b$, and $\mathbf{W}_b$ randomly
  $\hat{\mathbf{V}} \leftarrow \mathbf{WH} + \mathbf{W}_b\mathbf{H}_b$
  repeat
    for $j = 1, \ldots, J, g = 1, \ldots, G_j$ **do**
      $\mathbf{P}_{(j,g)} \leftarrow \frac{\lambda_j}{\epsilon + \|\mathbf{H}_{(j,g)}\|_1}$
      $\mathbf{Q}_{(j,g)} \leftarrow \frac{\lambda_j G_j \gamma_j}{\|\mathbf{H}_{(j)}\|_1}$
    end for
    $\mathbf{P} = [\mathbf{P}_{(1,1)}^T, \ldots, \mathbf{P}_{(1,G_1)}^T, \ldots, \mathbf{P}_{(J,1)}^T, \ldots, \mathbf{P}_{(J,G_J)}^T]^T$
    $\mathbf{Q} = [\mathbf{Q}_{(1,1)}^T, \ldots, \mathbf{Q}_{(1,G_1)}^T, \ldots, \mathbf{Q}_{(J,1)}^T, \ldots, \mathbf{Q}_{(J,G_J)}^T]^T$
    $\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^T(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2}) + \mathbf{Q}}{\mathbf{W}^T(\hat{\mathbf{V}}^{\cdot -1}) + \mathbf{P}} \right)^{\cdot \eta}$
    $\mathbf{H}_b \leftarrow \mathbf{H}_b \odot \left(\frac{\mathbf{W}_b^T(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2})}{\mathbf{W}_b^T \hat{\mathbf{V}}^{\cdot -1}} \right)^{\cdot \eta}$
    $\mathbf{W}_b \leftarrow \mathbf{W}_b \odot \left(\frac{(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2})\mathbf{H}_b^T}{\hat{\mathbf{V}}^{\cdot -1}\mathbf{H}_b^T} \right)^{\cdot \eta}$
    Normalize $\mathbf{W}_b$ and $\mathbf{H}_b$ component-wise
    $\hat{\mathbf{V}} \leftarrow \mathbf{WH} + \mathbf{W}_b\mathbf{H}_b$
  until convergence

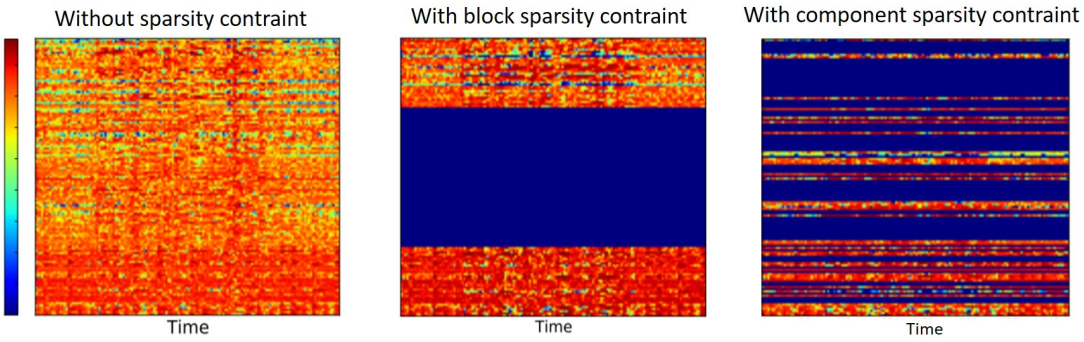


Figure 1.3: Estimated activation matrix $\mathbf{H}$ for two sources in a mixture where two training examples for each source were used for constructing the GSSM: (left) without a sparsity constraint, (middle) with a block sparsity-inducing penalty (blocks corresponding to poorly fitting models are zero), and (right) with a component sparsity-inducing penalty (rows corresponding to poorly fitting spectral components from different models are zero).

introduced a *relative* sparsity-inducing penalty and formulated it as:

$$\Psi_{\text{rel}}(\mathbf{H}) = \sum\nolimits_{j=1}^{J} \lambda_j \sum\nolimits_{g=1}^{G_j} \log\left(\frac{\epsilon + \|\mathbf{H}_{(j,g)}\|_1}{\|\mathbf{H}_{(j)}\|_1^{\gamma_j}}\right), \tag{1.15}$$

where γ_j are some non-negative constants. The penalty (1.15) can also be rewritten as

$$\Psi_{\text{rel}}(\mathbf{H}) = \Psi_{\text{gr}}(\mathbf{H}) - \sum\nolimits_{j=1}^{J} \lambda_j \gamma_j G_j \log\left(\|\mathbf{H}_{(j)}\|_1\right). \tag{1.16}$$

One can easily see that, while the new penalty keeps the group sparsity property thanks to $\Psi_{\text{gr}}(\mathbf{H})$ defined in (1.14), it prevents (when $\gamma_j > 0$) the supergroups from vanishing since if $\|\mathbf{H}_{(j)}\|_1$ tends to zero, then $-\log\left(\|\mathbf{H}_{(j)}\|_1\right)$ tends to $+\infty$. This formulation generalizes the group sparsity constraint in the sense that (1.15) reduces to (1.14) for $\gamma_j = 0$. One can then introduce either the *relative block sparsity-inducing penalty* or the *relative component sparsity-inducing penalty* by defining a group $\mathbf{H}_{(j,g)}$ to be either a block or a row in $\mathbf{H}$. Note that while we presented relative group sparsity within the context of NMF, the idea can also be extended to other dictionary decomposition schemes.

1.3.5 Algorithms for parameter estimation and results

In NMF formulation, multiplicative update (MU) rules are usually used for the parameter estimation as they are simple and they guarantee a the non-increasing value of the optimization function after each iteration [LS01, FBD09]. The derivation of such MU rules for the group sparsity is straightforward and almost identical to the one proposed in [LBF11], except that in our case the groups are defined differently and $\mathbf{W}$ is not updated. The overall algorithms for supervised case (criterion (1.12)) and semi-supervised case (criterion (1.13)) are summarized in Algorithms 1.1 and 1.2, respectively, without considering the formulas in red color. In these algorithms $\eta > 0$ is a constant parameter, $\mathbf{P}_{(j,g)}$ is a matrix of the same size as $\mathbf{H}_{(j,g)}$ whose entries have the same value, and $\mathbf{P}$ is a matrix concatenating all $\mathbf{P}_{(j,g)}$. When relative group sparsity is applied, some modifications in red color are added to take into account the effect of the group normalization.

We reported on-the-fly source separation results (including a user test) with the use of the (relative) group sparsity constraints in [8]. Later, these proposed group sparsity constraints were investigated in the context of single-channel speech enhancement in [7, 40]. More details about the algorithm derivation and experimental results can be found in our corresponding papers.

1.4 Multichannel audio source separation exploiting the GSSM

In multichannel setting, *i.e.,* when more microphones are available, additional information about the source locations can be exploited thanks to the phase and intensity differences of signals recorded at different microphones. Such *spatial* cues play an important role and are usually combined with *spectral* models to offer better source separation performance compared to the single-channel case. In my PhD, I proposed a spatial model named full-rank source spatial covariance matrices and investigated it within a Gaussian modeling framework for multichannel audio source separation [39]. The work was continued for some time after my PhD and we published a journal paper [42] where some prior knowledge about the source location is considered. In that work, we proposed two alternative probabilistic priors over the spatial covariance matrices, which are consistent with the theory of statistical room acoustics, and we derived EM algorithms for maximum a posteriori (MAP) estimation.

Motivated by the success of both the GSSM (for single-channel audio mixtures)

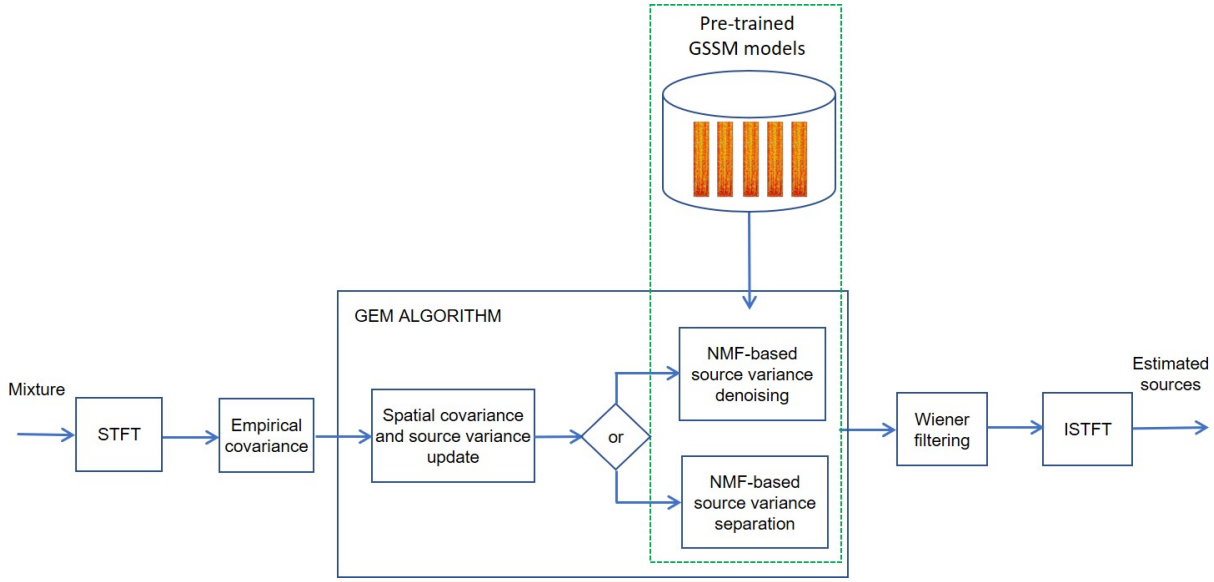


Figure 1.4: General workflow of the proposed approach. Green dashed boxes indicate the novelty compared to the existing works [OVB12, FSO17][41]

.

and the source spatial covariance model (for multichannel mixtures), we investigated their combination in multichannel audio source separation [11]. The general workflow is shown in Figure 1.4 and the contributions of the work are summarized as follows:

- We proposed two criteria to constrain the source variances in the GSSM-based Gaussian modeling framework.
- We derived algorithms for the parameter estimation, and studied their convergence and stability with respect to the parameter settings.
- We validated the effectiveness of the proposed approach in speech enhancement scenario using a benchmark dataset from the 2016 Signal Separation Evaluation Campaign (SiSEC 2016).

1.4.1 Local Gaussian modeling

Let us denote by $\mathbf{s}_j(n,f)$ the $I \times 1$ vector of the STFT coefficients of the contribution of j-th source at I microphones, where n is time frame index and f is the frequency bin. The mixing model in equation (1.1) is written in the frequency domain and in the

multichannel setting as:

$$\mathbf{x}(n,f) = \sum_{j=1}^{J} \mathbf{s}_j(n,f). \tag{1.17}$$

In the LGM, $\mathbf{s}_j(n,f)$ is modeled as a zero-mean complex Gaussian random vector with covariance matrix $\mathbf{\Sigma}_j(n,f) = \mathbb{E}(\mathbf{s}_j(n,f)\mathbf{s}_j^H(n,f))$, where ^{H} indicates the conjugate transposition. Such a covariance matrix is then factorized as

$$\mathbf{\Sigma}_j(n,f) = v_j(n,f)\,\mathbf{R}_j(f), \tag{1.18}$$

where $v_j(n,f)$ are scalar time-dependent *variances* encoding the spectro-temporal power of the sources and $\mathbf{R}_j(f)$ are time-independent $I \times I$ *spatial covariance matrices* encoding their spatial characteristics when sources and microphones are assumed to be static. Under the assumption that the source images are statistically independent, the mixture vector $\mathbf{x}(n,f)$ also follows a zero-mean multivariate complex Gaussian distribution with the covariance matrix computed as

$$\mathbf{\Sigma}_\mathbf{x}(n,f) = \sum_{j=1}^{J} v_j(n,f)\,\mathbf{R}_j(f). \tag{1.19}$$

With a further assumption that the mixture STFT coefficients at all time-frequency (T-F) bins are independent, the likelihood of the set of observed mixture vectors $\mathbf{x} = \{\mathbf{x}(n,f)\}_{n,f}$ given the set of parameters $\theta = \{v_j(n,f), \mathbf{R}_j(f)\}_{j,n,f}$ is given by

$$P(\mathbf{x}|\theta) = \prod_{n,f} \frac{1}{\det\left(\pi \mathbf{\Sigma}_\mathbf{x}(n,f)\right)} e^{-\mathrm{tr}(\mathbf{\Sigma}_\mathbf{x}^{-1}(n,f)\widehat{\mathbf{\Psi}}_\mathbf{x}(n,f))}, \tag{1.20}$$

where det represents the determinant of a matrix, tr() stands for matrix trace, and $\widehat{\mathbf{\Psi}}_\mathbf{x}(n,f) = \mathbb{E}(\mathbf{x}(n,f)\mathbf{x}^H(n,f))$ is the empirical covariance matrix, which can be numerically computed by local averaging over neighborhood of each T-F bin (n',f') as [43]. The parameters are then estimated by minimizing the negative log-likelihood:

$$\mathcal{L}(\theta) = \sum_{n,f} \mathrm{tr}\big(\mathbf{\Sigma}_\mathbf{x}^{-1}(n,f)\widehat{\mathbf{\Psi}}_\mathbf{x}(n,f)\big) + \log\det\big(\pi\mathbf{\Sigma}_\mathbf{x}(n,f)\big), \tag{1.21}$$

Under this model, once the parameters θ are estimated, the STFT coefficients of the source images are obtained in the minimum mean square error (MMSE) sense by

multichannel Wiener filtering as

$$\hat{\mathbf{s}}_j(n,f) = v_j(n,f)\,\mathbf{R}_j(f)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f). \tag{1.22}$$

Finally, the expected time-domain source images $\hat{\mathbf{s}}_j(t)$ are obtained by the inverse STFT of $\hat{\mathbf{s}}_j(n,f)$.

1.4.2 NMF-based source variance model

As can be seen in the previous section, NMF has been widely applied to single channel audio source separation where the mixture spectrogram is factorized into two latent matrices characterizing the spectral basis and the time activation [VBGB14, FBD09]. When adapting NMF to the considered LGM framework, the nonnegative source variances $v_j(n,f)$ can be approximated as [41]

$$v_j(n,f) = \sum_{k=1}^{K_j} w_{jfk} h_{jkn}, \tag{1.23}$$

where w_{jfk} is an entry of the spectral basis matrix $\mathbf{W}_{(j)}$, h_{jkn} is an entry of the activation matrix $\mathbf{H}_{(j)}$.

1.4.3 Source variance fitting with GSSM and group sparsity constraint

We proposed two strategies for the source variance fitting as follows.

Source variance denoising. The activation matrix is estimated by optimizing the criterion:

$$\underset{\mathbf{H}_{(j)}\geq 0}{min}\; D(\widehat{\mathbf{V}}_j \| \mathbf{W}_{(j)}\mathbf{H}_{(j)}) + \lambda\Psi(\mathbf{H}_{(j)}), \tag{1.24}$$

where $\widehat{\mathbf{V}}_j = \{v_j(n,f)\}_{n,f} \in \mathbb{R}_+^{F\times N}$ is the matrix of j-th source variances, $\mathbf{W}_{(j)}$ is constructed as (1.11) and fixed, $\Psi(\mathbf{H}_{(j)})$ represents a penalty function imposing sparsity on the activation matrix $\mathbf{H}_{(j)}$ (1.14), and λ is a trade-off parameter determining the contribution of the penalty.

Source variance separation. Let $\widetilde{\mathbf{V}} = \sum_{j=1}^{J} \widehat{\mathbf{V}}_j$ be the matrix of the total source variance estimate, it is decomposed by solving the following optimization problem

$$\underset{\mathbf{H}\geq 0}{min}\; D(\widetilde{\mathbf{V}} \| \mathbf{W}\mathbf{H}) + \lambda\Psi(\mathbf{H}) \tag{1.25}$$

where $\mathbf{H} = [\mathbf{H}_{(1)}^\top, ..., \mathbf{H}_{(J)}^\top]^\top \in \mathbb{R}_+^{K \times N}$, $K = \sum_{j=1}^J P_j$ the total number of rows in $\mathbf{H}$. This criterion can be seen as an additional NMF-based separation step applied on the source variances, while criterion (1.24) and other existing works [41][OVB12, FSO17] do not perform any additional separation of the variances, but more like denoising of the already separated variances.

Inspired by the advantage of two penalty functions inducing block and component sparsity (1.14) presented in Section 1.3, we investigated their combination in a more general form as

$$\Psi(\mathbf{H}) = \gamma \sum_{p=1}^{P} \log(\epsilon + \|\mathbf{H}_p\|_1) + (1-\gamma) \sum_{k=1}^{K} \log(\epsilon + \|\mathbf{h}_k\|_1), \tag{1.26}$$

where the first term on the right hand side of the equation represents the block sparsity-inducing penalty, the second term represents the component sparsity-inducing penalty, and $\gamma \in [0, 1]$ weights the contribution of each term. In (1.26), $\mathbf{h}_k \in \mathbb{R}_+^{1 \times N}$ is a row (or component) of $\mathbf{H}$, $\mathbf{H}_p$ is a subset of $\mathbf{H}$ representing the activation coefficients for p-th block, P is the total number of blocks, ϵ is a non-zero constant, and $\|.\|_1$ denotes ℓ_1-norm operator. In the considered setting, a block represents one training example for a source and P is the total number of used examples (*i.e.*, $P = \sum_{j=1}^J P_j$). Similar formula can also be written for the $\Psi(\mathbf{H}_{(j)})$ in (1.24).

1.4.4 Algorithms for parameter estimation and results

By putting (1.26) into (1.25), we now have a complete criterion for estimating the activation matrix $\mathbf{H}$ given $\widetilde{\mathbf{V}}$ and the pre-trained spectral model $\mathbf{W}$ in the source variance separation case. Similar procedure can be derived for the source variance denoising case (1.24). Within the LGM, a generalized EM algorithm has been used to estimate the parameters $\{v_j(n,f), \mathbf{R}_j(f)\}_{j,n,f}$ by considering the set of hidden STFT coefficients of all the source images $\{\mathbf{c}_j(n,f)\}_{n,f}$ as the *complete data*. For the proposed approach as far as the GSSM concerned, the E-step of the algorithm remains the same. In the M-step, we additionally perform the optimization defined either by (1.24) or by (1.25). This is done by the MU rules so that the estimated intermediate source variances $v_j(n,f)$ are further updated with the supervision of the GSSM.

We validated the performance and properties of the proposed approach in speech enhancement use case where we know already two types of sources in the mixture: speech and noise. For better comparison with the state of the art, we used the benchmark development dataset of the "Two-channel mixtures of speech and real-world background

Methods	Ca1		Sq1		Su1		Average	
	SDR OPS	SIR IPS	SDR OPS	SIR IPS	SDR OPS	SIR IPS	SDR OPS	SIR IPS
Liu*	-1.0 9.5	4.9 16.8	-8.5 14.2	-2.9 18.9	-12.8 21.2	-8.0 15.7	-7.0 14.2	-1.4 17.5
Le Magoarou* [MOD14]	9.2 31.3	11.6 29.3	4.0 38.9	6.2 45.2	-5.2 22.9	-4.5 24.6	3.7 32.8	5.6 35.3
Rafii* [RP13]	8.8 29.2	13.0 27.3	6.2 34.6	9.6 38.7	-2.7 23.9	-2.7 21.6	5.1 30.4	8.0 31.1
Ito* [IAN13]	7.2 -	25.9 -	23.7 -	9.1 -	5.6 -	- -	7.4 -	- -
Wood* [WR16]	3.0 33.7	9.4 60.7	1.9 38.6	2.4 60.5	0.2 25.9	-2.6 47.6	1.9 34.1	3.6 57.7
Arberet [41][OVB12]	9.1 13.3	10.0 10.9	3.3 8.3	3.3 10.5	-0.2 10.2	-1.2 3.7	4.4 10.4	4.6 9.1
GSSM + SV denoising $(\lambda = 10, \gamma = 0.2)$	10.5 8.4	11.8 12.7	7.0 8.5	8.5 14.7	5.1 11.3	5.6 7.8	7.7 18.1	9.0 12.5
GSSM + SV separation $(\lambda = 10, \gamma = 0.2)$	10.6 11.4	13.5 13.0	7.8 31.6	11.1 31.4	5.0 23.7	7.1 27.8	**8.1** **23.1**	**11.0** **24.5**

Table 1.1: Speech separation performance obtained on the devset of the BGN task of the SiSEC campaign. * indicates submissions by the authors and "-" indicates missing information.

noise" (BGN) task[1] within the SiSEC 2016 [LSR+17]. This devset contains stereo mixtures of 10 second duration and 16 kHz sampling rate. They were mixtures of male/female speeches and real-world noises recorded from different public environments: cafeteria (Ca), square (Sq), and subway (Su). Table 1.1 shows the speech separation performance in terms of the signal-to-distortion ratio (SDR), the signal to interference ratio (SIR), the overall perceptual score (OPS), and the target-related perceptual score (TPS) [VGF06, EVHH12], the higher the better, obtained by the proposed approaches and other state-of-the-art methods in the SiSEC campaign.

We also investigated the algorithm convergence by varying the number of EM and MU iterations, and observed that with 10 to 25 MU iterations, the algorithm converges nicely and saturates after about 10 EM iterations. We further investigated the separation results with different choice of the hyper-parameters λ and γ. The algorithm is less sensitive to the choice of γ, while more sensitive to the choice of λ and $\lambda > 10$ greatly decreases the separation performance. The best choice for these parameters in

[1] https://sisec.inria.fr/sisec-2016/bgn-2016/

term of the SDR are $\lambda = 10, \gamma = 0.2$. Please refer to the paper [11] for details about the algorithm derivation and evaluations.

1.5 Other contributions in audio source separation

1.5.1 Text-informed source separation

In this work, we presented a novel text-informed framework in which textual information in form of text transcript associated with speech source is used to guide its separation from other sources in the mixture. The separation workflow is as follows. First, a speech example is artificially generated via either a speech synthesizer or by a human reading the text. Then, this example is used to guide source separation. For that purpose, a new variant of the non-negative matrix partial co-factorization (NMPCF) model based on an excitation-filter channel speech model is introduced. Such a modeling allows coupling the linguistic information between the speech example and the speech in the mixture. The corresponding multiplicative update (MU) rules are eventually derived for the estimation of the parameters. We performed extensive experiments to assess the effectiveness of the proposed approach in terms of source separation and alignment performance [2].

1.5.2 Interactive user-guided source separation

In order to boost the source separation performance in real-world post-production application, we considered a temporal annotation of the source activity along the mixture given by a user. We then proposed weighting strategies incorporated in the NMF formulation so as to better exploit such annotation to guide the separation process [4]. A video demonstration is online[1]. In another work [5], we proposed an interactive source separation framework that allows end-users to provide feedback at each separation step so as to gradually improve the result. A prototype graphical user interface (GUI) is developed to help users annotating time-frequency regions where a source can be labeled as either active, inactive, or well-separated within the displayed spectrogram. Such user feedback information is then taken into account in an uncertainty-based learning algorithm to constraint the source estimates in a next separation step. Both the considered approaches were based on non-negative matrix factorization and were shown to be effective in real-world settings.

[1] https://www.youtube.com/watch?v=EjpLKvphpMot=16s

1.5.3 Informed source separation via compressive graph signal sampling

In this work, we investigated a novel informed source separation method for audio object coding based on a recent sampling theory for smooth signals on graphs. At the encoder, we assume to know the original sources, and thus the ideal binary time-frequency (T-F) mask considering only one source is active at each T-F point. This ideal mask is then sampled with a compressive graph signal sampling strategy that guarantees accurate and stable recovery in order to perform source separation at the decoder side. The graph can be built using feature vectors, computed using non-negative matrix factorization at both encoder and decoder sides. We show in our paper [9] that the proposed approach performs better than the state-of-the-art methods at low bitrate.

1.6 Conclusion

In this chapter we have presented the application of NMF model in audio source separation. We have considered single-channel case where some novel sparsity-inducing constraints were proposed to extract relevant spectral patterns from an over-complete source spectral dictionary. We have also extended the work to multi-channel settings within the local Gaussian modeling framework. Some other works on informed audio source separation have also been mentioned, which was done in close collaboration with Technicolor production services for the real use cases.

Chapter 2

Audio-visual scene analysis

Supervision: Sanjeel Parekh (PhD student)
Main collaborators: Alexey Ozerov (Technicolor, InterDigital), Slim Essid (Telecoms ParisTech), Patrick Pérez (Technicolor, Valeo.ai), Gaël Richard (Telecoms ParisTech).

This chapter summarizes part of the work done during the PhD of Sanjeel Parekh (2016-2019) and presented in the journal paper [22]. The work focuses on multimodal machine learning approach for audio-visual event identification and localization, and visual informed audio source separation. The organization of the chapter is as follows. We begin by briefly discussing the motivation and related works in Section 2.1. Then the proposed weakly supervised representation learning framework and its application for tackling classification and localization is described in Section 2.2. This is followed by some implementation details in Section 2.3. Results on benchmark datasets (*i.e.*, the DCASE smart cars challenge [MHD+17a] and the instrument dataset [KCS+17]) are discussed in Section 2.4. Finally, we conclude in Section 2.5.

2.1 Motivation and related works

Audio and visual cues appear everywhere in real life, and as humans we have great ability to perceive such information in order to analyse and understand the surrounding scenes. As an example, when a car passes by, we can instantly identify both audio and visual components that characterize this event. In many cases, information from audio cues can help better perceiving visual information and vice versa. For building machines with such scene analysis and understanding capabilities, it is important to design systems for jointly exploiting both audio and visual cues. Such approaches

should be able to learn meaningful audio-visual (AV) representations from large-scale real-world data. This work presents a step in that potential direction. We formulated the considered AV problem (shown in Figure 2.1) as follows. Given a video labeled as "train horn", we would like to:

- (1) identify the event (classification problem);
- (2) localize its visual presences and the associated temporal audio segment(s) (localization problem);
- (3) separate the target sound source from the others (source separation problem).

To seek a unified solution, we opt for a weakly supervised learning approach which exploits audio-visual data with only general video-level event labels without temporal and spatial information about the AV events. As the train horn may sound before or after the train is visible, the targeting model, when designed, must be able to deal with such *unsynchronized* AV events.

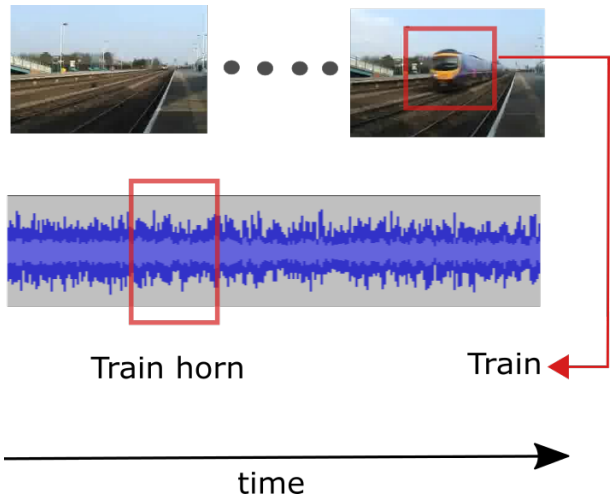


Figure 2.1: Pictorial representation of the considered problem (figure is from [22]).

When this work was started, to our best knowledge, there was no existing literature that use a weakly supervised multimodal deep learning framework to address all three targets mentioned above. However, there were relevant works in audio community for audio scene analysis [MHDV15, ZZHJH10, MHD+17b] and computer vision community for visual object localization and classification [ZPV06, BPT14, OBLS15, KOCL16, BV16]. Another line of work exploits audio-visual correlations thanks to feature space transformation techniques such as canonical correlation analysis (CCA) for localization

and representation learning in general [ISS13, KSE05]. We also started some works on motion-informed and video-guided audio source separation [18, 19]. It is worth noting that the field has been growing rapidly. Some parallel works have learnt meaningful representation through audio–image correlations in an unsupervised manner [AZ17, OIM+16, OWM+16]. Other works have exploited the multiple instant learning (MIL) and attention-based architectures [KYI+19, CWSB19, WLM19] for the similar weakly supervised learning paradigm. Similarly for audio source separation, parallel progress reveals several visual-guided systems using deep learning paradigm [ZGR+18, EML+18, GFG18]. While we share the high-level goal of weakly-supervised representation learning with some existing works, our multimodal design and our audio sub-module, as discussed in the next section, is significantly different.

2.2 Weakly supervised representation learning framework

In order to tackle simultaneously all three tasks mentioned above, in [22] we propose a novel multimodal framework based on multiple instance learning (MIL). Such MIL-based framework uses class-agnostic proposals from both video frames and audio, and exploits weak labels from data to learn a robust joint audio-visual representation. The general workflow is shown in Figure 2.2 and comprises four major steps: (a) audio/visual proposal extraction, (b) higher-level feature extraction and learning, (c) proposal scoring and aggregation, and finally (d) fusion and classification. Thanks to the cascaded deep neural network (DNN) layers, the model can be trained end-to-end using only video-level event labels without any timing information in a supervised way.

Let us model a video V as a bag of M selected visual proposals $\mathcal{R} = \{r_1, r_2, \ldots, r_M\}$ (*e.g.,* image regions obtained from sub-sampled frames in our experiment), and S audio proposals, $\mathcal{A} = \{a_1, a_2, \ldots, a_S\}$ (*e.g.,* temporal segments from the original sound track or from the separated signals constructed by NMF components in our experiment). Such visual and audio proposals are passed through DNN blocks to extract the corresponding features. Then we adopt two-stream (localization W_{loc} and classification W_{cls}) architecture proposed by Bilen *et al.* [BV16] for scoring each of the feature with respect to the classes. This architecture allows the localization layer to choose the most relevant proposals for each class. Subsequently, the classification stream output is multiplied with an attention weight through element-wise multiplication. The class scores over the video are obtained by summing the resulting weighted scores from all proposals. After performing the above stated operations for both audio and visual sub-modules, in the final step, the global video-level scores are ℓ_2 normalized and added.

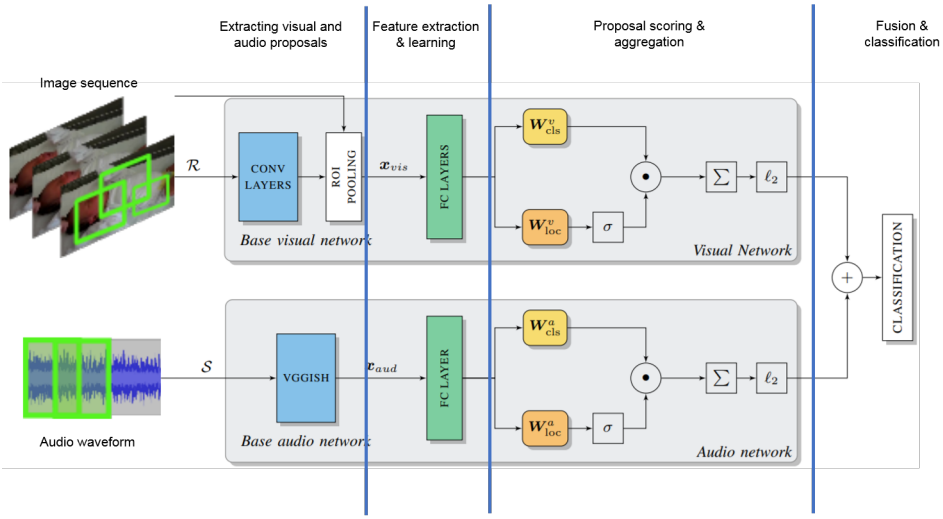


Figure 2.2: General workflow of the proposed weakly supervised representation learning approach.

Given a set of L training videos and corresponding groundtruth labels organized in C classes, $\{(V^{(l)}, y^{(l)})\}_{l=1}^{L}$, where $y \in \mathcal{Y} = \{-1, +1\}^C$ with the class presence denoted by $+1$ and absence by -1, we implement a multi-label classification problem. Both audio and visual sub-modules are trained jointly using the multi-label hinge loss on a batch of size B:

$$\mathcal{L}(w) = \frac{1}{CB} \sum_{l=1}^{B} \sum_{c=1}^{C} \max\Big(0, 1 - y_c^{(l)} \phi_c(V^{(l)}; w)\Big). \tag{2.1}$$

where w denotes all network parameters (weights and biases), $\phi_c(V^{(l)}; w) \in \mathbb{R}^C$ is the predicted output label.

2.3 Some implementation details and variant

Visual proposals and feature extraction. We sub-sampled frame sequences of each video at a rate of 1 frame per second, then generated class-agnostic region proposals on each extracted frame using EdgeBoxes algorithm [ZD14]. EdgeBoxes additionally generated a confidence score for each bounding box characterizing a region proposal. To reduce the computational load and redundancy, we used this score to select the top M_{img} proposals from each sampled frame and use them for feature extraction. A fixed-length feature vector is obtained from each region proposal using Fast-RCNN implementation

[Gir15] (a CNN with a region-of-interest (RoI) pooling layer). The feature vectors extracted after RoI pooling layer are passed through two fully connected layers, which are learned during training.

Audio proposals and feature extraction. For the localization and classification task, we extracted M_{aud} audio proposals as temporal segments with fixed-length window along the audio log-Mel spectrogram. The window length is 960 ms with 50% overlapping. Each log-Mel spectrogram proposal $a_s \in \mathbb{R}^{96\times 64}$ with 64 Mel-bands and 96 temporal frames is passed through a VGGish deep network [HCE+17b] to generate a 128 dimensional embedding as a base audio feature. This network was pre-trained on a YouTube-8M dataset [AEHKL+16] for audio classification based on video tags. Similar to visual part, prior to proposal scoring by two-stream module, the generated audio embeddings are passed through a fully-connected layer that is learned during training.

For source separation task as a variant, the STFT magnitude spectrogram $\mathbf{X}$ of the original audio track is first decomposed into K non-negative components as

$$\mathbf{X} \approx \sum_{k=1}^{K} \mathbf{w}_k \mathbf{h}_k, \tag{2.2}$$

where $\mathbf{w}_k$ and $\mathbf{h}_k$ represent k-th spectral pattern and its temporal activation, respectively. Then K NMF tracks are obtained from $\mathbf{w}_k, \mathbf{h}_k$ for $k \in [1, K]$ by the inverse STFT with the phase of the original track. These tracks are chunked into temporal segments similar to the original tracks to obtain more $M_{\text{aud}} \times K$ audio proposals. Such proposals are passed through the same VGGish audio network to generate embeddings. When performing NMF blindly, we do not know which NMF components belong to an audio source. However, the two stream architecture in the considered system will help to score each NMF component with respect to its relevance for a particular class $c \in C$. These relevance scores can then be appropriately aggregated to obtain a global score (denoted as α_k^c), which weights the contribution of k-th NMF component on a targeted audio source. Several aggregation strategies are discussed in detail in [22]. Finally source separation can be obtained as:

$$\mathbf{S}_c = \frac{\sum_{k=1}^{K} \alpha_k^c \mathbf{w}_k \mathbf{h}_k}{\sum_{k=1}^{K} \mathbf{w}_k \mathbf{h}_k} \mathbf{X} \tag{2.3}$$

Here $\mathbf{S}_c$ is the estimate of c-th source and is converted back to the time domain using the inverse STFT.

Table 2.1: Results on DCASE smart cars task test set [MHD+17a]. We report here the averaged F1 score, precision and recall values, and compare with state-of-the-art approaches. TS is an acronym for two-stream (table is from [22]).

	System	F1	Precision	Recall
(a)	AV Two Stream	**64.2**	59.7	**69.4**
(b)	Sync. AV Two Stream	62.0	57.2	67.6
(c)	TS Audio-Only	57.3	53.2	62.0
(d)	TS Video-Only	47.3	48.5	46.1
(e)	TS Video-Only WSDDN-Type [BV16]	48.8	47.6	50.1
(f)	AV One Stream	55.3	50.4	61.2
(g)	CVSSP - Fusion system [XKWP17]	55.6	**61.4**	50.8
(h)	CVSSP - Gated-CRNN-logMel [XKWP17]	54.2	58.9	50.2

2.4 Results

In the experiment with different settings in the paper [22], we showed that the learnt representations are useful for performing several tasks such as event/object classification, audio event detection, audio source separation, and visual object localization. We also demonstrated the model's capacity to learn from unsynchronized audio-visual events. State-of-the-art classification results are achieved on the Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 smart cars dataset [MHD+17a] with promising generalization to diverse object types such as musical instruments. As an example, Table 2.1 compares the event classification performance on the DCASE dataset of the proposed approach with several strong baselines (audio-only two stream architecture, video-only two stream architectures, AV one stream architecture) and state-of-the-art systems [XKWP17, BV16].

Figure 2.3 shows qualitative examples of the localization result (displayed through a yellow bounding box) in extreme asynchronous conditions. The video frames are placed above while the normalized audio localization heatmap at the bottom displays the scores assigned to each temporal audio segment. As expected in this temporal asynchronous case **A**, the system does not detect any yellow edges in the first frame. But the car is detected much later when it is completely visible. **B** depicts an example, where due to extreme lighting conditions the visual object is not visible. Here too, the system localizes the audio object and correctly predicts the 'motorcycle' class.

Following the PhD work of Sanjeel Parekh, with Valentin Bilot, a Master's intern,

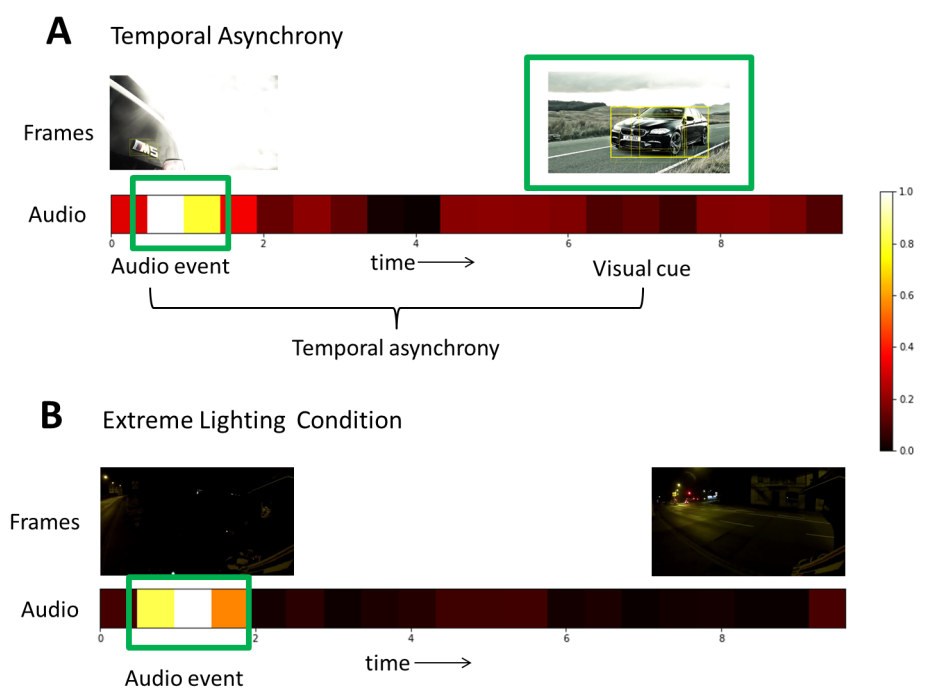


Figure 2.3: Qualitative results for unsynchronized AV events. For both the cases A and B, the heatmap at the bottom denotes audio localization over segments for the class under consideration. For heatmap display, the audio localization vector has been scaled to lie between [0,1]. The top row depicts video frames roughly aligned to the audio temporal axis. (A) Top: Here we show a video where the visual object of interest appears after the audio event. This is a 'car' video from the validation split. The video frames show bounding boxes where edge opacity is controlled by the box's detection score. In other words, higher score implies better visibility (B) Bottom: This is a case from the evaluation data where due to lighting conditions, the visual object is not visible. However the system correctly localizes in audio and predicts the 'motorcycle' class (figure is from [22]).

we used the presented two-stream framework for audio scene classification task and participated in the DCASE 2019 challenge. Classification results obtained with several settings and a late fusion of several models were reported in [38].

2.5 Conclusion

We have presented a deep AV scene understanding approach that can be trained jointly using weak labels. The proposed framework can perform several tasks (*i.e.,* classification, localization, source separation) simultaneously. State-of-the-art event classification

and detection performance is reported on the DCASE 2017 smart cars data and qualitative result confirms that the framework can tackle challenging case of unsynchronized AV events. As research in DNN advances quickly, the considered framework can benefit from other novel model-specific feature extraction techniques and modality fusion networks.

Chapter 3

Media interestingness and memorability

Supervision: Roman Cohendet (postdoc), Karthik Yadati (PhD intern), Yuesong Shen (MSc intern), Eloïse Berson (MSc intern), Hammad Squalli-Houssain (MSc intern)
Main collaborators: Claire-Hélène Demarty (Technicolor, InterDigital), Mats Sjöberg (Aalto University), Bogdan Ionescu (University Politehnica of Bucharest), Thanh-Toan Do (University of Liverpool).

Understanding and predicting user perceptions on visual content is a very active research domain in multimedia and computer vision communities. It offers a wide range of practical applications in content retrieval, education, summarization, advertising, content filtering, recommendation systems, *etc.* With recent advances in machine learning, such visual perceptions have moved from low-level features (such as intensity, colors, saliency) [FRC10] towards challenging high level subjective concepts such as visual aesthetics [BSS10, DOB11], emotion [RDP+11, SCKP09], popularity [KSH13, CMS15], interestingness [GGR+13, ZdJSJ16, DSC+17, CRZI19], and memorability [IXTO11, SZM+17a, EVC20]. In this chapter, we will summarize our pioneering works in two emerging concepts: interestingness and memorability, both for image and video. The contributions cover dataset construction, analysis, and prediction via machine learning models. Major impacts to the research community include the release of the large-scale datasets such as Interestingness10k [31, 34] and VideoMem10k [29]; the organization of two MediaEval benchmark campaigns: media interestingness prediction task (2016, 2017) [30, 31] and video memorability prediction task (2018, 2019) [32, 33]; and the pioneering work on video memorability [28, 29].

The chapter is organized as follows. We first briefly summarize our work on image and video interestingness in Section 3.1. We then discuss more in-depth work on video memorability in Section 3.2. Finally we draw conclusions in Section 3.3.

3.1 Image and video interestingness

Interestingness usually refers to arousing interest, curiosity, as well as the *ability of holding or catching attention* [Ber60]. Existing studies in psychology and vision research [CDP01, EI08] revealed that interest is determined by certain factors like *novelty*, *uncertainty*, *conflict*, *complexity*, and their combinations. This finding was also supported in the appraisal theory presented in [Sil05], where the author explained that appraisals like the novelty, the comprehensibility, and the complexity of an event are likely to arouse interest in this event. However, understanding and predicting visual interestingness remains challenging as its judgment is highly subjective and usually context-dependent. Following the literature, we distinguished two different notions, namely *socially-driven* interestingness and *content-driven* interestingness. The former is derived from media sharing websites such as Flickr[1] and Pinterest[2], where contextual information may greatly affect the judgement. The latter refers to human annotations that assess interestingness solely on the perceived media content. As an example, in our annotation protocol for content-driven interestingness, users only view two images or videos side by side on the screen and vote for which one they are more interested [30]. Our work brings contributions in both these notions and are summarized as follows:

- In [30, 31] we proposed a Predicting Media Interestingness Task in 2016 and 2017 within the MediaEval benchmarch[3]. For this purpose, we built a first publicly available[4] content-driven interestingness dataset for both images and videos based on a real-world Video on Demand (VOD) use case scenario. This Interestingness10k dataset contains 9,831 images and more than 4 hours of video, interestigness scores determined based on more than 1M pair-wise annotations of about 800 trusted annotators around the world. Figure 3.1 and Figure 3.2 depict examples of the images/videos, their interestingness scores, and the annotation agreements from the dataset. The interestingness prediction task greatly attracted the multimedia research community as shown by the largest number of

[1] https://www.flickr.com/
[2] https://www.pinterest.com/
[3] http://www.multimediaeval.org/
[4] https://www.interdigital.com/data_sets/intrestingness-dataset

international participants compared to other MediaEval tasks in the same years. By analizing the dataset and the prediction systems, we provided an in-depth analysis of the crucial components for visual interestingness prediction in a book chapter [25] and a journal paper in revision [34].

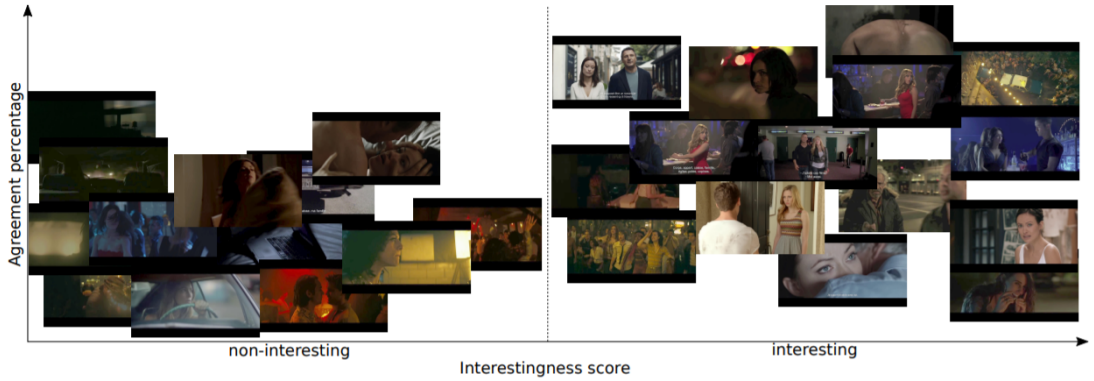


Figure 3.1: Examples from the Interestingnes10k image dataset: images annotated as interesting are on the right, whereas non-interesting images are on the left. Images at the top have higher annotation agreement, while images at the bottom have lower annotation agreement (figure is from [34]).

Figure 3.2: Examples from the Interestingnes10k video dataset: videos annotated as interesting are on the right, whereas non-interesting videos are on the left. Videos at the top have higher annotation agreement, while videos at the bottom have lower annotation agreement. Each video is depicted with a key-frame (figure is from [34]).

- In [44, 23] we applied DNN techniques for multimodal video interestingness prediction on both socially-driven dataset (*i.e.,* Flickr videos) and content-driven Interestingness10k dataset [30]. The workflow of the investigated approach is shown in Figure 3.3. We tested various deep neural DNN architectures, includ-

ing our proposed one combining several recurrent neural networks (RNNs), so as to handle several temporal samples at the same time. We then investigated different strategies for dealing with unbalanced dataset to improve the prediction results. We found that multimodality, as the mid-level fusion of audio and visual information, brings benefit to the task.

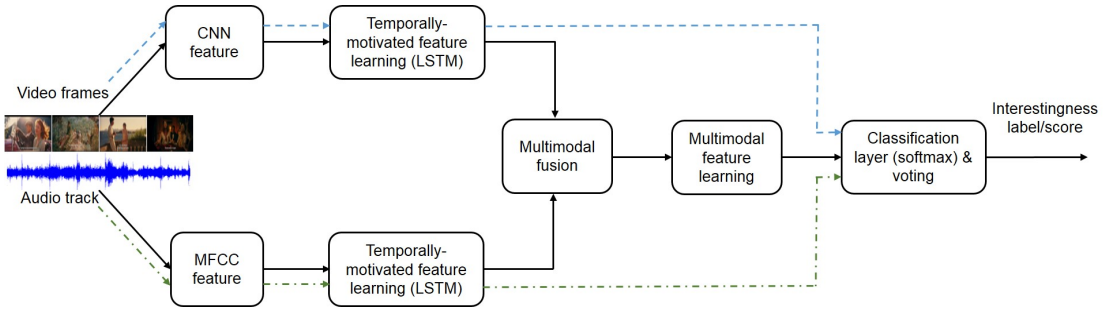


Figure 3.3: Proposed computational models for video interestingness prediction. Black arrows represent the workflow for our multimodal approach, whereas blue dash lines and green dash-dot lines represent monomodal workflows for visual-based and audio-based systems, respectively (figure is from [23]).

- In [24, 26] we focused on socially-driven interestingness and collected a large-scale interestingness dataset (LaFin) from Flickr website (images and their associated metadata), which was released for public use[1]. Given this LaFin dataset, where the ground-truth interestingness label is given by Flickr, we studied factors that may affect the users' interest such as image tags, titles, comments, and built deep learning-based computational models for the interestingness prediction. Our DNN models exploited different types of features such as VGG16 [SZ14], semantic features derived from an image captioning system [KSZ14], Word2Vec features [MCCD13] computed from Flickr tags, and their combinations. Table 3.1 shows that exploiting relevant contextual information derived from social metadata (*i.e.,* Flickr tags) could greatly improve the prediction result, and offer the accuracy over 92% in both training, validation, and test set.

3.2 Video memorability

While some contents have the power to be stored in our memories for a long time, others are quickly forgotten. Memorability of media contents such as images and videos has

[1]https://www.interdigital.com/data_sets/intrestingness-dataset

Inputs and *features*	Accuracy (%)		
	Train	Validation	Test
Image *VGG16*	78.07	75.54	76.45
Image *IC*	76.3	75.53	76.35
Flickr tags *Word2Vec*	89.96	89.68	89.63
Generated tags *Word2Vec*	65.42	63.47	65.12
Image *VGG16+IC*	85.27	83.34	83.59
Image + Generated tags *VGG16+Word2Vec*	78.82	76.48	75.65
Image+Flickr tags *VGG16+Word2Vec*	**92.76**	**90.99**	**91.08**
Image+Flickr tags *VGG16+IC+Word2Vec*	**93.72**	**92.46**	**92.59**

Table 3.1: Prediction results in terms of accuracy obtained by models with different sets of input features on our constructed LaFin dataset [26]. Here VGG16 and Image Captioning (IC) features are used for image, Word2Vec feature is used for Flickr tags and generated tags. The best performing system (last row) is multimodal where both VGG16, IC, and Word2Vec features are exploited.

become an important research subject not only for psychologists, behavior specialists, but also for computer scientists. Psychological literature has highlighted several factors which have critical influence on long-term memory, including emotion [KS08], visual attention [Cow98], semantics [Qui66], demographic information [CGDSLC16], or passage of time [McG00]. These factors have indeed provided computer vision researchers with insights to craft valuable computational features for the media memorability prediction [MLM13, IXP+14]. In computer vision and machine learning, the seminal work of Isola *et al.* [IXTO11] defined image memorability (IM) as the probability for an image to be recognized a few minutes after a single view, when presented amidst a stream of images. This definition has been widely accepted within subsequent work [KYP13, CEE13, KRTO15, LEOEQ16], and was adapted to short-term video memorability (VM) in our work [29]. Various machine learning models have been investigated for the prediction of image memorability [IXTO11, KRTO15, FAMR18], and more recently video memorability [HCS+15, SSS+17]. Recent studies also showed that style

transfer can be used to increase image memorability [SZM+17b, SZM+19]. We first worked on image memorability [27] and obtained a better prediction performance than the state of the art on the well-known LaMem dataset [KRTO15]. However, our more significant and pioneer contributions are on video memorability (VM), which will be presented in the remainder of this section.

3.2.1 VM dataset creation

MovieMem660. As research on computational understanding of video memorability is in its early stage, there is no publicly available dataset for modelling purposes. A few previous attempts provided protocols to collect video memorability data that would be difficult to generalize [HCS+15, SSS+17]. In [28] we presented a very first work on long-term video memorability where we measured the memory performances of participants from weeks to years after memorization to build a dataset of 660 videos. The videos were chosen as follows. We first established a list of 100 occidental movies, taking care of mixing popularity and genres. We then manually selected seven videos of 10 seconds from each movie. To maintain a high intra video semantic cohesion, we did not make cuts that would impair the understanding of the scene, nor did we aggregate shots that belong to different scenes. 104 people (22 to 58 years of age; age average = 37.1; 26% female; mostly educated people) participated in the experiment in a well-controlled environment (a quiet room equipped with subdued lights, the videos with high quality were displayed on a 60 inch monitor). The participants were first asked to fill a questionnaire during about 20 minutes about whether they remembered watching fully the movie, their confidence about the answer, the number of times they saw the movie, and when was the last time they saw the movie. Based on the answers to the questionnaire, an algorithm automatically selected 80 targets (*i.e.,* videos from seen movies) and 40 fillers (*i.e.,* videos from never seen movies) among the movies associated with the highest degree of certitude, with a maximum of two videos from the same movie. The fillers enabled to quantify the reliability of the annotations. Given such 120 videos selected, participants performed a recognition task where they saw the videos separated by an inter-stimuli interval of 2 seconds. On average, each video of our dataset had been viewed as a target by 10.7 participants, which corresponds to the mean number of observations that enter into the calculation of a memorability score.

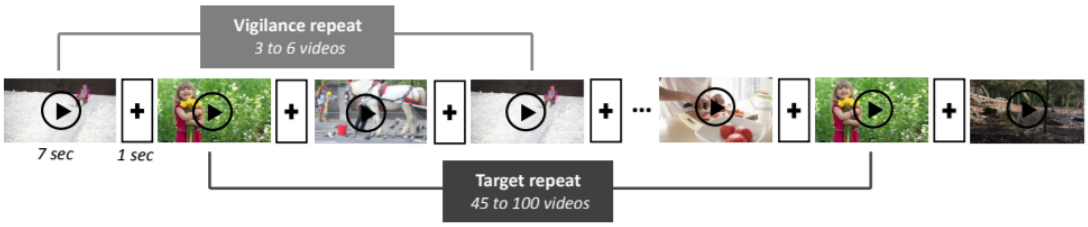
The memorability score assigned to each video is simply defined as the correct recognition rate of the video when viewed as target. Please refer to [28] for more details about the video selection, the annotation protocol, the memorability score calculation,

and the dataset analysis. This dataset, together with a list of pre-computed features (C3D[1], AudioSet[2], SentiBank[3], Affect [HX05], Image captions[4]), are made available for the research community[5].

VideoMem10k. This is the first large scale VM dataset, which is composed of 10,000 soundless videos of 7 seconds extracted from raw footage used by professionals when creating content. Videos contain only one semantic scene, but the scenes are varied (animal, food and beverages, nature, people, transportation, *etc*). Unlike the MovieMem660, our proposed protocol to annotate the VideoMem10k relies on crowdsourcing and is inspired by the protocol introduced in [IXP+14, IXTO11] for image. The protocol includes two steps to measure both human short-term and long-term memory performances for videos, and is shown in Figure 3.4. Step #1 consists of interlaced viewing and recognition tasks. Participants hired via Amazon Mechanical Turk (AMT) crowdsourcing platform watch a series of videos, some of them – the *targets* – repeated after a few minutes. Their task is to press the space bar whenever they recognize a video. Once the space bar is pressed, the next video is displayed, otherwise the current video continues till its end. Each participant watched 180 videos, that contain 40 *targets*, repeated once for memory testing, 80 *fillers* (*i.e.,* non target videos), and 20 so-called *vigilance fillers* which were repeated quickly after their first occurrence to monitor the participant's attention to the task. Step #2 took place 24 to 72 hours after step #1: the same participants performed similar recognition task to collect long-term annotations. They watched a new sequence of 120 videos, composed of 80 *fillers* (randomly chosen totally new videos) and 40 *targets* (randomly selected from the *non-vigilance fillers* of step #1). Note that, to guarantee the quality of the annotation, we used several controls: the vigilance task (step #1), a minimum correct recognition rate (15%, step #2), a maximum false alarm rate (30% for step #1; 40% for step #2), and a false alarm rate lower than the recognition rate (step #2 only). Finally, we had 9,402 participants for short-term, and 3,246 participants for long-term memorability who passed the vigilance controls. On average, a video was viewed as a repeated target 38 times (and at least 30 times) for the short-term task, and 13 times (at least 9 times) for the long-term task due to the lower number of participants in step #2.

We assigned a first raw memorability score to each video, defined as the percentage of correct recognitions by participants, for both short-term and long-term memorability.

[1] https://github.com/facebook/C3D
[2] https://github.com/tensorflow/models/tree/master/research/audioset
[3] http://www.ee.columbia.edu/ln/dvmm/vso/download/sentibank.html
[4] https://github.com/karpathy/neuraltalk2
[5] https://www.interdigital.com/data_sets/movie-memorability-dataset



(a) Step #1. Interlaced encoding and recognition tasks.

(b) Step #2. Second recognition task after 24 to 72 hours.

Figure 3.4: Proposed protocol to collect both short-term and long-term video memorability annotations. The second recognition task measures memory of videos viewed as fillers during step #1, to collect long-term memorability annotations (figure is from [29]).

Then, motivated by a similar work for image [IXP+14], the short-term raw scores were further refined by applying a linear transformation that takes into account the memory retention duration. Please refer to [29] for more details about the video selection, the annotation protocol, the memorability score calculation, and the dataset analysis. This dataset, together with a list of pre-computed features (C3D[1], HMP [CGR14], Inception-V3, Aesthetic visual features, *etc.*) are made available for the research community[2].

3.2.2 VM understanding

From two datasets, *i.e.,* MovieMem660 and VideoMem10k, we investigated a number of factors concerning the video memorability such as mean correct recognition rate, the memorability consistency over time, the memorability with respect to response time, the human vs. annotation consistency. Please refer to the papers [28, 29] for more details about the investigation. As an example, Figure 3.5 shows the mean correct recognition rate as a function of the retention interval between the memorization (*i.e.,* last view of video) and the measure of memorability performance for the MovieMem660 dataset (left) and the VideoMem10k dataset (middle for short-term VM, right for long-term VM). In line with other findings and as expected, recognition rate decreases linearly over time for the short-term, while long-term memory performances does not significantly

[1]https://github.com/facebook/C3D
[2]https://www.interdigital.com/data_sets/video-memorability-dataset

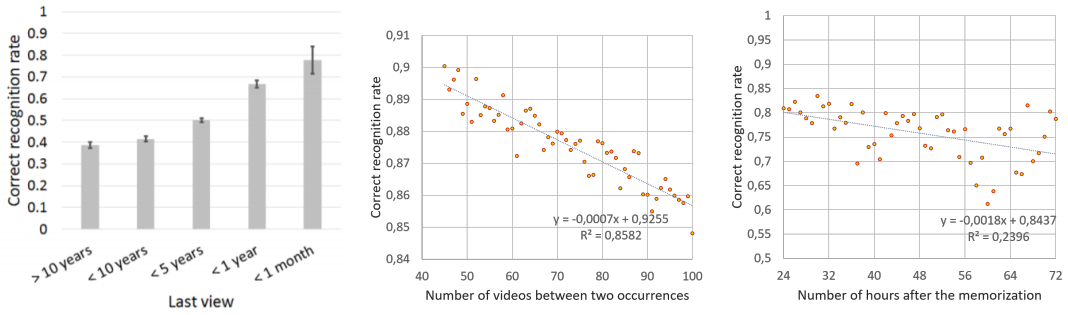


Figure 3.5: Mean correct recognition rate *vs.* the retention interval between the memorization and the measure of memory performance [28, 29]. Left: MovieMem660 dataset; Middle: Recognition rate decreases linearly over time for the short-term VM in VideoMem10k dataset; Right: long-term memory performances does not significantly change between 24 and 72 hours after memorization in VideoMem10k. Blue lines represent linear fitting.

change between 24 and 72 hours after memorization.

Human consistency vs. annotation consistency

We first followed the method proposed in [IXP+14] for IM to measure human consistency when assessing VM. For this purpose, we randomly split our participants in each dataset into two groups of same size and computed VM scores independently in each group. A Spearman's rank correlation between the two groups of scores was computed and averaged over 25 random half-split trials. Figure 3.6-left shows this human consistency as a function of the mean number of annotations per video for MovieMem660 dataset. The consistency of 0.70 is achieved from only about 18 annotations. This number is comparable to the maximal consistency obtained in image memorability (0.75 in [IXTO11] and 0.68 in [KRTO15] with 80 annotations), but with much less number of annotations. This may be explained by the fact that videos contain much more information than images, thus they are more memorable. For the VideoMem10k dataset, human consistency of 0.481 is observed for short-term memorability and of 0.192 for long-term memorability.

When splitting the number of participants into two groups, it is possible to have groups with unbalanced number of annotations per video. For this reason, we proposed a new metric named *annotation consistency*. We reproduced the previous process of human consistency computation but on videos which received at least N annotations and the split ensuring a balance number of annotations per video. By doing so, we obtained the annotation consistency as a function of the number of annotations per video, as presented in Fig. 3.6-middle and right for the VideoMem10k dataset. As can

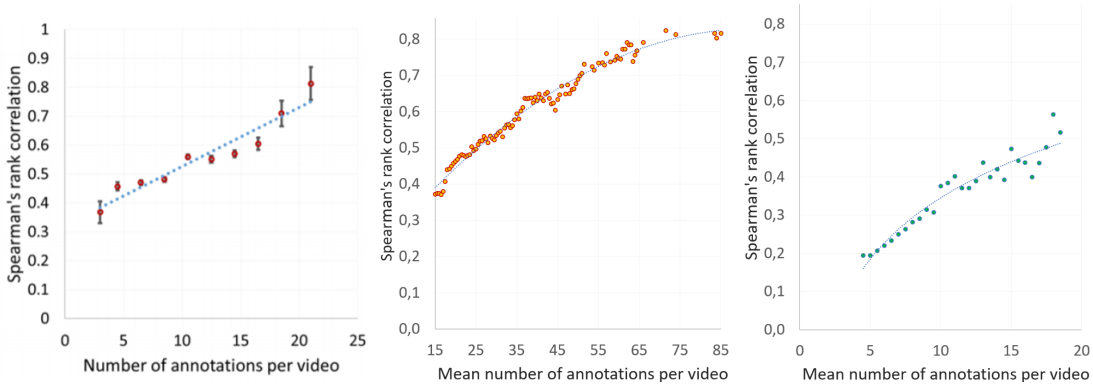


Figure 3.6: Consistency analysis [28, 29]. Left: human consistency in MovieMem660 dataset, middle: annotation consistency in VideoMem10k short-term VM, right: annotation consistency in VideoMem10k long-term VM. A greater number of annotations per video provides more reliable memorability scores.

be seen, the annotation consistency reaches 0.616 (respectively 0.364) for the short-term (resp. long-term) task for 38 (resp. 13) annotations. Again, this value of 0.616 for short-term memorability is to be compared to the one found in [KRTO15] (0.68) for images. We can also see that long-term consistency follows the same evolution as short-term consistency.

3.2.3 VM prediction

For the VM prediction with the VideoMem10k dataset, we investigated the use of various image based baselines (*i.e.,* MemNet [KRTO15], Squali *et al.,* [27], ResNet [HZRS16]), video-based models with spatio-temporal features (C3D [TBF+15], HMP [CGR14], ResNet3D [HKS17]), and the proposed semantic embedding-based model. For the latter, we fine-tuned a state-of-the-art visual semantic embedding pipeline used for image captioning [ECPC19], on top of which a 2-layer MLP is added, to regress the feature space to a single memorability score. The overall architecture is shown in Figure 3.7 and the training was done with a new ranking loss (*i.e.,* Spearman surrogate) proposed in [ECPC19]. Similar types of image-based and video-based features were investigated for the MovieMem660 dataset. Please refer to our papers [28, 29] for a detailed description of each considered systems and the discussion of the findings.

Table 3.1 shows the final prediction result in terms of the Spearman's rank correlation between predicted and groundtruth memorability scores on the validation and test sets, and on the 500 most annotated videos of the VideoMem10k dataset (test(500)). We also compare with the average and the best results obtained from the MediaEval

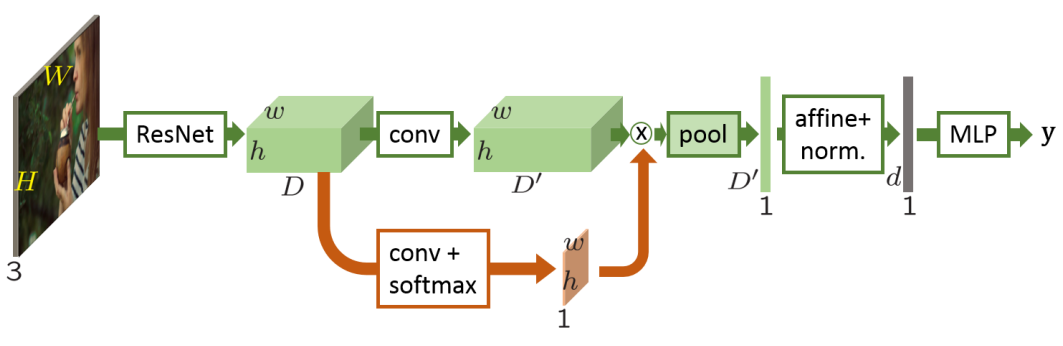


Figure 3.7: Semantic embedding model without (green pipeline) and with an attention mechanism (orange branch) (figure is from [29]).

2018 campaign for the same test set. As can be seen, baselines designed for IM prediction offer quite good results on VM prediction. This means that the memorability of a video is correlated to some extent with the memorability of its constituent frames. All the models show poorer performance at predicting long-term memorability compared with short-term VM prediction. This might be due to the fact that the memorability scores for long-term are based on a smaller number of annotations than for short-term, so they probably capture a smaller part of the intrinsic memorability. Finally, our proposed semantic embedding model outperforms all other systems for both short-term and long-term VM prediction.

Models	short-term memorability			long-term memorability		
	validation	test	test (500)	validation	test	test (500)
MemNet	0.397	0.385	0.426	0.195	0.168	0.213
Squalli *et al.*	0.401	0.398	0.424	0.201	0.182	0.232
IC-based model	0.495	0.441	0.517	0.233	0.204	0.199
ResNet101	0.498	0.46	0.527	0.222	0.218	0.219
C3D+HMP	0.424	0.337	0.412	0.324	0.121	0.120
ResNet3D	0.508	0.462	0.535	0.23	0.191	0.202
MediaEval'2018 - average	–	0.395	–	–	0.174	–
MediaEval'2018 - best	0.484	**0.497**	–	**0.261**	**0.257**	–
Semantic embedding model	**0.503**	0.494	**0.565**	0.260	0.256	**0.275**

Table 3.2: Results in terms of Spearman's rank correlation between predicted and ground-truth memorability scores, on the validation and test sets, and on the 500 most annotated videos of the dataset (test (500)) that were placed in the test set.

Intra-memorability visualization. To better understand what makes a video frame memorable, we visualize the 2D attention map obtained after the conv+softmax

layer (orange branch in Figure 3.7). This attention map is actually trained to learn which regions in each frame contribute more to the prediction. Some frame examples are shown in Figure 3.8. Our empirical study of the resulting attention maps tends to separate them in two categories. In the first one when frames contain roughly one main object, the model seems to focus on the main object and even, in the case of clear faces, on details of the faces, as if trying to remember the specific features of faces. In the second category that groups all other frames, with several main and secondary objects, textured background, *etc.,* it seems on the contrary that the model focuses on other little details that differentiate the image from another similar one. In other words, the second category shows results that might be interpreted as a second memorization process, once the first one – focusing on the main object – has already been achieved.

3.3 Conclusion

We have presented our pioneering works on media interestingness and memorability. Large-scale datasets about image/video interestingness and video memorability were constructed, analyzed, and released for the public use. Two series of MediaEval campaigns namely Media Interestingess Prediction Task (2016-2017), and Video Memorability Prediction Task (2018-2019) were organized and attracted great attention from the community. We also investigated various computational models for the media interestingness/memorability prediction and studied intrinsic factors that might make a content interesting or memorable.

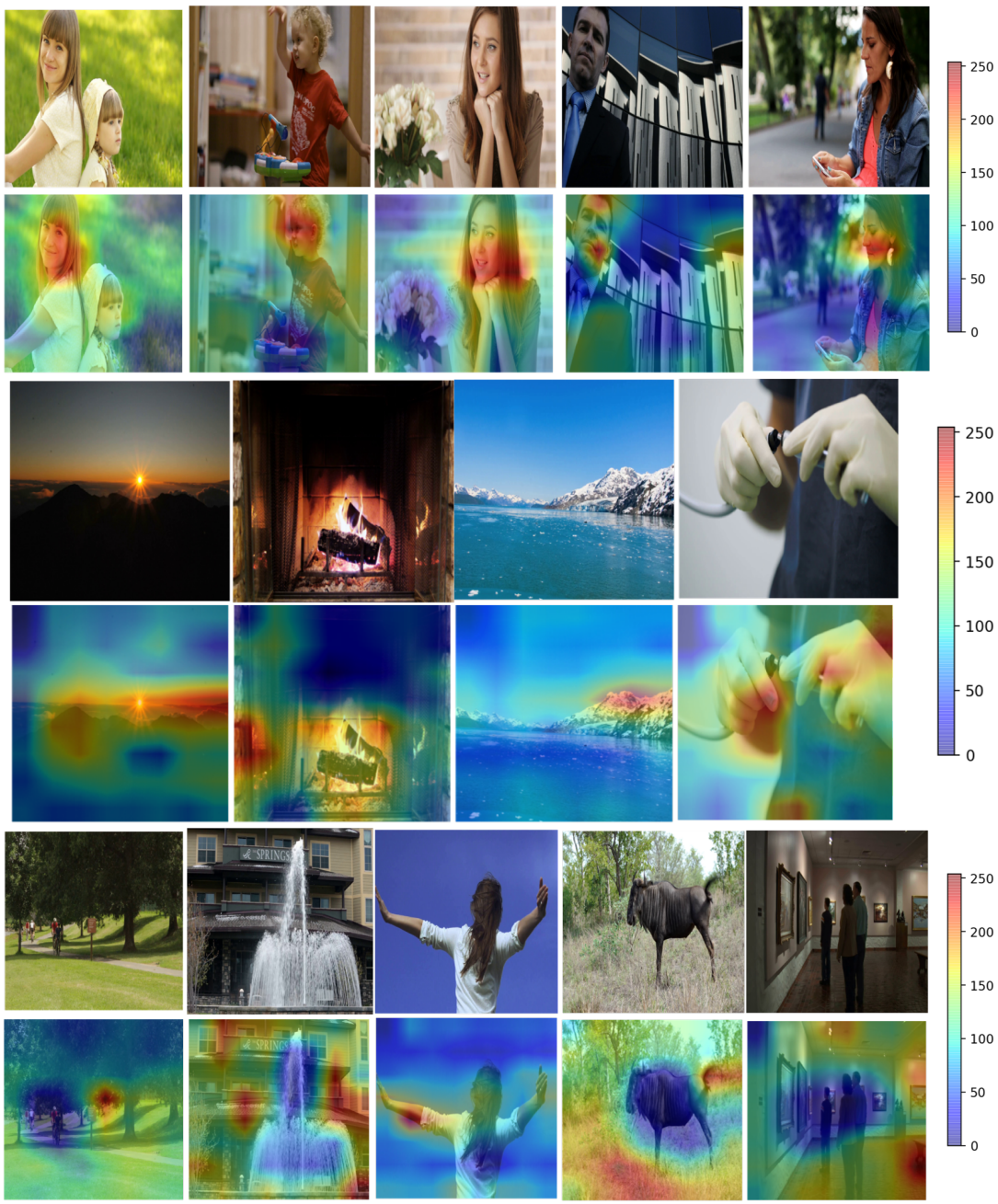


Figure 3.8: What makes a video/image memorable? Visualization of the attention mechanism's output for some video frames. The model focuses either on clear faces (first row) or main objects when background texture is dark or uniform or blurry (second row). In another type of images which contains several objects and textured background (third row), the model focuses rather on some details.

Chapter 4

Other contributions

Supervision: Pierre Prablanc (MSc intern), Gustavo Mafra (MSc intern), Eric Grinstein (MSc intern), Valentin Bilot (MSc intern), Ha Le (MSc intern), Van-Huy Vo (MSc intern)
Main collaborators: Alexey Ozerov (Technicolor, InterDigital), Patrick Pérez (Technicolor, Valeo.ai).

This chapter briefly summarizes our other contributions in applying signal processing and machine learning techniques to different problems.

4.1 Audio synchronization using fingerprinting

Consumers today often use their smartphones or tablets whilst watching TV. This has opened the door to personalized TV applications where additional services and related content can be accessed on the web to accompany the main TV view. Targeting such emerging applications, we looked for a technique to assure fast and accurate synchronization of media components streamed over different networks to different rendering devices. Focusing on audio processing, we considered fingerprinting techniques [Wan03] and generalized cross correlation [BS97], where the former can greatly reduce computational cost and the latter can offer sample-accurate synchronization. In [35], we proposed an approach combining these two techniques, where coarse frame-accurate synchronization positions were first found by fingerprint matching, and then a possible accurate synchronization position was verified by generalized cross correlation with phase transform (GCC-PHAT). Experimental results in a real-world setting confirmed the accuracy and the rapidity of the proposed approach.

In [13] we investigated another audio synchronization use case for movie: synchronizing multiple versions of the same movie, with an objective of automatically transferring metadata available on a reference version to other ones. We first adapted an existing audio fingerprinting technique [Wan03] to find all possible temporal matching positions between two audio tracks associated with two movie versions. We then proposed additional steps to refine the match and eliminate outliers. The proposed approach was shown to efficiently handle situations where temporal scene edits occur like scene addition, removal, and even the challenging scene re-ordering case. Experimental results over synthetic editorial data showed the effectiveness of the approach with respect to the state-of-the-art dynamic time warping (DTW) based solution.

4.2 Audio zoom via beamforming technique

This work focused on a practical application called audio zoom in smartphones, where sound capture focuses on the front direction while attenuating progressively surrounding sounds when recording a video. For this purpose, we first developed a novel approach that combines multiple Robust Minimum Variance Distortionless Response (RMVDR) beamformers [BS10, ML03] having different look directions with a post-processing algorithm. Then, spatial zooming effect is created by leveraging the microphone signals and the enhanced target source. The general workflow of the proposed audio zoom implementation is shown in Figure 4.1. Subjective test with real-world audio recordings using a mock-up simulating an usual shape of the smartphone confirms the rich user experience obtained by the proposed system [14]. A demo was presented at the ICASSP 2016 conference[1].

4.3 Audio classification

We have been interested in audio classification task for several years along with the emergence of deep learning. We participated in the 2016 Detection and Classification of Acoustic Scenes and Events (DCASE) challenges where we started from low-level feature representation for segmented audio frames and investigated different time granularity for feature aggregation. We studied the use of support vector machine (SVM) together with two popular neural network (NN) architectures, namely multi-layer perceptron (MLP) and convolutional neural network (CNN) and tested on benchmark datasets provided in the DCASE 2013 and 2016. We observed that a simple feature as averaged

[1] https://www2.securecms.com/ICASSP2016/ST-3.asp

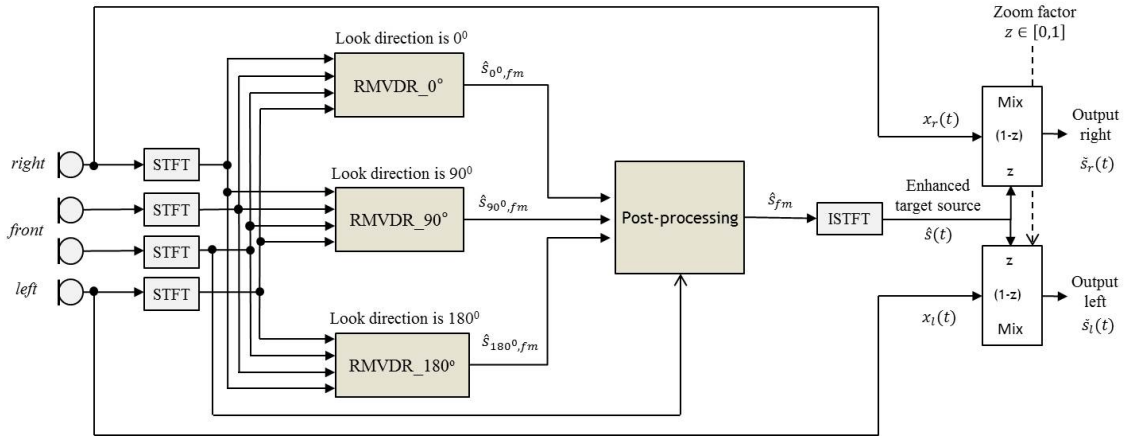


Figure 4.1: General workflow of the proposed audio zoom implementation (figure is from [14]).

Mel-log-spectrograms can obtain comparable performance with the best systems in the DCASE 2013 challenge [36].

Participating in the DCASE 2019 challenge with enlarged dataset, we used models exploiting multiple instance learning (MIL) method as a way of guiding the network attention to different temporal segments of a recording. We then proposed a simple late fusion of results obtained by the three investigated MIL-based models by multi-layer perceptron (MLP) layers. With such fusion, we obtained a better result on the development and the leaderboard dataset of the challenge [38].

In another work [37] we addressed a problem of discriminating the natural human voices and those played back by any kind of audio devices in the context of interactions with in-house voice user interface. The tackled problem finds relevant applications in (1) the far-field voice interactions of vocal interfaces such as Google Home, Facebook Portal, Amazon Echo, etc, and (2) the replay spoofing attack detection. The detection of loudspeaker emitted speech will help avoiding false wake-ups or unintended interactions with the devices in the first application, while eliminating attacks involve the replay of recordings collected from enrolled speakers in the second one. In this work, we first collected a real-world dataset under well-controlled conditions containing two classes: recorded speeches directly spoken by numerous people (considered as the natural speech), and recorded speeches played back from various loudspeakers (considered as the loudspeaker emitted speech). We then built DNN-based prediction models exploiting the combination of features extracted from different existing state-of-the-art DNN architectures. The combination of audio embeddings extracted from SoundNet

[YAT16] and VGGish [HCE+17a] network yields the classification accuracy up to about 90% and thus confirms the feasibility of the task.

4.4 Audio style transfer

Image style transfer has recently emerged with success and has become a very popular technology thanks to the power of convolution neural networks (CNNs). In this work we investigated the analogous problem in the audio domain: How to transfer the style of a reference audio signal to a target audio content? To the best of our knowledge, our paper [15] was one of the earliest formal publications in this topic. We proposed a flexible framework for the task, which uses a sound texture model to extract statistics characterizing the reference audio style, followed by an optimization-based audio texture synthesis to modify the target content. In contrast to mainstream optimization-based visual transfer method, the proposed process is initialized by the target content instead of random noise, and the optimized loss is only about texture, not structure. In order to extract features of interest, we investigated different architectures, whether pre-trained on other tasks, as done in image style transfer, or engineered based on the human auditory system. The overall framework is shown in Figure 4.2. In this figure, given an audio texture extraction model (artificial neural net or auditory model), the content sound is iteratively modified such that its audio texture matches well the one of the style sound. If required by texture model, raw signals are mapped to and from a suitable representation space by pre/post-processing. Experimental results on different types of audio signal confirm the potential of the proposed approach/ transfer in our experiments.

4.5 Speech inpainting

Audio inpainting in general consists in filling in missing portions of an audio signal. It exists in different forms such as audio declipping, clicks removal, and bandwidth extension. The problem of speech inpainting specifically consists in recovering some parts in a speech signal that are missing for some reasons. To our best knowledge none of the existing methods allows satisfactory inpainting of missing parts of large size such as one second and longer. In this work we addressed this challenging scenario with the assumption that the full text uttered in the speech signal is known (as in the case of such long missing parts entire words can be lost). We thus formulated a new concept of text-informed speech inpainting and proposed a method that is based on synthesizing the

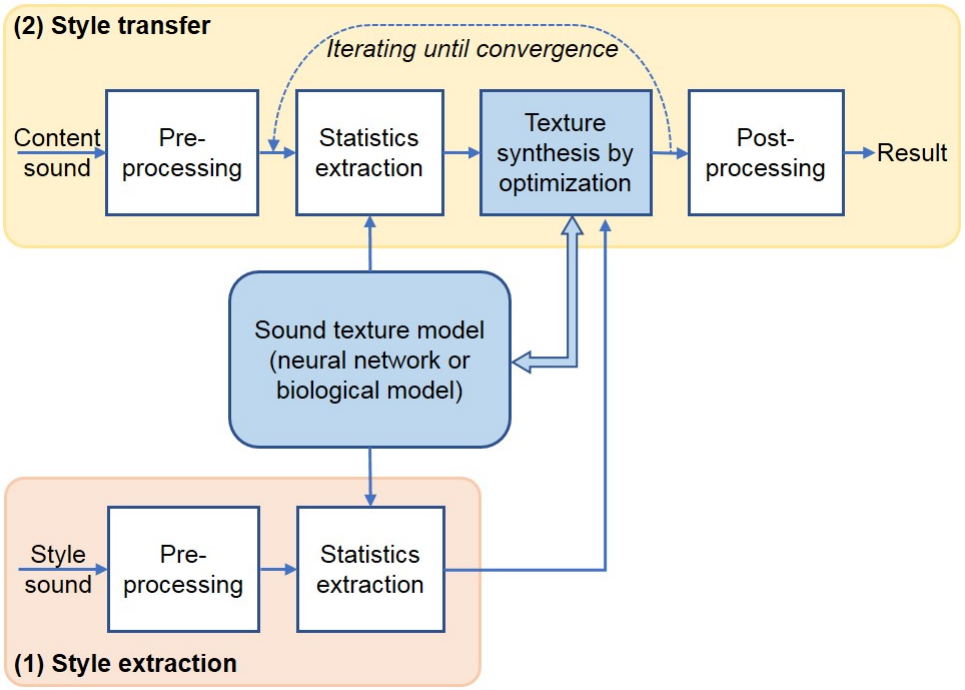


Figure 4.2: Proposed audio style transfer framework (figure is from [15]).

missing speech by a speech synthesizer, on modifying its vocal characteristics via a voice conversion method, and on filling in the missing part with the resulting converted speech sample [16]. We carried subjective listening tests to compare the proposed approach with two baseline methods.

4.6 Image inpainting

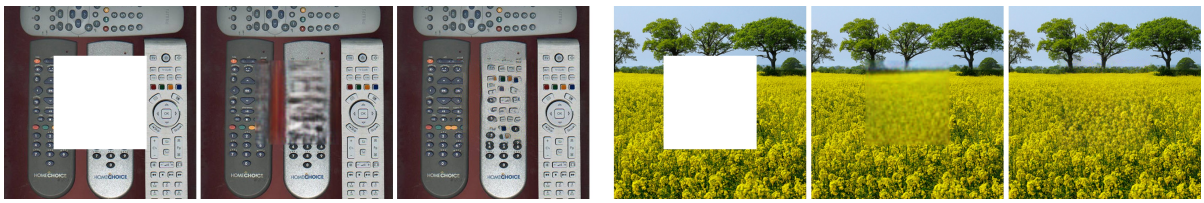


Figure 4.3: Given an incomplete scene (left image), the proposed context encoder produces a plausible structural completion (middle image), which can be subsequently refined for texture and details with a patch-based inpainting method (right image) (figure is from [17]).

Visual inpainting or scene completion, is the task of filling in a plausible way a

region in an image [BSCB00]. Successful approaches for small and medium missing regions include patch-based inpainting [CPT04, BDTL15] and iterative optimization-based approaches [AFCS11, BSFG09]. Recently, with the power of CNNs, Pathak *et al.* [PKD+16] have introduced convolutional "context encoders" (CEs) for unsupervised feature learning through image completion tasks. With the additional help of adversarial training, CEs turned out to be a promising tool to complete complex structures in real inpainting problems. In this work we proposed to push further this approach by relying on perceptual reconstruction losses at training time [17]. The overall workflow of the proposed structural CE is shown in Figure 4.4. We showed on various visual scenes the merit of the approach for structural inpainting, and confirmed it through a user study. Combined with the optimization-based refinement of [YLL+16] with neural patches, our context encoder opens up new opportunities for prior-free visual inpainting. Example of the qualitative results are shown in Figure 4.3.

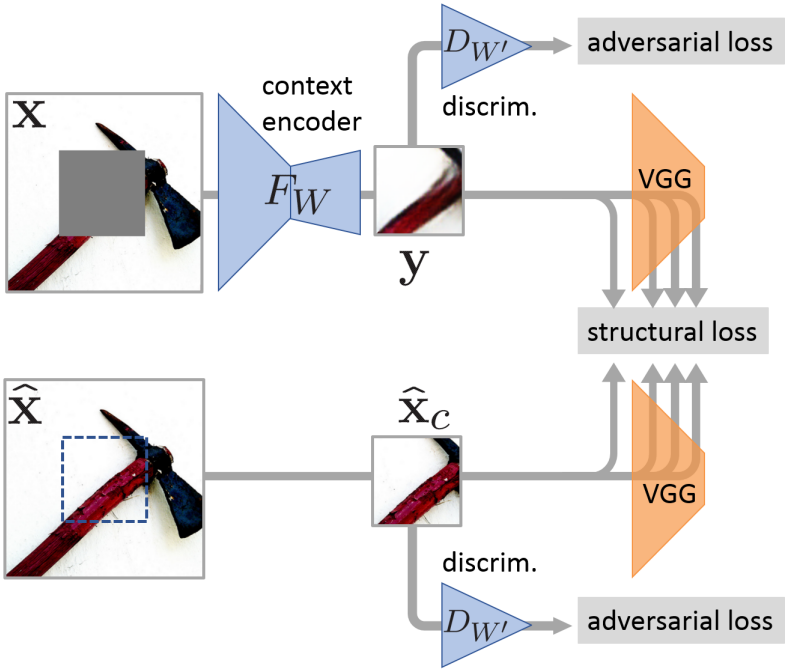


Figure 4.4: Proposed structural CE: The encoder-decoder architecture is trained with a structural loss that compares the reconstructed central image part with the original one through deep features. In a second training stage, the adversarial loss is added to the total loss, with a co-trained network in charge of declaring whether an input image patch is natural or produced by the competing CE. Learnable nets are in blue, orange ones are fixed (figure is from [17]).

Chapter 5

Conclusion

5.1 Achievements

This thesis summarizes my work in collaboration with other colleagues and students after the PhD. Particularly, we solved various problems ranging from different types of data (*i.e.,* audio, speech, text, image, and video). These works mostly exploited statistical signal processing and machine learning techniques for audio and multimodal data modeling and analysis. Concerning audio source separation, while my PhD focused on spatial source model (*i.e.,* the full-rank spatial covariance model), we mostly investigated the use of spectral source model based on NMF with sparsity constraints for a wide range of *informed* use cases. Such use cases cover supervised/semi-supervised, text-informed, user-guided, annotation-based, interactive, motion-informed, and video-informed audio source separation.

In recent years, data-driven paradigm has bloomed and found wide applications in most domains thanks to the deep learning. In line with this, we investigated the use of deep neural networks for various audio and visual tasks, which were found to be difficult for conventional model-based approaches before. These tasks include *e.g.,* audio style transfer, large-hole image inpainting, audio classification, audio-visual object detection and localization, image/video interestingness, and image/video memorability. For the media interestingness and memorability, we have led several activities in the research community such as constructing new datasets, organizing international challenges, and proposing computational models for prediction.

As a researcher in industry, my research projects have been driven by the actual needs of the company. Alongside the supervision of two PhD students and one postdoc, other projects usually aimed to investigate practical challenges over a few months. Thus,

I have had great opportunities to study various problems and extend my research, originally in audio processing, to areas in multimedia and computer vision with which I was not familiar. For each target problem, we were mostly successful to have at least one patent submission and one scientific paper.

5.2 Future directions

My research program aims to understand the real-world environment via multimodal data (*e.g.,* audio-visual scene) analysis. From this understanding, smart devices will be able to respond to users' needs. A specific direction could target multimodal question answering system (MQA), which finds great application in intelligent assistants such as Amazon Alexa, Google Home, Apple Siri, etc. While new frontiers in visual question answering (VQA) and multimodal fusion have been achieved recently[1], most state-of-the-art approaches treat VQA as a conventional classification problem where the list of possible answers is fixed in the dataset [GKSS+17, KMK+19]. This makes VQA systems less flexible in responding to the real needs of the users. Besides, multimodal sensors have been well investigated in different tasks such as activity recognition, but less investigated in the context of VQA. Thus, with a future PhD student, we would like to build a novel semantic multimodal question answering (MQA) framework that exploits the latest advances in deep learning research and multimodal fusion. This framework may go beyond the conventional classification approach thanks to the integration of reasoning techniques [YSY+17, ZBFC18].

We are now in the deep learning era where powerful DNN models for various complicated tasks are being published almost every month, or even week. However, most DNN architectures are fixed during the design, training, and inference stages so they are not easily adapted to the possible variation of the hardware and/or computational resources. Thus, another research direction I am exploring focuses on *flexible* DNN architectures, which offer memory efficiency and can be easily adapted to the available resources on the fly. Such flexible models (*e.g.,* motivated from the multi-scale dense networks [HCL+18] or slimmable networks [YH]) can be configured to fit to *e.g.,* edge computing or IoT devices with limited computing power when needed. The use of in-place knowledge distillation (IPKD) motivated from the teacher assistant strategy [MFL+19] to guide the efficient training of such flexible models is currently under investigation. Yet another interesting line of research is the unsupervised or weakly-supervised learning, where robust data-driven machine learning models can be built

[1] https://visualqa.org/workshop.html

without or with a small amount of labeled data.

Appendices

Appendix A

Paper 1: On-the-Fly Audio Source Separation—A Novel User-Friendly Framework

On-the-fly audio source separation - a novel user-friendly framework

Dalia El Badawy, Ngoc Q. K. Duong, *Member, IEEE* and Alexey Ozerov, *Member, IEEE*

***Abstract*—This article addresses the challenging problem of single-channel audio source separation. We introduce a novel user-guided framework where source models that govern the separation process are learned *on-the-fly* from audio examples retrieved online. The user only provides the search keywords that describe the sources in the mixture. In this framework, the generic spectral characteristics of each source are modeled by a *universal sound class model* learned from the retrieved examples via non-negative matrix factorization. We propose several group sparsity-inducing constraints in order to efficiently exploit a relevant subset of the universal model adapted to the mixture to be separated. We then derive the corresponding multiplicative update rules for parameter estimation. Separation results obtained from automated and user tests on mixtures containing various types of sounds confirm the effectiveness of the proposed framework.**

***Index Terms*—On-the-fly audio source separation, user-guided, non-negative matrix factorization, group sparsity, universal sound class model.**

I. Introduction

AUDIO source separation is a desired processing step within many real-world applications such as sound post-production, robotics, and audio enhancement [1]. However, it has remained a challenging task especially when the input is a single-channel mixture. Indeed, in this case the problem is highly ill-posed and, in contrast to multichannel mixing case, additional spatial information about the sources is not available. Earlier approaches usually assume that the sources are sparse in the short-time Fourier transform (STFT) domain and estimate the predominant source's STFT coefficients via *e.g.* binary masking [2] or ℓ_1-minimization [3], [4]. The separation performance achievable by these techniques is very limited in reverberant environments [5], [6] where the sources' STFT coefficients are quite overlapped. A more recent class of algorithms known as *informed* source separation [7], [8] utilizes prior information about the sources to guide the separation process, and was shown to be successful in many contexts using different types of prior information. For instance, such information may include musical scores of the corresponding music sources [7], [9], [10] or text of the corresponding speech sources [8]. In some approaches this symbolic information is then converted to audio using a MIDI synthesizer for musical scores [9], [10] or a speech synthesizer for text [8]. These synthesized signals (that may also include cover tracks as in [11]) called *deformed references* in [12] can be used to roughly learn the spectral and temporal characteristics of one or more sources in the mixtures so as to guide the separation process [8]–[10], [12]. A subclass of informed source separation approaches is *user-guided* separation methods where the prior information is provided by a user. Such information can be *e.g.*, user-"hummed" sounds that mimic the sources in the mixture [13] or source activity annotation along time [14] or in a time-frequency plane [15]; the annotation information is then used, instead of training data, to guide the separation process. Furthermore, recent publications disclose an interactive strategy [16], [17] where the user can perform annotations on the spectrogram of intermediate separation results to gradually correct the remaining errors. Note however that most of the existing approaches need to use prior information which may not be easy to acquire in advance (*e.g.*, musical score, text transcript), is difficult to produce (*e.g.*, user-hummed examples), or simply requires very experienced users while being very time consuming (*e.g.*, time-frequency annotations).

The main motivation of this work is to introduce a simple framework that enables everyone to easily perform source separation. We hence present the new concept of *on-the-fly* source separation inspired by on-the-fly visual search methods [18], [19] from the computer vision research community. More specifically, the proposed framework only requires the user to listen to the mixture and type some keywords that describe the sources to be separated; in other words, the user interaction is now carried out at a higher semantic level. For instance, a user would request to separate the "wind noise" (source 1 description) from the "bird song" (source 2 description). The given descriptions or keywords are then used to search the internet for similar audio examples that will be employed to govern the separation process. For this purpose, supervised approaches based on *e.g.*, nonnegative matrix factorization (NMF) [20], [21] or its probabilistic formulation known as Probabilistic Latent Component Analysis (PLCA) [13], [22], where retrieved examples can be used to pre-learn the spectral dictionaries of the corresponding sources, are of great interest. Other methods in the prior art that couple the decomposition of the reference signals together with the mixture could also be considered [8], [11], [12], [23]. Regardless of the approach, several challenges, as detailed in Section II, arise in this on-the-fly framework due to (i) the unknown quality of the retrieved examples and (ii) possible lack of source descriptions (*i.e.* some sources may not be described by the user). In our preliminary work [24], we investigated several strategies to handle issues with the quality of the retrieved

D. El Badawy is a student at EPFL, Switzerland, e-mail: (dalia.elbadawy@epfl.ch).

N. Q. K. Duong and A. Ozerov are with Technicolor R&I, France, e-mail: (quang-khanh-ngoc.duong@technicolor.com; alexey.ozerov@technicolor.com).

This work was done while the first author was an intern at Technicolor.

Manuscript received Month xx, 2016; revised Month xx, 2016.

examples and found that the one using a *universal sound class model (USCM)*[1] learned from examples via NMF with a *group sparsity* constraint is generally more efficient than the others. Note that since the USCM is actually an over-complete dictionary, a sparsity constraint is needed to help fit the most relevant spectral patterns to the sources in the mixture.

This article extends our preliminary work [24], [27] by providing the algorithms along with their mathematical derivations in addition to new results from a user test. Altogether, the main contributions of our proposed on-the-fly paradigm work are four-fold:

- We introduce a general framework for on-the-fly audio source separation which greatly simplifies the user interaction.
- We propose a novel so called *relative group sparsity* constraint and show its benefit in the semi-supervised case where training examples for some sources are missing.
- We derive several algorithms for parameter estimation when different group sparsity-inducing penalties and relative group sparsity-inducing penalties are used.
- We perform a range of evaluations, including both supervised and semi-supervised scenarios, and a user-test to validate the benefit of the proposed framework.

The remainder of this paper is organized as follows. Section II gives an overview of the on-the-fly framework and the related challenges. In Section III, we recall some background on supervised source separation based on NMF. We then propose several algorithms for parameter estimation with the use of different sparsity-inducing constraints in Section IV. Evaluation results with a user-test are presented in Section V. Finally, we conclude in Section VI.

II. On-the-fly framework and challenges

A. Overview and challenges

The proposed framework only requires minimal user input enabling inexperienced users to apply source separation to essentially any mixture.It is applicable as well when relevant training examples for some sources are either not readily available offline or not representative enough, which is likely the case for uncommon sounds such as animal or environmental sounds. The general workflow is shown in Fig. 1. The user inputs a few keywords specifying the sources in the mixture (*e.g.*, "dog barking", "wind", etc.,), then a search engine retrieves relevant source examples accordingly. The source spectral models are then learned on-the-fly and used for supervising the separation. This approach is actually analogous to on-the-fly methods in visual search where a user types a persons name (e.g., "Clint Eastwood") [18] or an objects description (e.g., "car") [19] and a classifier is trained using example images retrieved via Google Image Search.

Although the on-the-fly approach simplifies the user interaction and eliminates the need for offline training samples, there are several challenges that need to be addressed as follows:

- (C_1) *Handling irrelevant examples*: Some retrieved examples may contain sounds with entirely different spectral characteristics than those of the source in the mixture, *e.g.*, searching for "bird chirps" and obtaining some "chirp signal" examples too. Those examples should not be used in training.
- (C_2) *Handling noisy examples*: Some retrieved examples are actually mixtures of relevant and irrelevant sounds, *e.g.*, "female speech" with a dog barking in the background. Those examples may still be useful and should not be discarded entirely.
- (C_3) *Handling missing examples*: This may happen when the user describes only the sources of interest and ignores the remaining sources or when the search engines do not return results for some of the provided keywords. We refer to this challenge as the semi-supervised case where all non-described sources that possibly appear in the mixture should be grouped as one background source.

In fact in our previous work [24] to handle the first challenge, we investigated the use of a simple example preselection scheme based on the spectral similarity between the examples and the mixture to discard irrelevant examples. Thus, one can imagine having additional user interaction after specifying the keywords. For instance, the user may screen the list of retrieved examples and subjectively select a more relevant subset for training. This is the "Examples Refinement" step in Fig. 1.

B. Graphical User Interface

We implemented the system along with a graphical user interface (GUI) as shown in Fig. 2 and employed it for our user-tests. It features the ability to listen to a mixture and input one or more keywords describing the different sources. Then, per source, an online search for audio is performed. Next, the user can listen to the list of retrieved audio examples as well as view their waveforms or spectrograms (useful for the more advanced users). The optional example selection is then done by ticking the corresponding checkboxes. USCMs are then learned on-the-fly to guide the separation. The last step is to output the separated sources. A video showing a demo is available online at http://youtu.be/mBmJW7cy710/. On the practical side, the data transferred between the user and the server consists of the keywords and the mixture file as well as the separated sources which are sent back to the user. On the server, each example file requires computing the STFT followed by NMF; the examples are independent and these operations can thus be done in parallel. Then once the USCMs have all been constructed, the separation step is faster as the multiplicative updates are performed only one time. The overall complexity thus mostly depends on the number of training examples and the size of USCMs. Thus on an average PC, it would take from 30 seconds to a few minutes to get the separation results back.

[1]The term "universal speech model" was introduced in [25] for the separation of speech and noise, and was inspired by the term "universal background model" used for speaker verification [26]. We here extend it to "universal sound class model", since our framework deals with the separation of sources belonging to any sound class.

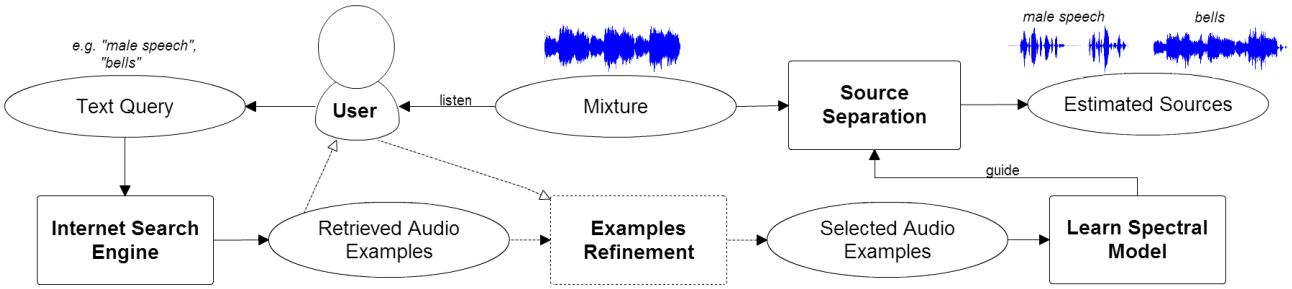
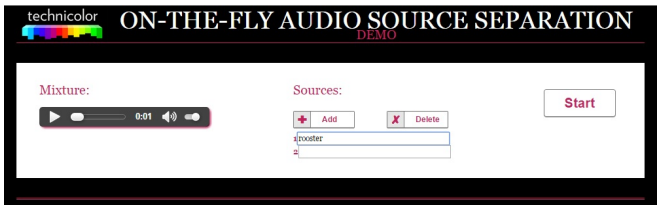
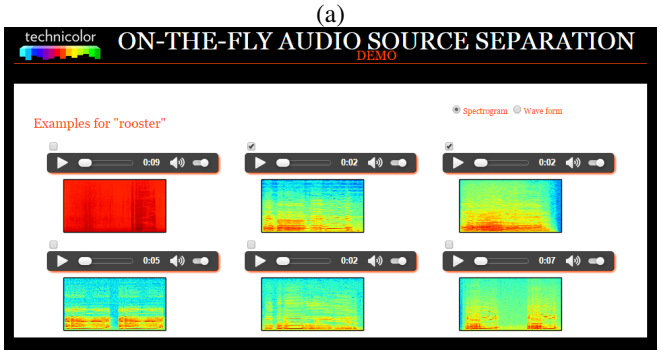


Fig. 1. General workflow of the proposed on-the-fly framework. A user listens to the mixture and types some keywords describing the sources. These keywords are then used to retrieve examples to learn spectral models for the described sources. Optionally, the user may listen to the retrieved examples and discard irrelevant ones.



(a)



(b)

Fig. 2. Screenshots of the proposed GUI. (a) The user can listen to a mixture, and then type keywords describing different sources within it. (b) A set of retrieved examples (waveforms or spectrograms can be displayed) for each source with checkboxes so that the user can select the most appropriate ones to be used for training the source spectral model.

III. Background on NMF-based Source Separation

A. Conventional supervised approach

We discuss in this section a standard supervised source separation approach. We base our framework on NMF since it is one of the most popular and well-suited models in the state of the art on audio source separation. As per *e.g.* [22], [25], first source spectral models are learned on-the-fly from training data retrieved online. Then these models are used to supervise the separation.

Assuming J sources, let $\mathbf{X} \in \mathbb{C}^{F\times N}$ and $\mathbf{S}_j \in \mathbb{C}^{F\times N}$ be the STFT coefficients of the single channel mixture signal and the j-th source signal, respectively, where F is the number of frequency bins and N the number of time frames. Usual additive mixing is assumed as

$$\mathbf{X} = \sum_{j=1}^{J} \mathbf{S}_j. \tag{1}$$

Let $\mathbf{V} = |\mathbf{X}|^{\cdot 2}$ be the power spectrogram of the mixture, where $\mathbf{X}^{\cdot p}$ is the matrix with entries $[\mathbf{X}]_{il}^{p}$, $^{\cdot p}$ denotes an element-wise operation. Then, NMF algorithms construct two non-negative matrices $\mathbf{W} \in \mathbb{R}^{F\times K}$ and $\mathbf{H} \in \mathbb{R}^{K\times N}$ such that $\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH}$. The factorization is usually done by solving the following optimization problem [20], [28]

$$\mathbf{W}^*, \mathbf{H}^* = \underset{\mathbf{H}\geq 0, \mathbf{W}\geq 0}{\arg\min}\ D(\mathbf{V}\|\mathbf{WH}), \tag{2}$$

where

$$D(\mathbf{V}\|\hat{\mathbf{V}}) = \sum_{f,n=1}^{F,N} d(\mathbf{V}_{fn}\|\hat{\mathbf{V}}_{fn}) \tag{3}$$

and $d(\cdot\|\cdot)$ is a scalar divergence measure. We use the Itakura-Saito (IS) divergence defined as

$$d_{IS}(x\|y) = \frac{x}{y} - \log\left(\frac{x}{y}\right) - 1 \tag{4}$$

which is appropriate for audio signals due to its scale invariance [20]. The parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ are usually initialized with random non-negative values and are iteratively updated via multiplicative update (MU) rules [20], [28].

In the training step of the supervised setting, a spectral model for each source j, denoted by $\mathbf{W}_{(j)}$, is first learned from the corresponding training examples concatenated together by optimizing criterion (2). Then the spectral model for all sources $\mathbf{W}$ is obtained by concatenating the source models as:

$$\mathbf{W} = [\mathbf{W}_{(1)}, \ldots, \mathbf{W}_{(J)}]. \tag{5}$$

Then in the separation step, the time activation matrix $\mathbf{H}$ is estimated via the MU rules optimizing (2) [20], while $\mathbf{W}$ is kept fixed. Note that the activation matrix is also partitioned into horizontal blocks as

$$\mathbf{H} = [\mathbf{H}_{(1)}^T, \ldots, \mathbf{H}_{(J)}^T]^T, \tag{6}$$

where $\mathbf{H}_{(j)}$ denotes the block characterizing the time activations for the j-th source.

Once the parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ are obtained, Wiener filtering is applied to compute the source STFT coefficients as

$$\hat{\mathbf{S}}_j = \frac{\mathbf{W}_{(j)}\mathbf{H}_{(j)}}{\mathbf{WH}} \odot \mathbf{X}, \tag{7}$$

where $\odot$ denotes the element-wise Hadamard product and the division is also element-wise. Finally, the inverse STFT is computed to produce the time domain source estimates.

B. USCM-based approach

The conventional supervised approach as described in Section III-A assumes using all retrieved (or user-selected) examples for a given source to learn the source spectral model. This may not be suitable in the current framework due to the challenges mentioned in Section II where the noisy examples may lead to a poor spectral model. Thus, in this section we propose an efficient and flexible approach to better utilize the examples, when available, for guiding the separation, while also handling the case of missing examples. In the following, the training examples refer to either the full list of retrieved examples or the user-selected examples in case of user intervention. We employ a so-called *universal sound class model*, learned in advance from training examples, with sparsity constraints on the activation matrix $\mathbf{H}$ in order to enforce the selection of only few representative spectral patterns during the model fitting. In the following, we first present the USCM construction, and then the optimization criterion for model fitting.

1) USCM construction: Assuming that the j-th source is described by the user and some examples are retrieved for it, we denote by $\mathbf{V}_{jp}$ the spectrogram of the p-th example corresponding to the j-th source. First, $\mathbf{V}_{jp}$ is used to learn the NMF spectral model, denoted by $\mathbf{W}_{jp}$, by optimizing the criterion (similar to (2)):

$$\mathbf{H}^*_{jp}, \mathbf{W}^*_{jp} = \underset{\mathbf{H}_{jp}\geq 0, \mathbf{W}_{jp}\geq 0}{\arg\min}\ D(\mathbf{V}_{jp} \| \mathbf{W}_{jp}\mathbf{H}_{jp}), \tag{8}$$

where $\mathbf{H}_{jp}$ is the corresponding time activation matrix. Given $\mathbf{W}_{jp}$ for all examples, the USCM for the j-th source is constructed as

$$\mathbf{W}_{(j)} = [\mathbf{W}_{j1}, \ldots, \mathbf{W}_{jP_j}] \tag{9}$$

where P_j is the number of retrieved examples for the j-th source.

2) Model fitting for supervised source separation: In the supervised setting, we assume having source models for all the sources in the mixture, that is to say that for every source, the user gave its description and examples were successfully retrieved. It can be seen that the USCM $\mathbf{W}_{(j)}$ constructed in (9) becomes a large matrix when the number of examples increases, and it is often redundant since different examples may share similar spectral patterns. Therefore, in the NMF decomposition of the mixture, the need for a sparsity constraint arises to fit only a subset of each $\mathbf{W}_{(j)}$ to the source in the mixture. In other words, the mixture is decomposed in a supervised manner, given $\mathbf{W}$ constructed from $\mathbf{W}_{(j)}$ as in (5) and fixed, by solving the following optimization problem

$$\mathbf{H}^* = \underset{\mathbf{H}\geq 0}{\arg\min}\ D(\mathbf{V} \| \mathbf{WH}) + \Psi(\mathbf{H}) \tag{10}$$

where $\Psi(\mathbf{H})$ denotes a penalty function imposing sparsity on the activation matrix $\mathbf{H}$. Different penalties can be chosen, as will be discussed in Section IV, resulting in a sparse matrix $\mathbf{H}$ as visualized in Fig. 3b and Fig. 3c.

3) Model fitting for semi-supervised source separation: We describe in this section a so-called semi-supervised setting where not all of the source models can be learned in advance [25]. This occurs either when the user only describes the sources of interest and not all of them or when the search engine fails to retrieve examples for a given query.

We propose to model all the "missing" sources as one background source whose spectrogram can be approximately factorized as $\mathbf{W}_\mathrm{b}\mathbf{H}_\mathrm{b}$, where $\mathbf{W}_\mathrm{b}$ and $\mathbf{H}_\mathrm{b}$ are the corresponding spectral model and activation matrix, respectively. The parameter $\boldsymbol{\theta}_\mathrm{b} = \{\mathbf{W}_\mathrm{b}, \mathbf{H}_\mathrm{b}\}$ can be randomly initialized with a small number of components (*i.e.* number of columns in $\mathbf{W}_\mathrm{b}$) K_b. All the other sources, for which some examples are available, are modeled as in the supervised case by $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ (see Fig. 4e and Fig. 4f). The parameters are estimated altogether by optimizing the following criterion

$$\mathbf{H}^*, \mathbf{W}^*_\mathrm{b}, \mathbf{H}^*_\mathrm{b} = \underset{\mathbf{H}\geq 0, \mathbf{W}_\mathrm{b}\geq 0, \mathbf{H}_\mathrm{b}\geq 0}{\arg\min}\ D(\mathbf{V} \| \mathbf{WH} + \mathbf{W}_\mathrm{b}\mathbf{H}_\mathrm{b}) + \Psi(\mathbf{H}). \tag{11}$$

We see that in contrast to criterion (10) $\mathbf{W}_\mathrm{b}$ is updated as well and there is no group sparsity-inducing penalty on $\mathbf{H}_\mathrm{b}$. The reason is that, as opposed to $\mathbf{W}$, $\mathbf{W}_\mathrm{b}$ is neither an overcomplete dictionary nor has an underlying structure that can be exploited for regularization.

IV. Sparsity constraints and algorithms

In this section we consider two classes of sparsity constraints, namely *group sparsity* and a newly proposed *relative group sparsity* for the optimization problem (10) and (11). In each case, two variations are considered: a *block* sparsity-inducing penalty and a *component* sparsity-inducing penalty. For every constraint, we give the corresponding algorithm for estimating the parameters.

A. Group sparsity constraints and parameter estimation algorithm

We consider a group sparsity-inducing penalty defined as

$$\Psi_\mathrm{gr}(\mathbf{H}) = \sum\nolimits_{j=1}^{J} \lambda_j \sum\nolimits_{g=1}^{G_j} \log\left(\epsilon + \|\mathbf{H}_{(j,g)}\|_1\right), \tag{12}$$

where $\mathbf{H}_{(j,g)}$ ($g = 1, \ldots, G_j$) are the groups within the activation sub-matrix $\mathbf{H}_{(j)}$ corresponding to the j-th USCM (see equation (6) for the definition of $\mathbf{H}_{(j)}$), G_j the total number of groups for the j-th source, $\|\cdot\|_1$ denotes the ℓ_1 matrix norm, $\epsilon > 0$ and $\lambda_j \geq 0$ are trade-off parameters determining the contribution of the penalty for each source. Note that in the remainder of the paper, $\mathbf{H}_{(j,g)}$ should not be confused with $\mathbf{H}_{jp}$ in (8). We introduce two options for defining the groups $\mathbf{H}_{(j,g)}$ and derive the corresponding MU rules for the parameter estimation as follows.

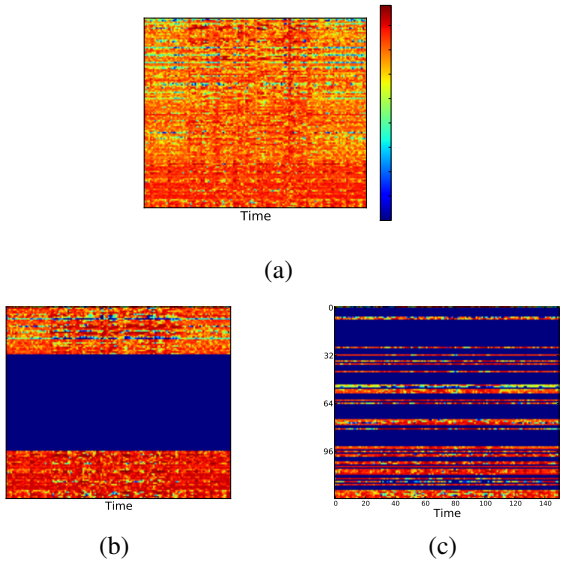


Fig. 3. Estimated activation matrix $\mathbf{H}$ for two sources in a mixture containing a rooster and bird chirps where two retrieved examples for each source were used for training the USCMs: (a) without a sparsity constraint, (b) with a block sparsity-inducing penalty (blocks corresponding to poorly fitting models are zero), and (c) with a component sparsity-inducing penalty (rows corresponding to poorly fitting spectral components from different models are zero).

1) Block sparsity-inducing penalty: As in [25], we restrict the groups to be sub-matrices of $\mathbf{H}_{(j)}$ corresponding to the spectral models $\mathbf{W}_{(j,p)}$ trained using the p-th example (see (8) for the definition of $\mathbf{W}_{(j,p)}$). In that case the indices g and p coincide and $G_j = P_j$. This so-called *block sparsity*-inducing strategy allows filtering out irrelevant spectral models $\mathbf{W}_{(j,l)}$, thus dealing with irrelevant retrieved examples (challenge C_1 in Section II). An illustration for the estimated activation matrix $\mathbf{H}$ for that case is shown in Fig. 3b where blocks corresponding to irrelevant examples for each source are set to zero.

2) Component sparsity-inducing penalty: As an alternative solution to fitting the universal model, we restrict the groups to be rows of $\mathbf{H}_{(j)}$ corresponding to different spectral components (in that case the number of groups G_j simply equals to the number of rows in $\mathbf{H}_{(j)}$). This so-called *component sparsity*-inducing strategy allows filtering out irrelevant spectral components, thus dealing with noisy retrieved examples (challenge C_2 in Section II). Fig. 3c shows an estimated activation matrix $\mathbf{H}$ where rows corresponding to irrelevant spectral components for each source are set to zero.

3) MU rules for parameter estimation: MU rules for the optimization of criterion (10) (supervised case) and (11) (semi-supervised case) are summarized in Algorithms 1 and 2, respectively, where $\eta > 0$ is a constant parameter, $\mathbf{P}_{(j,g)}$ is a matrix of the same size as $\mathbf{H}_{(j,g)}$ whose entries have the same value, and $\mathbf{P}$ is a matrix concatenating all $\mathbf{P}_{(j,g)}$. This algorithm is almost identical to the one proposed in [21], except that the groups are defined differently and $\mathbf{W}$ is not updated. It is proven in [21] using a majorization-minimization [29] formulation that these updates with $\eta = 1/2$ are *monotonic*, i.e., the cost function is non-increasing after each iteration.

Algorithm 1 MU rules for NMF with group sparsity in supervised case

Input: $\mathbf{V}$, $\mathbf{W}$, $\boldsymbol{\lambda}$
Output: $\mathbf{H}$
  Initialize $\mathbf{H}$ randomly
  $\hat{\mathbf{V}} = \mathbf{W}\mathbf{H}$
  repeat
    for $j = 1, \ldots, J, g = 1, \ldots, G_j$ **do**
      $\mathbf{P}_{(j,g)} \leftarrow \frac{\lambda_j}{\epsilon + \|\mathbf{H}_{(j,g)}\|_1}$
    end for
    $\mathbf{P} = [\mathbf{P}_{(1,1)}^T, \ldots, \mathbf{P}_{(1,G_1)}^T, \ldots, \mathbf{P}_{(J,1)}^T, \ldots, \mathbf{P}_{(J,G_J)}^T]^T$
    $\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^T(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2})}{\mathbf{W}^T \hat{\mathbf{V}}^{\cdot -1} + \mathbf{P}} \right)^{\cdot \eta}$
    $\hat{\mathbf{V}} \leftarrow \mathbf{W}\mathbf{H}$
  until convergence

Algorithm 2 MU rules for NMF with group sparsity in semi-supervised case

Input: $\mathbf{V}$, $\mathbf{W}$, $\boldsymbol{\lambda}$
Output: $\mathbf{H}, \mathbf{H}_b, \mathbf{W}_b$
  Initialize $\mathbf{H}$, $\mathbf{H}_b$, and $\mathbf{W}_b$ randomly
  $\hat{\mathbf{V}} = \mathbf{W}\mathbf{H} + \mathbf{W}_b\mathbf{H}_b$
  repeat
    for $j = 1, \ldots, J, g = 1, \ldots, G_j$ **do**
      $\mathbf{P}_{(j,g)} \leftarrow \frac{\lambda_j}{\epsilon + \|\mathbf{H}_{(j,g)}\|_1}$
    end for
    $\mathbf{P} = [\mathbf{P}_{(1,1)}^T, \ldots, \mathbf{P}_{(1,G_1)}^T, \ldots, \mathbf{P}_{(J,1)}^T, \ldots, \mathbf{P}_{(J,G_J)}^T]^T$
    $\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^T(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2})}{\mathbf{W}^T(\hat{\mathbf{V}}^{\cdot -1}) + \mathbf{P}} \right)^{\cdot \eta}$
    $\mathbf{H}_b \leftarrow \mathbf{H}_b \odot \left(\frac{\mathbf{W}_b^T(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2})}{\mathbf{W}_b^T \hat{\mathbf{V}}^{\cdot -1}} \right)^{\cdot \eta}$
    $\mathbf{W}_b \leftarrow \mathbf{W}_b \odot \left(\frac{(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2})\mathbf{H}_b^T}{\hat{\mathbf{V}}^{\cdot -1}\mathbf{H}_b^T} \right)^{\cdot \eta}$
    Normalize $\mathbf{W}_b$ and $\mathbf{H}_b$ component-wise (see, e.g., [20])
    $\hat{\mathbf{V}} \leftarrow \mathbf{W}\mathbf{H} + \mathbf{W}_b\mathbf{H}_b$
  until convergence

Note that the updates of $\mathbf{H}$ are identical in both the supervised and semi-supervised cases. Additionally, in the semi-supervised case, since the derivations of (11) with respect to $\mathbf{W}_\mathrm{b}$ and $\mathbf{H}_\mathrm{b}$ are not affected by the sparsity constraint $\Psi(\mathbf{H})$, the updates of $\mathbf{W}_\mathrm{b}$ and $\mathbf{H}_\mathrm{b}$ are straightforwardly derived as in [30].

B. Relative group sparsity constraints and parameter estimation algorithm

For the separation to be feasible, we require that every learned source model has a corresponding non-zero activation; however, this constraint is not enforced by the group sparsity penalty in (12) where it can happen that a group of different sources are fit together using the same source model, instead of separately using their designated models, rendering their separation impossible. We observed this "source vanishing" phenomenon in practice as illustrated in Fig. 4a (in case of

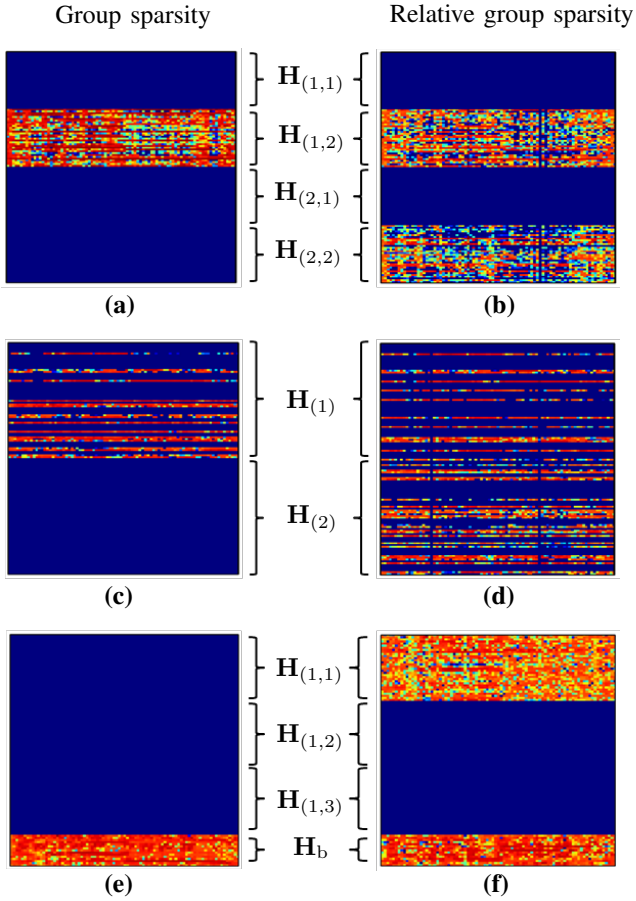


Fig. 4. Examples of estimated activation matrices $\mathbf{H}$ for two sources in a mixture containing a rooster and bird chirps where two retrieved examples for each source were used for training the USCMs. Left column: **(a)** block sparsity in the supervised case, **(c)** component sparsity in the supervised case, and **(e)** block sparsity in the semi-supervised case. Right column: same settings as in the left column, but for the proposed relative block/component sparsity.

using the block sparsity-inducing penalty) and Fig. 4c (in case of using the component sparsity-inducing penalty) in the supervised case. Moreover, the problem worsens in the semi-supervised case, depicted in Fig. 4e for the block sparsity case, where the entire mixture is fit by the estimated background model only (the same effect occurs for the component sparsity case). This is due to the fact that $\mathbf{W}_{\mathrm{b}}$ and $\mathbf{H}_{\mathrm{b}}$ are now fully unconstrained in (11), whereas $\mathbf{W}$ is fixed and $\mathbf{H}$ is constrained by the group sparsity-inducing penalty. It can also be seen that increasing the trade-off parameters λ_j in the penalty (12) (thus decreasing the number of active groups) increases the chances of source vanishing in both the supervised and semi-supervised cases. In this section, we present a novel sparsity-inducing penalty which helps prevent this problem completely.

1) Relative group sparsity-inducing penalties: This observation motivates us to introduce a general solution based on a new notion, namely *relative group sparsity*. While we present it here within the context of NMF, the idea extends to other dictionary decomposition schemes. We assume that the groups are organized into so-called *supergroups* (*i.e.* $\mathbf{H}_{(j)}$ corresponding to a USCM is considered as a supergroup), and we characterize the relative group sparsity constraint $\Psi(\mathbf{H})$ by the following properties

- It induces the sparsity of the groups (as in group sparsity), and at the same time
- It induces the anti-sparsity of the supergroups (*i.e.* prevents them from vanishing entirely).

In other words, the group sparsity property is now considered *relative* to the corresponding supergroup $\mathbf{H}_{(j)}$ and not within the full set of coefficients in $\mathbf{H}$. It is formulated as [27]

$$\Psi_{\mathrm{rel}}(\mathbf{H}) = \sum\nolimits_{j=1}^{J} \lambda_j \sum\nolimits_{g=1}^{G_j} \log\left(\frac{\epsilon + \|\mathbf{H}_{(j,g)}\|_1}{\|\mathbf{H}_{(j)}\|_1^{\gamma_j}}\right), \quad (13)$$

where γ_j are some non-negative constants. Penalty (13) can be also rewritten as

$$\Psi_{\mathrm{rel}}(\mathbf{H}) = \Psi_{\mathrm{gr}}(\mathbf{H}) - \sum\nolimits_{j=1}^{J} \lambda_j \gamma_j G_j \log\left(\|\mathbf{H}_{(j)}\|_1\right), \quad (14)$$

and one can easily see that, while the new penalty keeps the group sparsity property thanks to $\Psi_{\mathrm{gr}}(\mathbf{H})$ defined in (12), it prevents (when $\gamma_j > 0$) the supergroups from vanishing since if $\|\mathbf{H}_{(j)}\|_1$ tends to zero, then $-\log\left(\|\mathbf{H}_{(j)}\|_1\right)$ tends to $+\infty$. This formulation generalizes the group sparsity constraint in the sense that (13) reduces to (12) for $\gamma_j = 0$. So while we only require that γ_j is non-zero for the relative group sparsity to be active, in our experiments we show results for $\gamma_j = 1$ and $\gamma_j = \frac{1}{G_j}$. The latter was chosen to act as a normalization such that the effect of the penalty is even across the USCMs regardless of their size.

Note also that one can introduce either the *relative block sparsity-inducing penalty* or the *relative component sparsity-inducing penalty* by defining a group $\mathbf{H}_{(j,g)}$ to be either a block or a row in $\mathbf{H}$ similar to what has been presented in Section IV-A.

2) MU rules for parameter estimation: MU rules for parameter estimation when using the new penalty $\Psi_{\mathrm{rel}}(\mathbf{H})$ are derived in the same way as the rules for group sparsity in Section IV-A. The resulting algorithms for both supervised and semi-supervised cases are summarized in Algorithm 3 and Algorithm 4, respectively. Details on the derivation of Algorithm 3 are given in the Appendix, and that of Algorithm 4 is very similar. Note that $\mathbf{P}_{(j,g)}$ and $\mathbf{Q}_{(j,g)}$ are matrices of the same size as $\mathbf{H}_{(j,g)}$ whose entries have the same value, and $\mathbf{P}$ and $\mathbf{Q}$ are concatenations of $\mathbf{P}_{(j,g)}$ and $\mathbf{Q}_{(j,g)}$, respectively.

V. Experiments

We start by describing the data set, parameter settings, and evaluation metrics in Section V-A. We then evaluate the performance of the proposed supervised and semi-supervised on-the-fly audio source separation algorithms in Section V-B. The sensitivity of the different algorithms with respect to the choice of the trade-off parameter λ_j which determines the contribution of the sparsity penalty is presented in Section V-C. We finally present user-test results in Section V-D.

A. Data, parameter settings, and evaluation metrics

We evaluated the performance of the proposed on-the-fly algorithms on a data set of 15 single-channel mixtures of

Algorithm 3 MU rules for NMF with relative group sparsity in the supervised case

Input: $\mathbf{V}$, $\mathbf{W}$, $\boldsymbol{\lambda}$
Output: $\mathbf{H}$
  Initialize $\mathbf{H}$ randomly
  $\hat{\mathbf{V}} = \mathbf{WH}$
  repeat
    for $j = 1, \ldots, J, g = 1, \ldots, G_j$ **do**
      $\mathbf{P}_{(j,g)} \leftarrow \frac{\lambda_j}{\epsilon + \|\mathbf{H}_{(j,g)}\|_1}$
      $\mathbf{Q}_{(j,g)} \leftarrow \frac{\lambda_j G_j \gamma_j}{\|\mathbf{H}_{(j)}\|_1}$
    end for
    $\mathbf{P} = [\mathbf{P}^T_{(1,1)}, \ldots, \mathbf{P}^T_{(1,G_1)}, \ldots, \mathbf{P}^T_{(J,1)}, \ldots, \mathbf{P}^T_{(J,G_J)}]^T$
    $\mathbf{Q} = [\mathbf{Q}^T_{(1,1)}, \ldots, \mathbf{Q}^T_{(1,G_1)}, \ldots, \mathbf{Q}^T_{(J,1)}, \ldots, \mathbf{Q}^T_{(J,G_J)}]^T$
    $\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^T(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2}) + \mathbf{Q}}{\mathbf{W}^T(\hat{\mathbf{V}}^{\cdot -1}) + \mathbf{P}} \right)^{\cdot \eta}$
    $\hat{\mathbf{V}} \leftarrow \mathbf{WH}$
  until convergence

Algorithm 4 MU rules for NMF with relative group sparsity in the semi-supervised case

Input: $\mathbf{V}$, $\mathbf{W}$, $\boldsymbol{\lambda}$
Output: $\mathbf{H}$
  Initialize $\mathbf{H}$, $\mathbf{H}_b$, and $\mathbf{W}_b$ randomly
  $\hat{\mathbf{V}} \leftarrow \mathbf{WH} + \mathbf{W}_b\mathbf{H}_b$
  repeat
    for $j = 1, \ldots, J, g = 1, \ldots, G_j$ **do**
      $\mathbf{P}_{(j,g)} \leftarrow \frac{\lambda_j}{\epsilon + \|\mathbf{H}_{(j,g)}\|_1}$
      $\mathbf{Q}_{(j,g)} \leftarrow \frac{\lambda_j G_j \gamma_j}{\|\mathbf{H}_{(j)}\|_1}$
    end for
    $\mathbf{P} = [\mathbf{P}^T_{(1,1)}, \ldots, \mathbf{P}^T_{(1,G_1)}, \ldots, \mathbf{P}^T_{(J,1)}, \ldots, \mathbf{P}^T_{(J,G_J)}]^T$
    $\mathbf{Q} = [\mathbf{Q}^T_{(1,1)}, \ldots, \mathbf{Q}^T_{(1,G_1)}, \ldots, \mathbf{Q}^T_{(J,1)}, \ldots, \mathbf{Q}^T_{(J,G_J)}]^T$
    $\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^T(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2}) + \mathbf{Q}}{\mathbf{W}^T(\hat{\mathbf{V}}^{\cdot -1}) + \mathbf{P}} \right)^{\cdot \eta}$
    $\mathbf{H}_b \leftarrow \mathbf{H}_b \odot \left(\frac{\mathbf{W}_b^T\left(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2}\right)}{\mathbf{W}_b^T \hat{\mathbf{V}}^{\cdot -1}} \right)^{\cdot \eta}$
    $\mathbf{W}_b \leftarrow \mathbf{W}_b \odot \left(\frac{\left(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2}\right)\mathbf{H}_b^T}{\hat{\mathbf{V}}^{\cdot -1}\mathbf{H}_b^T} \right)^{\cdot \eta}$
    Normalize $\mathbf{W}_b$ and $\mathbf{H}_b$ component-wise (see, e.g., [20])
    $\hat{\mathbf{V}} \leftarrow \mathbf{WH} + \mathbf{W}_b\mathbf{H}_b$
  until convergence

two sources artificially mixed at 0 dB signal to noise ratio (SNR). Note that during the mixing, we made sure that two sources had more or less the same duration so that in all the mixtures both sources appear most of the time. The mixtures were sampled at either 16000 Hz or 11025 Hz and their duration varies between 1 and 13 seconds. The sources in the mixtures were selected as follows: (*female speech, traffic*), (*female speech, cafe*), (*male speech, bells*), (*male speech, car*), (*woman singing, restaurant*), (*drums, guitar*), (*applause, electric guitar*), (*piano, ringtone*), (*violin, cough*), (*bat, owl*), (*chirps, rooster*), (*chirps, river*), (*siren, dog*), (*cat, dog*), and (*ocean, cricket*). The speech samples (*female speech*, *male speech*) were obtained from the "American English" ITU-T P.501[2] dataset. The following sources *cafe*, *car*, and *restaurant* were obtained from DEMAND[3] from one channel out of the 16 channels. The music instruments (*drums*, *electric guitar*, *guitar*, *piano*, *violin*, *woman singing*) were obtained from QUASI[4]. The remainder were from various websites, mostly www.grsites.com/archive/sounds/ (*bells*, *cat*, *chirps*, *dog*, *rooster*, *river*, *traffic*), but also www.sounddogs.com (*bat*, *owl*) and www.wavlist.com (*cricket*), among others. The diversity in the types of sources should demonstrate the advantage of the proposed on-the-fly strategy since, as opposed to speech where pre-trained models are fairly common, having a pre-trained model for every possible sound class is not viable. In the implementation of the framework, sound examples for training were retrieved from www.findsounds.com, a search engine for audio, as well as from www.freesound.org, a database of user-uploaded sounds. Note that these two websites are different from the ones used to get the sources in the mixtures; thus the possibility that the training set contains a source from the mixtures is very small. The retrieved files were restricted to those with sampling rates at least as high as that of the mixture, and the ones with higher sampling rates were down-sampled accordingly. For retrieval in our experiments, we differentiate between two types of search keywords: i) *reference* keywords given by an expert (the first author) who prepared the dataset and thus had also listened to the separate sources and not only the mixtures and ii) *user* keywords given by non-expert users in our user test. It is important to note that the reference keywords are not the only "correct" keywords since other synonyms can be used. Table I lists the reference keywords and the corresponding user keywords along with the number of times a keyword was given by the users. Note that some reference keywords like "male speech" or "female speech" are repeated in more than one mixture, thus the count of their corresponding user keywords is more than the number of users.

Other parameters were set as follows. The STFT was calculated using a sine window and a frame length of 47 ms with 50% overlap, the number of iterations for MU updates was 200 for learning the USCM $\mathbf{W}_{(j)}$ and 100 for separating the mixture, and the number of NMF components for each spectral model learned from one example $\mathbf{W}_{(j,p)}$ was set to 32. In the semi-supervised case, the number of NMF components for the background source was $K_b = 10$ to avoid overfitting since $\mathbf{W}_b$ is unconstrained. Additionally, since the number of training examples P_j per source was different depending on the availability of the data (search results), the trade-off parameter λ_j determining the contribution of the sparsity-inducing penalty was set to $\lambda_0 FNP_j$ (where λ_0 is a constant) so that λ_j is greater when more examples are available. The intuition here is that the smaller the USCM $\mathbf{W}_{(j)}$ is, which happens when few examples are available, the lower the level of sparsity that should be imposed in the decomposition. We found these settings to generally result in a good separation performance.

[2]http://www.itu.int/rec/T-REC-P.501
[3]http://parole.loria.fr/DEMAND/
[4]http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/

TABLE I
REFERENCE KEYWORDS AND THE CORRESPONDING USER KEYWORDS.

Reference keywords	User keywords
applause	background noise, cheers, concert, concert crowd, crowd, crowd cheering (2), crowd concert, people cheering
bat	bird (2), bird cackling, bird chirping, birds sound, monkey (2), jungle, night animal
bells	bell tone, bells, bells church, church bell (3), church bells (3)
cafe	chattering, crowd, crowd speech, crowd talking (2), many people talking, party, people, people talking
car	aeroplane noise, ambient noise, boat motor, calm noise, car, drive, nothing, thunder storm, wind
cat	cat (7), cat meow, cat meowing
chirps	bird (4), bird chirping (6), birds (2), birds chirping (2), birds sing (2), night creatures, sparrows
cough	caugh, cough (2), coughing (5), man caughing
cricket	bird, birds, birds sing, cricket (2), night, night animal, night creatures, tweet-tweet
dog	dog (5), dog bark, dog barking (2), dogs
drums	bass drum, drum (3), drum beats, drums, percussion, rythmic, tap beats
electric guitar	electric guitar, guitar (2), guitar concert, music (3), music playing, riff guitar
female speech	female speech (3), female voice (2), female voice english, girl read, girl talking (5), woman, woman read, woman speak, woman speaking, woman speech, woman talking (2), woman voice
guitar	acoustic guitar, electronic organ, guitar (7)
male speech	male english speech (3), male speech english, man reading, man speak, man speaking (2), man speech, man talking (4), man voice, men speech, poetry recitation, read (2)
ocean	car, driving a car, road traffic, sea waves, storm, street, traffic noice, waterfall, waves
owl	dog, dog moaning, owl (4), owl hooting, pigeon, woodpecker
piano	pianist, piano (4), piano music (3), soft piano strings
restaurant	chattering ambiance, crowd (2), crowd noise with photo clicks, crowd speech, crowd talking, people noise, people talking (2)
ringtone	jingle phone ringing, mobile ringtone (2), phone ringing, phone ringtone, ring, ringing, ringtone, smartphone ring
river	motor engine noise, river, river flowing, sea, stream, water (2), water boiling, water flowing
rooster	cock (2), cock cluck, cock-a-doodle-do, hen (2), rooster (3)
siren	ambulance, police, police car, police siren (5), siren
traffic	car, car passing, car running, road traffic, road with cars, street traffic, traffic noice, traffic noise, traffic sound
violin	cello strings orchestra, music album, music (2), piano, soundtrack, violin (3)
woman singing	brasilian woman singing, brazilian song, girl singing (3), singing woman, woman singing (3)

The source separation performance was evaluated in terms of the normalized signal-to-distortion ratio (NSDR), which measures the overall signal distortion, and the normalized signal-to-interference ratio (NSIR) which measures the leakage of the other sources [31], [32]. Recall that the normalized values are computed by subtracting the SDR and SIR of the original mixture signal from those of the separated sources [32]. The normalization serves to show the gain of using the proposed source separation system as opposed to a naive method that simply assigns the mixture as a source estimate. These metrics are measured in dB and are averaged over all sources and all mixtures for the different algorithms.

B. Separation results using reference keywords

In this experiment, we use the reference keywords given by the expert for retrieval. The goal is to evaluate and compare the performance of the different algorithms. For the supervised case (*i.e.*, Algorithm 1 and Algorithm 3), two keywords were used to retrieve examples for both sources in the mixture, while only one keyword was used for the semi-supervised case (*i.e.*, Algorithm 2 and Algorithm 4). Note that, in the semi-supervised scenario, we tested two cases as follows (i) a keyword was provided for source 1 only and (ii) a keyword was provided for source 2 only; we then averaged the obtained separation results.

We compare the average separation performance obtained using the four sparsity-inducing penalties presented in the paper: *block sparsity* as the baseline [25], the proposed *component sparsity*, *relative block sparsity*, and *relative component sparsity*. Results for the supervised case (*i.e.,* Algorithm 1 for *block sparsity* and *component sparsity*, and Algorithm 3 for *relative block sparsity*, and *relative component sparsity*) are shown in Table II, while those for the semi-supervised case (*i.e.,* Algorithm 2 and Algorithm 4) are shown in Table III. In each case, we run the algorithms with different values of the trade-off parameter λ_0, and the value resulting in the highest average NDSR is chosen and shown in the tables along with the corresponding NSDR and NSIR. Note that the result shown in Table II is 1.8 dB NDSR higher than that reported in our previous work [24]. The reason is that: (1) the dataset (training and testing set) is enlarged by the size and the variation of the sound sources; and (2) the parameter λ_j is here adapted per mixture and not constant for the whole dataset. Also, for easier reading here, we do not compare again the separation performance with the standard supervised NMF setting without using USCM model nor with some other baselines as it has been investigated in our previous study [24]. For the relative sparsity cases, we tested two values for the hyper-parameter γ_j: a fixed $\gamma_j = 1$ as a natural choice, and $\gamma_j = \frac{1}{G_j}$ such that the denominator term $\|\mathbf{H}_{(j)}\|_1^{\gamma_j}$ in the penalty (13) is adaptively normalized with respect to the size of the group G_j.

First, as expected, the results obtained in the supervised case are much better than those achieved in the semi-supervised

TABLE II
SUPERVISED CASE: AVERAGE SOURCE SEPARATION PERFORMANCE.

Method	NSDR	NSIR
Block sparsity [25] ($\lambda_0 = 1 \times 10^{-4}$)	5.14	9.80
Component sparsity ($\lambda_0 = 1 \times 10^{-6}$)	5.91	10.67
Rel. block sparsity ($\gamma_j = 1, \lambda_0 = 1 \times 10^{-4}$)	4.78	9.27
Rel. component sparsity ($\gamma_j = 1, \lambda_0 = 1 \times 10^{-6}$)	**6.15**	10.70
Rel. block sparsity ($\gamma_j = \frac{1}{G_j}, \lambda_0 = 1 \times 10^{-4}$)	5.03	9.44
Rel. component sparsity ($\gamma_j = \frac{1}{G_j}, \lambda_0 = 1 \times 10^{-6}$)	5.96	**10.72**

TABLE III
SEMI-SUPERVISED CASE: AVERAGE SOURCE SEPARATION PERFORMANCE.

Method	NSDR	NSIR
Block sparsity ($\lambda_0 = 1 \times 10^{-4}$)	0.74	4.66
Component sparsity ($\lambda_0 = 2 \times 10^{-8}$)	1.98	6.22
Rel. block sparsity ($\gamma_j = 1, \lambda_0 = 4 \times 10^{-4}$)	1.68	6.03
Rel. component sparsity ($\gamma_j = 1, \lambda_0 = 5 \times 10^{-7}$)	**2.31**	**6.64**
Rel. block sparsity ($\gamma_j = \frac{1}{G_j}, \lambda_0 = 1 \times 10^{-4}$)	1.09	5.74

case where examples for one source are missing. Second, using an adaptive $\gamma_j = \frac{1}{G_j}$ in the supervised case improved the NSDR for the relative block sparsity by 0.25 dB but had no significant effect on the relative component sparsity; in contrast it negatively affected the performance in the semi-supervised case. Third, we note that the proposed component sparsity-inducing penalty achieves a better separation performance than the block sparsity-inducing penalty which was exploited in [25], in both supervised and semi-supervised cases. A possible explanation is that the former offers more flexibility by exploiting the most representative spectral patterns from different spectral models that match the mixture. Last, it is worth noting that the proposed relative component sparsity-inducing penalty performs the best in both supervised and semi-supervised cases in terms of both NSDR and NSIR, the advantage being more significant in the semi-supervised case likely because the source vanishing problem is more severe. We note that the corresponding average signal-to-artifact ratio (SAR) for the different algorithms was on the order of 11 dB. In particular, the SAR corresponding to the relative component sparsity-inducing penalty was 10.98 dB and 11.35 dB for the adaptive γ_j.

In general, the methods would fail if the retrieved examples are quite dissimilar from the actual sources in the mixture. As an example, a mixture of of an electric guitar and applause (cheers and whistles) had low NSDR for both source estimates (0.54 dB and -0.49 dB respectively). In this case, we observed that the retrieved training files for the applause contained mostly just clapping sounds and as such the learned USCM did not capture the cheers; similarly for the guitar where most retrieved examples were not close to the chords in the mixture.

C. Separation results with different choices of λ_j

One of the most important parameters in the presented algorithms in the on-the-fly framework is the trade-off parameter λ_j determining the contribution of the sparsity-inducing penalty. We propose to set $\lambda_j = \lambda_0 FNP_j$ so that it is normalized with respect to the size of the USCM and is larger when more examples are used. In this experiment, we varied λ_0 and assessed the sensitivity of the different algorithms described in Section V-B to this choice in the semi-supervised scenario. The dataset and other parameter settings are the same as described before. The results are shown in Fig. 5 where $\lambda_0 = \{10^{-5}, 5 \times 10^{-5}, 10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$ for the block/relative block sparsity algorithms and $\lambda_0 = \{10^{-7}, 5 \times 10^{-7}, 10^{-6}, 5 \times 10^{-6}, 10^{-5}\}$ for the component/relative component sparsity algorithms. Note that the range of λ_0 is different for the (relative) block and component sparsity algorithms as they are different types of penalties so their optimal range is different.

As can be seen, the relative block sparsity and relative component sparsity algorithms are generally more stable than the block sparsity and component sparsity algorithms over a large range of λ_0 where the results obtained by the former algorithms drop sharply for the last point. Within a good range, *i.e.* the first four points, the relative block sparsity with $\gamma_j = \frac{1}{G_j}$ is the most stable one as its NSDR varies at most 0.2 dB. The relative component sparsity algorithm, which offers the highest performance in general, is not very sensitive to the considered parameter though it has more than 1 dB NSDR difference within the considered range.

D. Separation results for the user test

In the second experiment, our goal was to evaluate the performance of the proposed on-the-fly framework when practically used by non-expert users. We also test the effect of the examples refinement step on learning the USCM. The algorithms based on the proposed *relative* block/component sparsity-inducing penalties, which perform better than those using the block/component sparsity-inducing penalties as shown in Section V-B, were tested using the input from 9 different users who were of different age groups, technical backgrounds, and were all not native English speakers. The best parameter settings as determined from Section V-B were used. Using the GUI described in Section II-B, the users were asked to process each of the 15 mixtures as follows. First, they were asked to listen to the mixture and accordingly type keywords describing the two sources. They were instructed to change the keywords in case the search engine did not return results. Then, they were required to listen to the retrieved examples and select those that sound more similar to the sources in the mixture; at least one example was required to be selected. Given the recorded user input (keywords and selected examples), we examine two possibilities of using the examples in guiding the separation process as follows:

- All retrieved examples are used (All).
- Only the subset of examples selected by the user is used (Subset).

The source separation performance, averaged over all 9 users and over all mixtures per method, is shown in Table IV and Table V for the supervised and semi-supervised cases, respectively. We note that the results for the average user are mostly lower than those for the expert in the supervised case due to the following issues. As can be seen from the keywords in Table I, some sounds like *bat* and *owl* were sometimes

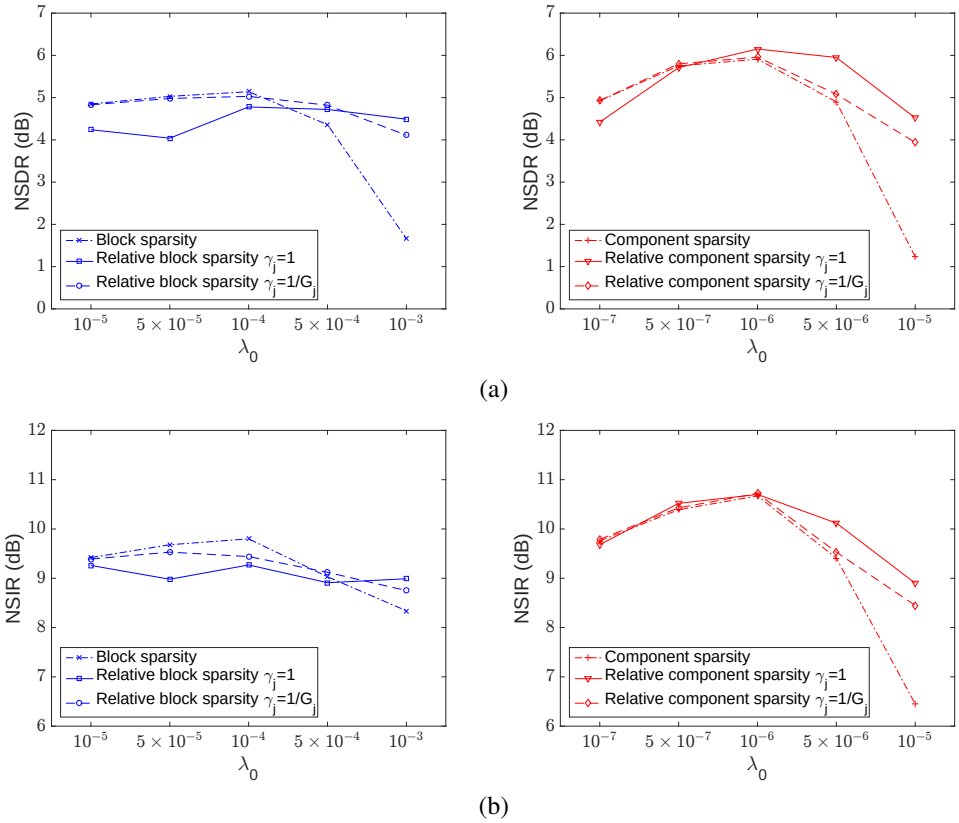


Fig. 5. Separation performance of the different algorithms, in terms of NSDR (a) and NSIR (b), as a function of λ_0.

not recognized by the users and were confused with other sounds (e.g., bird sounds). Also, some spelling mistakes can be found (e.g, *caugh* instead of *cough*). This may have negatively affected the results. Additionally, one of the mixtures included a popular ringtone composed of marimba notes; however, the retrieved examples mainly included classical telephone rings, perhaps "marimba" would have been a better choice for searching. In the semi-supervised case, the expert results are not better than the average user. The reason is likely that the guidance is reduced in this setting and the overall performance is quite lower compared to the supervised case.

Nevertheless, the performance globally follows the same trend as presented in Section V-B: relative component sparsity generally outperforms relative block sparsity, especially in the semi-supervised case, with the SAR on the order of 9 dB. It is interesting to observe the effect of selecting a subset of examples. As can be seen in Table IV, using a subset of examples selected by the users only improves the performance in the supervised case. However, in the semi-supervised case, such a pre-selection even negatively affects the results as can be observed in Table V. This is likely due to the fact that having few selected examples (only one in the extreme case) leads to having fewer components in the learned spectral model for which a sparse decomposition is not optimal. Thus, it seems to be better to keep all retrieved examples for the known source and let the relative component sparsity penalty induce the appropriate selection.

TABLE IV
USER TEST IN THE SUPERVISED CASE: AVERAGE SOURCE SEPARATION PERFORMANCE.

Method	NSDR	NSIR
Relative block sparsity (All)	2.42	7.53
Relative block sparsity (Subset)	**3.16**	**8.24**
Relative component sparsity (All)	2.91	7.75
Relative component sparsity (Subset)	2.98	8.19

TABLE V
USER TEST IN THE SEMI-SUPERVISED CASE: AVERAGE SOURCE SEPARATION PERFORMANCE.

Method	NSDR	NSIR
Relative block sparsity (All)	1.88	7.50
Relative block sparsity (Subset)	1.24	7.34
Relative component sparsity (All)	**2.78**	**8.04**
Relative component sparsity (Subset)	1.53	7.60

VI. CONCLUSION

In this paper, we presented the novel concept of on-the-fly audio source separation and described several algorithms to implement it. Specifically, we proposed using a universal sound class model learned by NMF from retrieved examples and imposing group sparsity-inducing constraints to efficiently handle the selection of the most representative spectral patterns. Additionally, we introduced the notion of relative group sparsity to overcome a so-called *source vanishing* problem that occurs in the considered on-the-fly paradigm. In contrast to

other state-of-the-art user-guided approaches, the considered framework greatly simplifies the user interaction with the system such that everyone, not necessarily an expert, can do source separation by just typing keywords describing the audio sources in the mixture. Experiments on mixtures containing various types of sounds confirm the potential of the proposed framework as well as the corresponding algorithms. Future work may be devoted to running real-world experiments, studying the use of a different group sparsity model that induces dynamic relationships between atoms or groups [33], as well as extending the framework to multichannel mixtures where *spatial* source models (*e.g.* those from [34] or [35]) may also be learned. Additionally, investigating the optimal USCM model size for different types of sound sources would be an interesting direction.

APPENDIX
DERIVATION OF MU RULES IN ALGORITHM 3

Let $C(\mathbf{H})$ denote the right part of criterion (10) with relative group sparsity penalty $\Psi(\mathbf{H}) = \Psi_{\mathrm{rel}}(\mathbf{H})$ defined as in (13) and $D(\cdot\|\cdot)$ being IS divergence specified as in equations (3) and (4). The partial derivative of $C(\mathbf{H})$ with respect to h_{kn} writes

$$\nabla_{h_{kn}} C(\mathbf{H}) = \sum_{f=1}^{F} w_{fk} \left(\frac{1}{[\mathbf{WH}]_{fn}} - \frac{v_{fn}}{[\mathbf{WH}]_{fn}^2} \right) + \frac{\lambda_j}{\epsilon + \|\mathbf{H}_{(j,g)}\|_1} - \frac{\lambda_j G_j \gamma_j}{\|\mathbf{H}_{(j)}\|_1} \quad (15)$$

Following a standard approach for MU rules derivation (see *e.g.*, [20], [28]), we represent $\nabla_{h_{kn}} C(\mathbf{H})$ as

$$\nabla_{h_{kn}} C(\mathbf{H}) = \nabla^+_{h_{kn}} C(\mathbf{H}) - \nabla^-_{h_{kn}} C(\mathbf{H}) \quad (16)$$

with $\nabla^+_{h_{kn}} C(\mathbf{H}), \nabla^-_{h_{kn}} C(\mathbf{H}) \geq 0$ defined as

$$\nabla^+_{h_{kn}} C(\mathbf{H}) \triangleq \sum_{f=1}^{F} w_{fk} \frac{1}{[\mathbf{WH}]_{fn}} + \frac{\lambda_j}{\epsilon + \|\mathbf{H}_{(j,g)}\|_1}, \quad (17)$$

$$\nabla^-_{h_{kn}} C(\mathbf{H}) \triangleq \sum_{f=1}^{F} w_{fk} \frac{v_{fn}}{[\mathbf{WH}]_{fn}^2} + \frac{\lambda_j G_j \gamma_j}{\|\mathbf{H}_{(j)}\|_1}, \quad (18)$$

and we update each parameter h_{kn} as

$$h_{kn} \leftarrow h_{kn} \left(\frac{\nabla^-_{h_{kn}} C(\mathbf{H})}{\nabla^+_{h_{kn}} C(\mathbf{H})} \right)^{.\eta}, \quad (19)$$

where $\eta = 0.5$ following the derivation in [21]. Rewritten in a matrix form, we obtain the updates of the activation matrix $\mathbf{H}$ in Algorithm 3.

ACKNOWLEDGMENT

The authors would like to thank all the colleagues at Technicolor's Rennes research center who participated in the experiments.

REFERENCES

[1] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*, Springer, 2007.
[2] O. Yılmaz and S. T. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.
[3] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007, article ID 24717.
[4] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, pp. 2353–2362, 2001.
[5] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The Signal Separation Campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.
[6] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *Proc. Latent Variable Analysis and Signal Separation*, 2015, pp. 387–395.
[7] S. Ewert, B. Pardo, M. Mueller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, 2014.
[8] L. L. Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, pp. 1–5, 2014.
[9] J. Ganseman, P. Scheunders, G. J. Mysore, and J. S. Abel, "Source separation by score synthesis," in *Proc. ICMC*, 2010, pp. 462–465.
[10] J. Fritsch and M. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 888–891.
[11] N. Souviraá-Labastie, E. Vincent, and F. Bimbot, "Music separation guided by cover tracks: designing the joint nmf model," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) 2015*, 2015.
[12] N. Souviraá-Labastie, A. Olivero, E. Vincent, and F. Bimbot, "Multi-channel audio source separation using multiple deformed references," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 11, no. 23, pp. 1775–1787, 2015.
[13] P. Smaragdis and G. J. Mysore, "Separation by humming: User-guided sound extraction from monophonic mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 69–72.
[14] N. Q. K. Duong, A. Ozerov, and L. Chevallier, "Temporal annotation-based audio source separation using weighted nonnegative matrix factorization," in *IEEE Int. Conf. on Consumer Electronics (ICCE-Berlin)*, 2014.
[15] A. Lefèvre, F. Bach, and C. Févotte, "Semi-supervised NMF with time-frequency annotations for single-channel source separation," in *Int. Conf. on Music Information Retrieval (ISMIR)*, 2012, pp. 115–120.
[16] N. J. Bryan and G. J. Mysore, "Interactive refinement of supervised and semi-supervised sound source separation estimates," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 883–887.
[17] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot, "An interactive audio source separation framework based on nonnegative matrix factorization," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 1586–1590.
[18] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "On-the-fly specific person retrieval," in *13th Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2012, pp. 1–4.
[19] K. Chatfield and A. Zisserman, "Visor: Towards on-the-fly large-scale object category retrieval," in *Asian Conference on Computer Vision*, ser. Lecture Notes in Computer Science. Springer, 2012, pp. 432–446.
[20] C. Févotte, N. Bertin, and J. Durrieu, "Non-negative matrix factorization with the Itakura-Saito divergence. with application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.
[21] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito non-negative matrix factorization with group sparsity," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 21–24.
[22] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2007, pp. 414–421.

[23] T. Gerber, M. Dutasta, L. Girin, and C. Févotte, "Audio source separation using multiple deformed references," in *International Society for Music Information Retrieval Conf. (ISMIR)*, 2012.
[24] D. El Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly audio source separation," in *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.
[25] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 141–145.
[26] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.
[27] D. El Badawy, A. Ozerov, and N. Q. K. Duong, "Relative group sparsity for non-negative matrix factorization with application to on-the-fly audio source separation," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, accepted.
[28] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural and Information Processing Systems 13*, 2001, pp. 556–562.
[29] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Stat.*, vol. 58, no. 1, pp. 30–37, Feb. 2004.
[30] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, Sep. 2011.
[31] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
[32] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 90–93.
[33] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Similarity induced group sparsity for non-negative matrix factorisation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4425–4429.
[34] N. Q. K. Duong, E. Vincent, and R. 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.
[35] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 727–739, 2014.

Dalia El Badawy is currently a doctoral student at Ecole polytechnique fédérale de Lausanne (EPFL). She received the M.Sc. in communication systems from EPFL in 2015 and the B.Sc. in digital media engineering and technology from the German University in Cairo in 2012. Her research interests include applications of audio signal processing and acoustics as well as machine learning.

Ngoc Q. K. Duong received the B.S. degree from Posts and Telecommunications Institute of Technology (PTIT), Vietnam, in 2004, and the M.S. degree in electronic engineering from Paichai University, Korea, in 2008. He obtained the Ph.D. degree at the French National Institute for Research in Computer Science and Control (INRIA), Rennes, France in 2011.

From 2004 to 2006, he was with Visco JSC as a System Engineer. He was also a Research Engineer for the acoustic echo/noise cancelation system at Emersys Company, Korea in 2008. He is currently a Senior Scientist at Technicolor R&D France where he has worked since Nov. 2011. His research interest concerns signal processing (audio, image, and video), machine learning, and affective computing. He has received several research awards, including the IEEE Signal Processing Society Young Author Best Paper Award in 2012 and the Bretagne Young Researcher Award in 2015. He is the co-author of more than 30 scientific papers and about 25 pending patents.

Alexey Ozerov holds a Ph.D. in Signal Processing from the University of Rennes 1 (France). He worked towards this degree from 2003 to 2006 in the labs of France Telecom R&D and in collaboration with the IRISA institute. Earlier, he received an M.Sc. degree in Mathematics from the Saint-Petersburg State University (Russia) in 1999 and an M.Sc. degree in Applied Mathematics from the University of Bordeaux 1 (France) in 2003. From 1999 to 2002, Alexey worked at Terayon Communicational Systems (USA) as a R&D software engineer, first in Saint-Petersburg and then in Prague (Czech Republic). He was for one year (2007) in Sound and Image Processing Lab at KTH (Royal Institute of Technology), Stockholm, Sweden, for one year and half (2008-2009) in TELECOM ParisTech / CNRS LTCI - Signal and Image Processing (TSI) Department, and for two years (2009 - 2011) with METISS team of IRISA / INRIA - Rennes. Now he is a Senior Scientist in Technicolor Research & Innovation at Rennes, France. Since 2016 he is a Distinguished member of the Technicolor Fellowship Network and currently he is a member of IEEE Signal Processing Society Audio and Acoustic Signal Processing Technical Committee. He received the IEEE Signal Processing Society Best Paper Award in 2014. His research interests include image processing, audio restauration, audio source separation, source coding, audio classification and automatic speech recognition.

A. PAPER 1: ON-THE-FLY AUDIO SOURCE SEPARATION—A NOVEL USER-FRIENDLY FRAMEWORK

Appendix B

Paper 2: Gaussian Modeling-Based Multichannel Audio Source Separation Exploiting Generic Source Spectral Model

Gaussian modeling-based multichannel audio source separation exploiting generic source spectral model

Thanh Thi Hien Duong, Ngoc Q. K. Duong *Senior Member, IEEE*, Phuong Cong Nguyen, and Cuong Quoc Nguyen

***Abstract*—As *blind* audio source separation has remained very challenging in real-world scenarios, some existing works, including ours, have investigated the use of a *weakly-informed* approach where generic source spectral models (GSSM) can be learned a priori based on nonnegative matrix factorization (NMF). Such approach was derived for single-channel audio mixtures and shown to be efficient in different settings. This paper proposes a multichannel source separation approach where the GSSM is combined with the source spatial covariance model within a unified Gaussian modeling framework. We present the generalized expectation-minimization (EM) algorithm for the parameter estimation. Especially, for guiding the estimation of the intermediate source variances in each EM iteration, we investigate the use of two criteria: (1) the estimated variances of each source are constrained by NMF, and (2) the total variances of all sources are constrained by NMF altogether. While the former can be seen as a source variance denoising step, the latter is viewed as an additional separation step applied to the source variance. We demonstrate the speech separation performance, together with its convergence and stability with respect to parameter setting, of the proposed approach using a benchmark dataset provided within the 2016 Signal Separation Evaluation Campaign.**

KEYWORDS

Multichannel audio source separation, local Gaussian model, nonnegative matrix factorization, generic spectral model, group sparsity constraint.

I. INTRODUCTION

Real-world recordings are often mixtures of several audio sources, which usually deteriorate the target one. Thus many practical applications such as speech enhancement, sound post-production, and robotics use audio source separation technique [1], [2] to extract individual sound sources from their mixture. However, despite numerous effort in the past decades, blind source separation performance in reverberant recording conditions is still far from perfect [3], [4]. To improve the separation performance, *informed* approaches have been proposed and emerged recently in the literature [5], [6]. Such approaches exploit side information about either the sources themselves or the mixing condition in order to guide the separation process. Examples of the investigated side information include deformed or hummed references of one (or more) source(s) in a given mixture [7], [8], text associated with spoken speeches [9], temporal annotation of the source activity along the mixtures [10], core associated with musical sources [11], [12], and motion associated with audio-visual objects in a video [13]. Following this trend, some recent works including ours have proposed to use a very abstract semantic information just about the types of audio sources existing in the mixture to guide the source separation. If one source in the mixture is known as "speech", then several speaker-independent speech examples can be used to create a universal speech model as presented in [14]; if several types of sound sources in the mixture are known (*e.g.,* birdsong, piano, waterfall), their audio examples found by internet search can be used to learn the corresponding universal sound class models as presented in [15]. Such universal models were shown to be effective in guiding the source separation algorithm and resulted in promising performance. Inspired by this idea, we have further investigated the use of generic speech and noise model for single-channel speech separation in [16] and shown its promising result in (a) the supervised case, where both speech GSSM and noise GSSM are learned during training phase, and (b) the semi-supervised case, where only the speech GSSM is pre-learned. Furthermore, we have proposed to combine the block sparsity constraint investigated in [14] with the component sparsity constraint presented in [17] in a common formulation in order to take into account the advantage of both of them [18].

It should be noted that the works cited above [9], [12], [16], [18] considered only a single channel case, where the mixtures are mono, and exploited non-negative matrix factorization (NMF) [19], [20] to model the spectral characteristics of audio sources. Some recent works have investigated the use of the deep neural networks (DNN) to model the source spectra, where basically the types of sources in the mixture also need to be known as a side information in order to collect training data. Such DNN-based approaches were shown to offer very promising results in single-channel speech and music separation [21]–[23], multichannel speech separation [24], [25]. However, they require a large amount of labeled data for training, which may not always be available and the training is usually computationally expensive.

When more recording channels are available thanks to the use of multiple microphones, a multichannel source separation algorithm should be considered as it allows to exploit important information about the spatial locations of audio

Thanh Thi Hien Duong is with International Research Institute MICA and Hanoi University of Mining and Geology, Vietnam, e-mail: (duongthihienthanh@humg.edu.vn).

Ngoc Q. K. Duong is with Technicolor R&I, France, e-mail: (quang-khanh-ngoc.duong@technicolor.com).

Phuong Cong Nguyen is with MICA and Hanoi University of Science and Technology, Vietnam, e-mail: (phuong.nguyencong@hust.edu.vn).

Cuong Quoc Nguyen is with Hanoi University of Science and Technology, Vietnam, e-mail: (cuong.nguyenquoc@hust.edu.vn).

sources. Such spatial information is reflected in the mixing process (usually with reverberation), and can be modeled by *e.g.,* the interchannel time difference (ITD) and interchannel intensity difference (IID) [26]–[29], the rank-1 time-invariant mixing vector in the frequency domain when following the narrowband assumption [30]–[33], or the full-rank spatial covariance matrix in local Gaussian model (LGM) where the narrowband assumption is relaxed [34]–[36].

In this paper, we present an extension of the previous works [15], [16], [18] to the multichannel case where the NMF-based GSSM is combined with the full-rank spatial covariance model in a Gaussian modeling paradigm. Around this LGM, existing works have investigated several source spectral models such as Gaussian mixture model (GMM) [37], NMF as a linear model with nonnegativity constraints [36], [38], continuity model [39], kernel additive model [40], heavy-tailed distributions-based model [41], [42], and recently DNN [24]. Focusing on NMF in this study, our work is most closely related to [38] and [36] as both of them use NMF within the LGM to constrain the source spectra in each EM iteration. However, our work is different from [38] in the sense that we use the pre-trained GSSM, so that potentially the algorithm is less sensitive to the parameter initialization, and it does not suffer from the well-known permutation problem. Our work is also different from [36] as we exploit the mixed group sparsity constraint to guide the optimization, which allows the algorithm to automatically select the most representative spectral components in the GSSM. In addition, instead of constraining the variances of each source by NMF as done in [36], [38], we propose to constrain the total variances of all sources altogether by NMF and show that this novel optimization criterion offers better source separation performance.While part of the work was presented in [43], this paper provides more details regarding the algorithm derivation and the parameter settings. Furthermore, the source separation performance analysis and the comparison with existing approaches are extended.

The rest of the paper is organized as follows. We discuss the problem formulation and the background in Section II. We present the proposed GSSM-based multichannel source separation approach in Section III. In this section, we first present two ways of constructing the GSSM based on NMF. Then, to constrain the intermediate source variance estimates, two optimization criteria are introduced, which can be seen as either performing source variance denoising or source variance separation. The generalized EM algorithm is derived for the parameter estimation. We finally validate the effectiveness of the proposed approach in speech enhancement scenario using a benchmark dataset from the 2016 Signal Separation Evaluation Campaign (SiSEC 2016) in Section IV. For this purpose, we first analyze the convergence of the derived algorithm and investigate its sensitivity to the parameter settings in terms of source separation performance. We then show that the proposed algorithm outperforms most state-of-the-art methods in terms of the energy-based criteria.

II. PROBLEM FORMULATION AND MODELING

In this section, we review the formulation and the Gaussian modeling framework for multichannel audio source separation.

Let us formulate the problem in a general setting, where J sources are observed by an array of I microphones. The contribution of each source, indexed by j, to the microphone array is denoted by a vector $\mathbf{c}_j(t) \in \mathbb{R}^{I\times 1}$ and the I-channel mixture signal is the sum of all source images as

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t). \tag{1}$$

The objective of source separation is to estimate the source images $\mathbf{c}_j(t)$ given $\mathbf{x}(t)$. As the considered algorithm operates in the frequency domain, we denote by $\mathbf{c}_j(n,f)$ and $\mathbf{x}(n,f)$ the complex-valued short-term Fourier transforms (STFT) of $\mathbf{c}_j(t)$ and $\mathbf{x}(t)$, respectively, where $n = 1,2,..,N$ is time frame index and $f = 1,2,...,F$ the frequency bin index. Equation (1) can be written in the frequency domain as

$$\mathbf{x}(n,f) = \sum_{j=1}^{J} \mathbf{c}_j(n,f). \tag{2}$$

A. Local Gaussian model

We consider the existing nonstationary LGM as it has been known to be robust in modeling reverberant mixing conditions and flexible in handling prior information [34], [37]. In this framework, $\mathbf{c}_j(n,f)$ is modeled as a zero-mean complex Gaussian random vector with covariance matrix $\mathbf{\Sigma}_j(n,f) = \mathbb{E}(\mathbf{c}_j(n,f)\mathbf{c}_j^H(n,f))$:

$$\mathbf{c}_j(n,f) \sim \mathcal{N}_c(\mathbf{0}, \mathbf{\Sigma}_j(n,f)), \tag{3}$$

where $\mathbf{0}$ is an $I \times 1$ vector of zeros and ^{H} indicates the conjugate transposition. Furthermore, the covariance matrix is factorized as

$$\mathbf{\Sigma}_j(n,f) = v_j(n,f)\,\mathbf{R}_j(f), \tag{4}$$

where $v_j(n,f)$ are scalar time-dependent *variances* encoding the spectro-temporal power of the sources and $\mathbf{R}_j(f)$ are time-independent $I \times I$ *spatial covariance matrices* encoding their spatial characteristics when sources and microphones are assumed to be static. Under the assumption that the source images are statistically independent, the mixture vector $\mathbf{x}(n,f)$ also follows a zero-mean multivariate complex Gaussian distribution with the covariance matrix computed as

$$\mathbf{\Sigma}_\mathbf{x}(n,f) = \sum_{j=1}^{J} v_j(n,f)\,\mathbf{R}_j(f). \tag{5}$$

Assuming that the mixture STFT coefficients at all time-frequency (T-F) bins are independent, the likelihood of the set of observed mixture vectors $\mathbf{x} = \{\mathbf{x}(n,f)\}_{n,f}$ given the set of variance and spatial covariance parameters $\theta = \{v_j(n,f), \mathbf{R}_j(f)\}_{j,n,f}$ is given by

$$P(\mathbf{x}|\theta) = \prod_{n,f} \frac{1}{\det\left(\pi \mathbf{\Sigma}_\mathbf{x}(n,f)\right)} e^{-\mathrm{tr}(\mathbf{\Sigma}_\mathbf{x}^{-1}(n,f)\widehat{\mathbf{\Psi}}_\mathbf{x}(n,f))}, \tag{6}$$

where det represents determinant of a matrix, tr() stands for matrix trace, and $\widehat{\mathbf{\Psi}}_\mathbf{x}(n,f) = \mathbb{E}(\mathbf{x}(n,f)\mathbf{x}^H(n,f))$ is the empirical covariance matrix. It can be numerically computed

by local averaging over neighborhood of each T-F bin (n', f') as [36], [44]:

$$\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f) = \sum_{n',f'} w_{nf}^2(n',f')\mathbf{x}(n',f')\mathbf{x}^H(n',f'), \quad (7)$$

where w_{nf} is a bi-dimensional window specifying the shape of the neighborhood such that $\sum_{n',f'} w_{nf}^2(n',f') = 1$. We use Hanning window in our implementation. The quadratic T-F presentation as $\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f)$ aims to improve the robustness of the parameter estimation as it exploits the observed data in several T-F points instead of a single one. The negative log-likelihood derived from (6) is

$$\mathcal{L}(\theta) = \sum_{n,f} \mathrm{tr}\big(\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f)\big) + \log\det\big(\pi\mathbf{\Sigma}_{\mathbf{x}}(n,f)\big), \quad (8)$$

Under this model, once the parameters θ are estimated, the STFT coefficients of the source images are obtained in the minimum mean square error (MMSE) sense by multichannel Wiener filtering as

$$\hat{\mathbf{c}}_j(n,f) = v_j(n,f)\,\mathbf{R}_j(f)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f). \quad (9)$$

Finally, the expected time-domain source images $\hat{\mathbf{c}}_j(t)$ are obtained by the inverse STFT of $\hat{\mathbf{c}}_j(n,f)$.

B. NMF-based source variance model

NMF has been a well-known technique for latent matrix factorization [19] and shown to be powerful in modeling audio spectra [6], [20]. It has been widely applied to single channel audio source separation where the mixture spectrogram is usually factorized into two latent matrices characterizing the spectral basis and the time activation [20]. When adapting NMF to the considered LGM summarized in Section II-A, the nonnegative source variances $v_j(n,f)$ can be approximated as

$$v_j(n,f) = \sum_{k=1}^{K_j} w_{jfk}h_{jkn}, \quad (10)$$

where w_{jfk} is an entry of the spectral basis matrix $\mathbf{W}_j \in \mathbb{R}_+^{F\times K_j}$, h_{jkn} is an entry of the activation matrix $\mathbf{H}_j \in \mathbb{R}_+^{K_j\times N}$, and K_j the number of latent components in the NMF model.

To our best knowledge, this NMF formulation for the source variances within the LGM was first presented in [38], and then further discussed in [36], [37]. However, in those works, the basis matrix $\mathbf{W}_j$ is not a GSSM as proposed in this article (presented in Section III-A), and thus the parameters $\{\mathbf{W}_j, \mathbf{H}_j\}$ were estimated differently.

C. Estimation of the model parameters

The set of parameters θ is estimated by minimizing the criterion (8) using a generalized EM algorithm (GEM) [45]. This algorithm consists in alternating between E step and M step. In the E step, given the observed empirical covariance matrix $\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f)$ and the current estimate of θ, the conditional expectation of the natural statistics is computed as [31]

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_j(n,f) &= \mathbf{G}_j(n,f)\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f)\mathbf{G}_j^H(n,f) + \\ &\quad (\mathbf{I} - \mathbf{G}_j(n,f))\mathbf{\Sigma}_j(n,f), \end{aligned} \quad (11)$$

where $\mathbf{G}_j(n,f) = \mathbf{\Sigma}_j(n,f)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)$ is the Wiener gain, $\mathbf{I}$ is an $I\times I$ identity matrix. Then in the M step, given $\widehat{\mathbf{\Sigma}}_j(n,f)$ the parameters $\theta_j = \{v_j(n,f), \mathbf{R}_j(f)\}_{n,f}$ associated to each j-th source are updated in the maximum likelihood sense by optimizing the following criterion [34]:

$$\mathcal{L}(\theta_j) = \sum_{n,f} \mathrm{tr}\big(\mathbf{\Sigma}_j^{-1}(n,f)\widehat{\mathbf{\Sigma}}_j(n,f)\big) + \log\det\big(\pi\mathbf{\Sigma}_j(n,f)\big). \quad (12)$$

By computing the derivatives of $\mathcal{L}(\theta_j)$ with respect to $v_j(n,f)$ and each entry of $\mathbf{R}_j(f)$ and equating them to zero, the iterative updates for these parameters are found as

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_j(n,f)}\widehat{\mathbf{\Sigma}}_j(n,f) \quad (13)$$

$$v_j(n,f) = \frac{1}{I}\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{\Sigma}}_j(n,f)) \quad (14)$$

At each EM iteration, once $v_j(n,f)$ is updated in the M step by (14), it will be further constrained by NMF as (10). For this purpose, given the matrix of the current source variance estimate $\mathbf{V}_j \in \mathbb{R}_+^{F\times N}$ whose entries are $v_j(n,f)$, the corresponding NMF parameters are estimated by minimizing the Itakura-Saito divergence, which offers scale-invariant property, as

$$\min_{\mathbf{H}_j\geq 0, \mathbf{W}_j\geq 0} D(\mathbf{V}_j\|\mathbf{W}_j\mathbf{H}_j), \quad (15)$$

where $D(\mathbf{V}_j\|\mathbf{W}_j\mathbf{H}_j) = \sum_{n=1}^{N}\sum_{f=1}^{F} d_{IS}\big(v_j(n,f)\|w_{jfk}h_{jkn}\big)$, and

$$d_{IS}(x\|y) = \frac{x}{y} - \log\left(\frac{x}{y}\right) - 1. \quad (16)$$

The parameters $\{\mathbf{W}_j, \mathbf{H}_j\}$ are usually initialized with random non-negative values and are iteratively updated via the well-known multiplicative update (MU) rules [19], [20].

III. PROPOSED GSSM-BASED MULTICHANNEL APPROACH

The global workflow of the proposed approach is depicted in Fig. 1. In the following, we will first review a training phase for the GSSM construction based on NMF in Section III-A. We then propose the NMF-based source variance model fitting with sparsity constraint in Section III-B. Finally, we derive the generalized EM algorithm for the parameter estimation in Section III-C. Note that we focus on NMF as the spectral model in this paper, however, the whole idea of the proposed approach can potentially be used for other spectral models than NMF such as GMM or DNN.

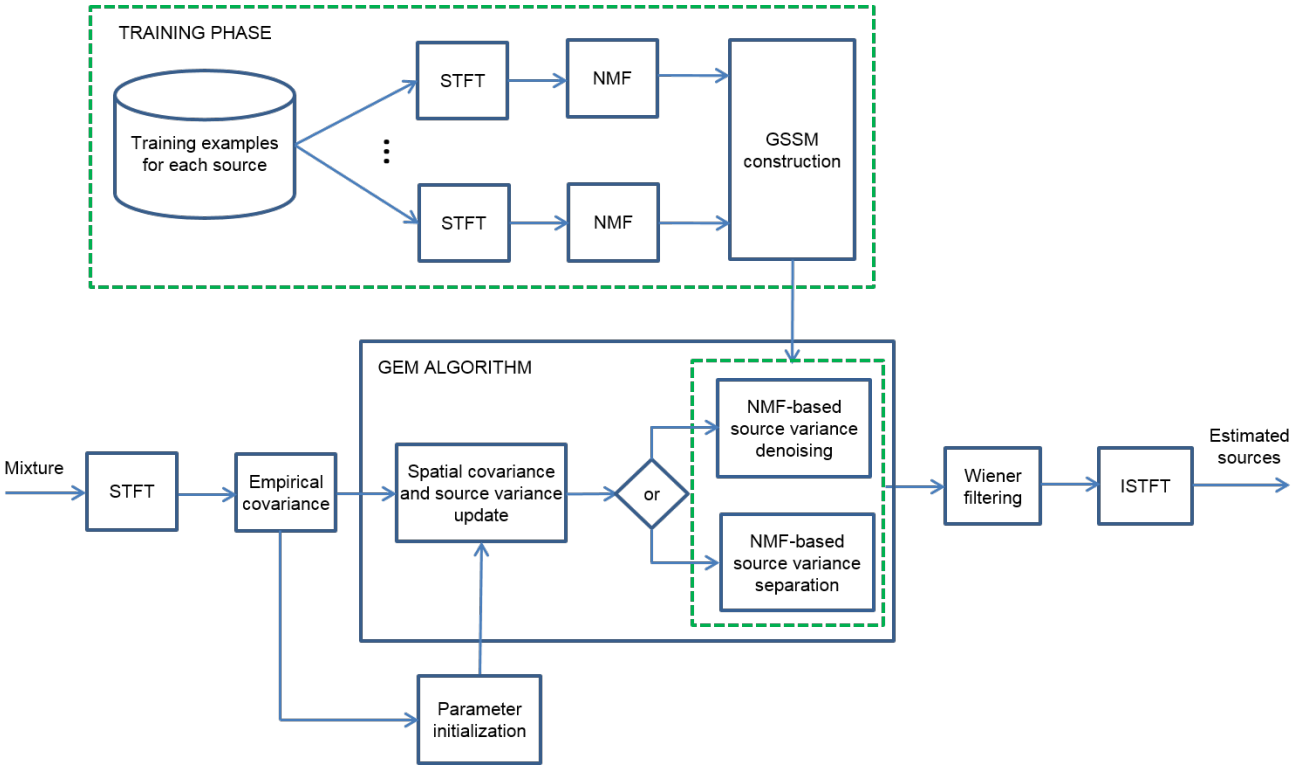


Fig. 1. General workflow of the proposed source separation approach. Top green dashed box decribles training phase for the GSSM construction. Bottom blue boxes indicate processing steps for source separation. Green dashed boxes indicate the novelty compared to the existing works [36]–[38].

A. GSSM construction

In this section, we review the GSSM construction, which was introduced in [14], [17]. We assume that the types of sources in the mixture are known and some recorded examples of such sounds are available. This is actually feasible in practice. For instance, in the speech enhancement, one target source is speech and another is noise and one can easily find speech and noise recordings. We need several examples for each type of source as one recording is usually not fully representative of the others and a source like "noise" is poorly defined. Let us denote by $s_j^l(t)$ a l-th single-channel learning example of j-th source and its corresponding spectrogram obtained by STFT $\mathbf{S}_j^l$. First, $\mathbf{S}_j^l$ is used to learn the corresponding NMF spectral dictionary, denoted by $\mathbf{W}_j^l$, by optimizing the similar criterion as (15):

$$\underset{\mathbf{H}_j^l \geq 0, \mathbf{W}_j^l \geq 0}{min} D(\mathbf{S}_j^l \| \mathbf{W}_j^l \mathbf{H}_j^l) \tag{17}$$

where $\mathbf{H}_j^l$ is the time activation matrix. Given $\mathbf{W}_j^l$ for all examples $l = 1, ..., L_j$ of the j-th source, the GSSM for the j-t source is constructed as

$$\mathbf{U}_j = [\mathbf{W}_j^1, \ldots, \mathbf{W}_j^{L_j}], \tag{18}$$

then the GSSM for all the sources is computed by

$$\mathbf{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_J]. \tag{19}$$

As an example for speech and noise separation, in the practical implementation, we may need several speech examples for different male voices and female voices (*e.g.,* 5 examples in total), and examples of different types of noise such as those from outdoor environment, cafeteria, waterfall, street, etc.,. (*e.g.,* 6 examples in total).

Note that as another variant investigated in this work, the GSSM $\mathbf{U}_j$ can be constructed differently by first concatenating all examples for each source ($\mathbf{S}_j = [\mathbf{S}_j^1, \ldots, \mathbf{S}_j^{L_j}]$), and then performing NMF on the concatenated spectrogram only once by optimizing the criterion

$$\underset{\mathbf{H}_j \geq 0, \mathbf{U}_j \geq 0}{min} D(\mathbf{S}_j \| \mathbf{U}_j \mathbf{H}_j). \tag{20}$$

We will show in the experiment that this way of constructing the GSSM does not provide as good source separation performance as the one presented before by (18).

B. *Proposed* source variance fitting with GSSM and mixed group sparsity constraint

As the GSSM is constructed to guide the NMF-based source variance constraint, we propose two fitting strategies as follows:

1) Source variance denoising: Motivated by the source variance model (10), when exploiting the GSSM model we propose a variant as

$$v_j(n,f) = \sum_{k=1}^{P_j} u_{jfk} \tilde{h}_{jkn}, \tag{21}$$

where u_{jfk} is an entry of $\mathbf{U}_j$, $\tilde{h}_{jkn}$ is an entry of the corresponding activation matrix $\tilde{\mathbf{H}}_j \in \mathbb{R}_+^{P_j \times N}$. This leads to

a straightforward extension of the conventional optimization criterion described by (15) where $\widetilde{\mathbf{H}}_j$ is now estimated by optimizing the criterion:

$$\min_{\widetilde{\mathbf{H}}_j \geq 0} D(\mathbf{V}_j \| \mathbf{U}_j \widetilde{\mathbf{H}}_j) + \lambda \Omega(\widetilde{\mathbf{H}}_j), \tag{22}$$

where $\mathbf{U}_j$ is constructed by (18) or (20) and fixed, $\Omega(\widetilde{\mathbf{H}}_j)$ presents a penalty function imposing sparsity on $\widetilde{\mathbf{H}}_j$, and λ is a trade-off parameter determining the contribution of the penalty. Note that as the GSSM $\mathbf{U}_j$ constructed in (18) becomes a large matrix when the number of examples L_j for each source increases, and it is actually a redundant dictionary since different examples may share similar spectral patterns. Thus to fit the source variances with the GSSM, sparsity constraint is naturally needed in order to activate only a subset of $\mathbf{U}_j$ which represents the spectral characteristics of the sources in the mixture [46]–[48].

2) Source variance separation: We propose another source variance model as

$$v(n,f) = \sum_{k=1}^{K} u_{fk} \tilde{h}_{kn}, \tag{23}$$

where $v(n,f) = \sum_{j=1}^{J} v_j(n,f)$, u_{fk} is an entry of the GSSM model $\mathbf{U}$ constructed as (19) and fixed, $K = \sum_{j=1}^{J} P_j$. Under this model, let $\widetilde{\mathbf{V}} = \sum_{j=1}^{J} \mathbf{V}_j$ be the matrix of the total source variance estimate, it is then decomposed by solving the following optimization problem

$$\min_{\widetilde{\mathbf{H}} \geq 0} D(\widetilde{\mathbf{V}} \| \mathbf{U}\widetilde{\mathbf{H}}) + \lambda \Omega(\widetilde{\mathbf{H}}) \tag{24}$$

where $\Omega(\widetilde{\mathbf{H}})$ presents a penalty function imposing sparsity on the activation matrix $\widetilde{\mathbf{H}} = [\widetilde{\mathbf{H}}_1^\top, ..., \widetilde{\mathbf{H}}_J^\top]^\top \in \mathbb{R}_+^{K \times N}$ the total number of rows in $\widetilde{\mathbf{H}}$. This criterion can be seen as an additional NMF-based separation step applied on the source variances, while criterion (22) and other existing works [36]–[38] do not perform any additional separation of the variances, but more like denoising of the already separated variances. For the sake of simplicity, in the following, we only present the algorithm derivation for the criterion (24), but a strong synergy can be found for the criterion (22).

Recent works in audio source separation have considered two penalty functions, namely *block* sparsity-inducing penalty [14] and *component* sparsity-inducing penalty [17]. The former one enforces the activation of *relevant examples* only while omitting irrelevant ones since their corresponding activation block in $\widetilde{\mathbf{H}}$ will likely converge to zero. The latter one, on the other hand, enforces the activation of *relevant components* in $\mathbf{U}$ only. It is motivated by the fact that only a part of the spectral model learned from an example may fit well with the targeted source in the mixture, while the remaining components in the model do not. Thus instead of activating the whole block, the component sparsity-inducing penalty allows selecting only the more likely relevant spectral components from $\mathbf{U}$. Inspired by the advantage of these penalty functions, in our recent work we proposed to combine them in a more general form as [18]

$$\Omega(\widetilde{\mathbf{H}}) = \gamma \sum_{p=1}^{P} \log(\epsilon + \|\mathbf{H}_p\|_1) + (1-\gamma) \sum_{k=1}^{K} \log(\epsilon + \|\mathbf{h}_k\|_1), \tag{25}$$

where the first term on the right hand side of the equation presents the block sparsity-inducing penalty, the second term presents the component sparsity-inducing penalty, and $\gamma \in [0,1]$ weights the contribution of each term. In (25), $\mathbf{h}_k \in \mathbb{R}_+^{1 \times N}$ is a row (or component) of $\widetilde{\mathbf{H}}$, $\mathbf{H}_p$ is a subset of $\widetilde{\mathbf{H}}$ representing the activation coefficients for p-th block, P is the total number of blocks, ϵ is a non-zero constant, and $\|.\|_1$ denotes ℓ_1-norm operator. In the considered setting, a block represents one training example for a source and P is the total number of used examples (*i.e.,* $P = \sum_{j=1}^{J} L_j$).

By putting (25) into (24), we now have a complete criterion for estimating the activation matrix $\widetilde{\mathbf{H}}$ given $\widetilde{\mathbf{V}}$ and the pre-trained spectral model $\mathbf{U}$. The derivation of MU rule for updating $\widetilde{\mathbf{H}}$ is presented in the Appendix.

C. Proposed multichannel algorithm

Within the LGM, a generalized EM algorithm used to estimate the parameters $\{v_j(n,f), \mathbf{R}_j(f)\}_{j,n,f}$ by considering the set of hidden STFT coeffients of all the source images $\{\mathbf{c}_j(n,f)\}_{n,f}$ as the *complete data*. The overview for the GEM derivation are presented in Section II-C, and more details can be found in [34], [37].

For the proposed approach as far as the GSSM concerned, the E-step of the algorithm remains the same as in [34]. In the M-step, we additionally perform the optimization defined either by (22) (for source variance denoising) or by (24) (for source variance separation). This is done by the MU rules so that the estimated intermediate source variances $v_j(n,f)$ are further updated with the supervision of the GSSM. The detail of overall proposed algorithm with source variance separation is summarized in Algorithm 1.

Note that this generalized EM algorithm requires the same order of computation compared to the existing method [37], [38] as sparsity constraint and bigger GSSM size does not significantly affect the overall computational time. As an example, for separating a 10-second long mixture presented in our experiment, both [38] and our proposed method (when non-optimally implemented in Matlab) take about 400 seconds when running in a laptop with Intel Core i5 Processor, 2.2 GHz, and 8 GB RAM.

IV. EXPERIMENTS

A. Dataset and parameter settings

We validated the performance of the proposed approach in an important speech enhancement use case where we know already two types of sources in the mixture: speech and noise. For a better comparison with the state of the art, we used the benchmark development dataset of the "Two-channel mixtures of speech and real-world background noise" (BGN) task[1] within the SiSEC 2016 [4]. This devset contains stereo

[1]https://sisec.inria.fr/sisec-2016/bgn-2016/

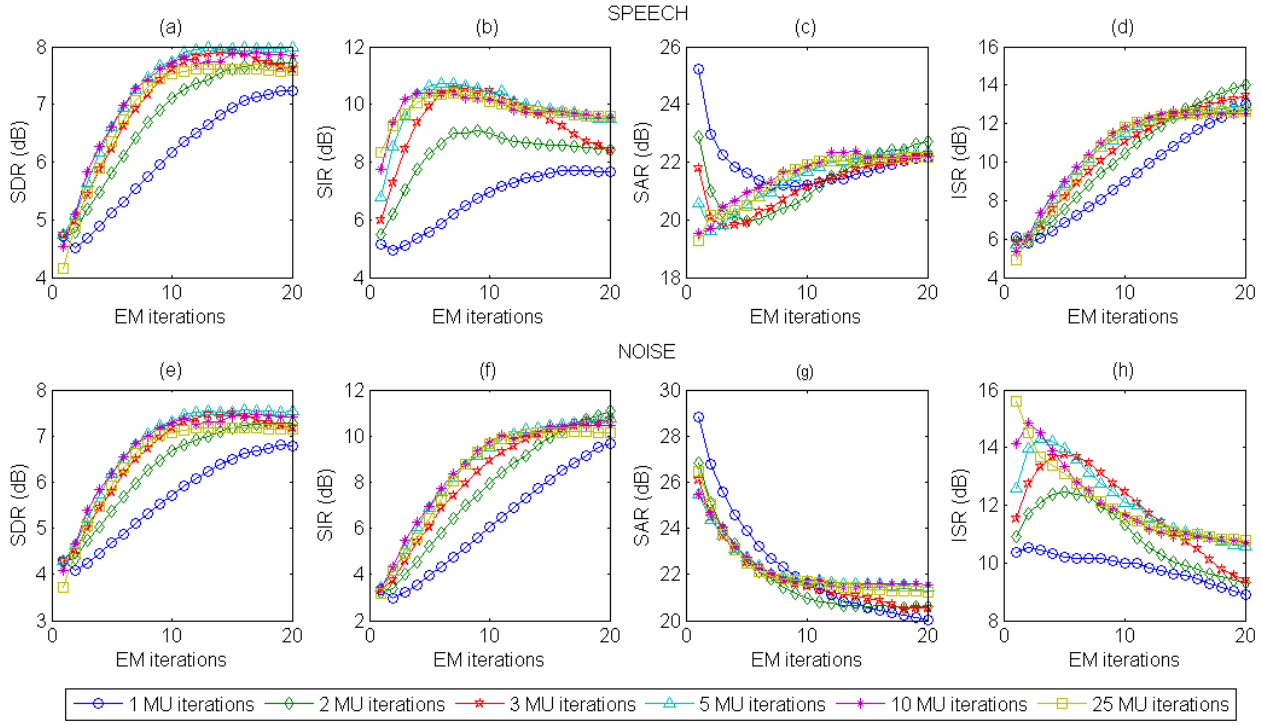


Fig. 2. Average separation performance obtained by the proposed method over stereo mixtures of speech and noise as functions of EM and MU iterations. (a): speech SDR, (b): speech SIR, (c): speech SAR, (d): speech ISR, (e): noise SDR, (f): noise SIR, (g): noise SAR, (h): noise ISR

mixtures of 10-second duration and 16 kHz sampling rate. They were the mixture of male/female speeches and real-world noises recorded from different public environments: cafeteria (Ca), square (Sq), and subway (Su). Overall there were nine mixtures: three with Ca noise, four with Sq noise, and two with Su noise. The signal-to-noise ratio was drawn randomly between -17 and +12 dB by the dataset creators.

Our works in single-channel case [16], [18] and preliminary tests on multichannel case show that only a few examples for each source could be enough to train an efficient GSSM. Thus, for training the generic speech spectral model, we took only one male voice and two female voices from the SiSEC 2015[2]. These three speech examples are also 10-second length. We performed the listening check to confirm that these examples used for the speech and noise model training are different from those in the devset, which were used for testing. For training the generic noise spectral model, we extracted five noise examples from the Diverse Environments Multichannel Acoustic Noise Database (DEMAND)[3]. Again they were 10-second length and contained three types of environmental noise: cafeteria, square, metro. The STFT window length was 1024 for all train and test files. The number of NMF components in $\mathbf{W}_j^l$ for each speech example was set to 32, while that for noise example was 16. These values were found to be reasonable in [15] and our work on single-channel case [18]. Each $\mathbf{W}_j^l$ were obtained by optimizing (17) with 20 MU iterations.

Initialization of the spatial covariance matrices: As suggested in [34], we firstly tried to initialize the spatial covariance matrix $\mathbf{R}_j(f)$ by performing hierarchical clustering on the mixture STFT coefficients $\mathbf{x}(n,f)$. But this strategy did not give us a good separation performance as the noise source in the considered mixtures is diffuse (*i.e.,* it does not come from a single direction). Thus we initialized the noise spatial covariance matrix based on the diffuse model where noise is assumed to come uniformly from all spatial directions. With this assumption, the diagonal entries of the noise spatial covariance matrix are one and the off-diagonal entries are real-valued computed as in [49]

$$r_{1,2}(f) = r_{2,1}(f) = \frac{\sin(2\pi f d/v)}{2\pi f d/v}, \tag{26}$$

where d is the distance between two microphones and $v = 334$ m/s the sound velocity. The spatial covariance matrix for the speech source was initialized by the full-rank direct+diffuse model detailed in [34] where the speech's direction of arrival (DoA) was set to 90 degrees. This DoA initialization was chosen for balancing the fact that the speech direction can vary between 0 degree and 180 degrees in each mixture and we did not have access to the ground truth information while performing the test.

The source separation performance for all approaches was evaluated by two sets of criteria. The four power-based criteria: the signal to distortion ratio (SDR), the signal to interference ratio (SIR), the signal to artifacts ratio (SAR), and the source image to spatial distortion ratio (ISR), measured in dB where

[2]https://sisec.inria.fr/sisec-2015/2015-underdetermined-speech-and-music-mixtures/.

[3]http://parole.loria.fr/DEMAND/.

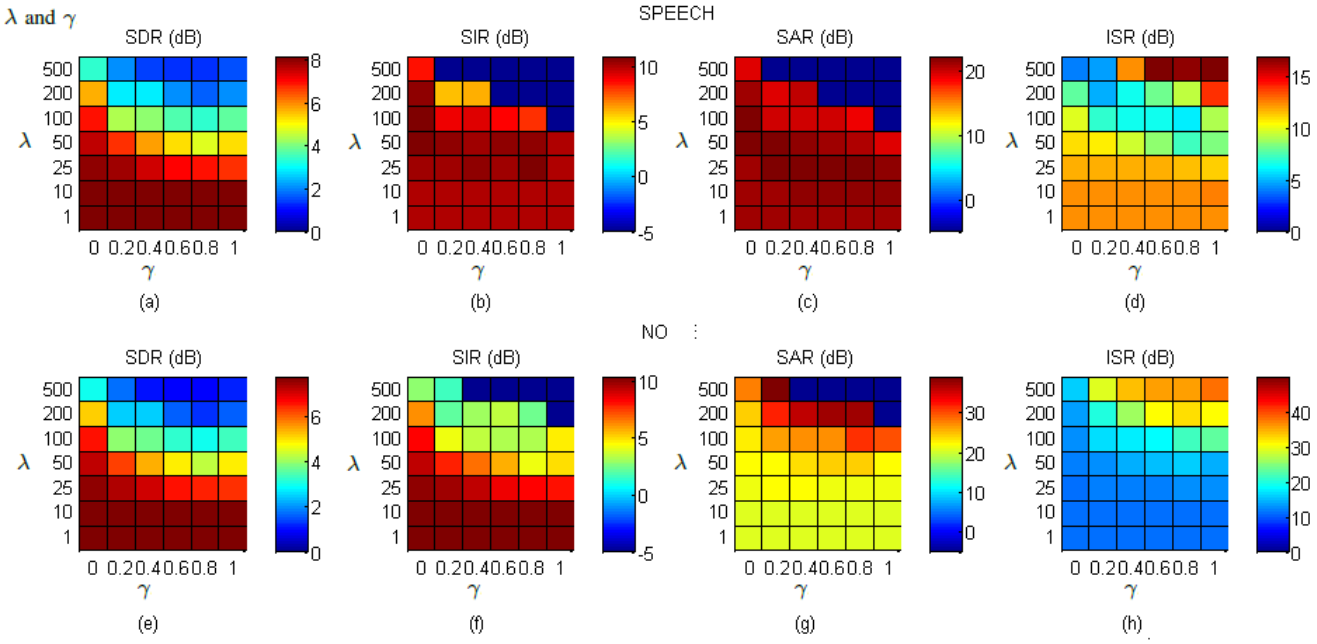


Fig. 3. Average separation performance obtained by the proposed method over stereo mixtures of speech and noise as functions of λ and γ. (a): speech SDR, (b): speech SIR, (c): speech SAR, (d): speech ISR, (e): noise SDR, (f): noise SIR, (g): noise SAR, (h): noise ISR

the higher the better [50]. The four perceptually-motivated criteria: the overall perceptual score (OPS), the target-related perceptual score (TPS), the artifact-related perceptual score (APS), and the interference-related perceptual score (IPS) [51], where a higher score is better. As power-based criteria are more widely used in source separation community, the hyper-parameters for each algorithm were chosen in order to maximize the SDR - the most important metric as it reflects the overall signal distortion.

B. Algorithm analysis

1) Algorithm convergence: separation results as functions of EM and MU iterations: We first investigate the convergence in term of separation performance of the derived Algorithm 1 by varying the number of EM and MU iterations and computing the separation results obtained on the benchmark BGN dataset. In this experiment, we set $\lambda = 10$ and $\gamma = 0.2$ as we will show in next section that these values offer both the stability and the good separation performance. The speech and noise separation results, measured by the SDR, SIR, SAR, and ISR, averaged over all mixtures in the dataset, illustrated as functions of the EM and MU iterations, are shown in Fig. 2.

As it can be seen, generally the SDR increases when the number of EM and MU iterations increases. With 10 or 25 MU iterations, the algorithm converges nicely and saturates after about 10 EM iterations. The best separation performance was obtained with 10 MU iterations and 15 EM iterations. It is also interesting to see that with a small number of MU iterations like 1, 2, or 3, the separation results are quite poor and the algorithm is less stable as it varies significantly even with a large number of EM iterations. This reveals the effectiveness of the proposed NMF constraint (24).

2) Separation results with different choices of λ and γ: We further investigate the sensitivity of the proposed algorithm to two parameters λ and γ, which determine the contribution of sparsity penalty to the NMF constraint in (24). For this purpose, we varied the values of these parameters, $\lambda = \{1, 10, 25, 50, 100, 200, 500\}$, $\gamma = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, and applied the corresponding source separation algorithm presented in the Algorithm 1 on the benchmark BGN dataset. The number of EM and MU iterations are set to 15 and 10, respectively, as these values guarantee the algorithm's convergence shown in Fig. 2. The speech and noise separation results, measured by the SDR, SIR, SAR, and ISR, averaged over all mixtures in the dataset, represented as functions of λ and γ, are shown in Fig. 3.

It can be seen that the proposed algorithm is less sensitive to the choice of γ, while more sensitive to the choice of λ, and the separation performance greatly decreases with $\lambda > 10$. The best choice for these parameters in term of the SDR are $\lambda = 10, \gamma = 0.2$. With the small value of λ (*e.g.,* $\lambda = 1$), varying γ does not really affect the separation performance as the evaluation criteria are quite stable. We noted that with $\gamma = 0.2$, the algorithm offers 0.2 dB and 1.0 dB SDR, which are higher than when $\gamma = 0$ and $\gamma = 1$, respectively. This confirms the effectiveness of the mixed sparsity penalty (25) in the multichannel setting.

C. Comparison with the state of the art

We compare the speech separation performance obtained on the BGN dataset of the proposed approach with its close prior art (*i.e.* Arberet's algorithm [38]) and other state-of-the-art methods presented at the SiSEC campaign over different

Algorithm 1 Proposed GSSM + SV separation algorithm

Require:
Mixture signal $\mathbf{x}(t)$
List of examples of each source in the mixture $\{s_j^l(t)\}_{j=1:J,l=1:L_j}$
Hyper-parameters λ, γ, MU-iteration
Ensure: Source images $\hat{\mathbf{c}}_j(t)$ separated from $\mathbf{x}(t)$

- Compute the mixture STFT coefficients $\mathbf{x}(n,f) \in \mathbb{C}^{F\times N}$ and then $\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f) \in \mathbb{C}^{I\times I}$ by (7)
- Construct the GSSM model $\mathbf{U}_j$ by (18), then $\mathbf{U} \in \mathbb{R}_+^{F\times K}$ by (19)
- Initialize the spatial covariance matrices $\mathbf{R}_j(f), \forall j, f$ (see Section IV-A)
- Initialize the non-negative time activation matrix for each source $\widetilde{\mathbf{H}}_j$ randomly, then $\widetilde{\mathbf{H}} = [\widetilde{\mathbf{H}}_1^\top, \ldots, \widetilde{\mathbf{H}}_J^\top]^v \in \mathbb{R}_+^{K\times N}$
- Initialize the source variance $v_j(n,f) = [\mathbf{U}_j\widetilde{\mathbf{H}}_j]_{n,f}$

// Generalized EM algorithm for the parameter estimation:
repeat
// E step (perform calculation for all j, n, f):
$\mathbf{\Sigma}_j(n,f) = v_j(n,f)\mathbf{R}_j(f)$ // eq. (4)
$\mathbf{\Sigma}_{\mathbf{x}}(n,f) = \sum_{j=1}^J v_j(n,f)\,\mathbf{R}_j(f)$ // eq. (5)
$\mathbf{G}_j(n,f) = \mathbf{\Sigma}_j(n,f)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)$ // Wiener gain
$\widehat{\mathbf{\Sigma}}_j(n,f) = \mathbf{G}_j(n,f)\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f)\mathbf{G}_j^H(n,f) + (\mathbf{I} - \mathbf{G}_j(n,f))\mathbf{\Sigma}_j(n,f)$ // eq. (11)

// M step: updating spatial covariance matrix and unconstrained source spectra
$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^N \frac{1}{v_j(n,f)}\widehat{\mathbf{\Sigma}}_j(n,f)$ // eq. (13)
$v_j(n,f) = \frac{1}{I}\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{\Sigma}}_j(n,f))$ // eq. (14)

$\mathbf{V}_j = \{v_j(n,f)\}_{n,f}$
$\widetilde{\mathbf{V}} = \sum_{j=1}^J \mathbf{V}_j$

// MU rules for NMF inside M step to further constrain source spectra by the GSSM
for $iter = 1, ...,$ MU-iteration **do**
for $p = 1, ..., P$ **do**
$\mathbf{Y}_p \leftarrow \frac{1}{\epsilon + \|\mathbf{H}_p\|_1}$
end for
$\mathbf{Y} = [\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_P^\top]^\top$
for $k = 1, ..., K$ **do**
$\mathbf{z}_k \leftarrow \frac{1}{\epsilon + \|\mathbf{h}_k\|_1}$
end for
$\mathbf{Z} = [\mathbf{z}_1^\top, \ldots, \mathbf{z}_K^\top]^\top$
// Updating activation matrix
$\widehat{\mathbf{V}} = \mathbf{U}\widetilde{\mathbf{H}}$
$\widetilde{\mathbf{H}} \leftarrow \widetilde{\mathbf{H}} \odot \left(\frac{\mathbf{U}^\top(\widetilde{\mathbf{V}}\odot\widehat{\mathbf{V}}^{\cdot -2})}{\mathbf{U}^\top(\widehat{\mathbf{V}}^{\cdot -1}) + \lambda(\gamma\mathbf{Y} + (1-\gamma)\mathbf{Z})}\right)^{\cdot\frac{1}{2}}$ // eq. (31)
end for
$v_j(n,f) = [\mathbf{U}_j\widetilde{\mathbf{H}}_j]_{n,f}$ // updating constrained spectra

until convergence

- Source separation by multichannel Wiener filtering (9)
- Time domain source images $\hat{\mathbf{c}}_j(t)$ are obtained by the inverse STFT of $\hat{\mathbf{c}}_j(n,f)$.

years since 2013. The results of these methods were submitted by the authors and evaluated by the SiSEC organizers [4], [52], [53]. All comparing methods are summarized as follows:

- Martinez-Munoz's method (SiSEC 2013) [52]: this algorithm exploits source-filter model for the speech source and the noise source is modeled as a combination of pseudo-stationary broadband noise, impulsive noise, and pitched interferences. The parameter estimation is based on the MU rules employed in non-negative matrix factorization.
- Wang's method [54] (SiSEC 2013): this algorithm performs well-known frequency domain independent component analysis (ICA). The associated permutation problem is solved by a novel region-growing permutation alignment technique.
- Le Magoarou's method [9] (SiSEC 2013): this approach uses text transcript of the speech source in the mixture as prior information to guide the source separation process. The algorithm is based on the nonnegative matrix partial co-factorization.
- Bryan's method [55] (SiSEC 2013): this interactive approach exploits human annotation on the mixture spectrogram to guide and refine the source separation process. The modeling is based on the probabilistic latent component analysis (PLCA), which is equivalent to NMF.
- Rafii's method [56] (SiSEC 2013): this technique uses a similarity matrix to separate the repeating background from the non-repeating foreground in a mixture. The underlying assumption is that the background is dense and low-ranked, while the foreground is sparse and varied.
- Ito's method [57] (SiSEC 2015): this is a permutation-free frequency-domain blind source separation algorithm via full-band clustering of the time-frequency (T-F) components. The clustering is performed via MAP estimation of the parameters with EM algorithm.
- Liu's method [4] (SiSEC 2016): the algorithm performs Time Difference of Arrival (TDOA) clustering based on GCC-PHAT.
- Wood's method [58] (SiSEC 2016): this recently proposed algorithm first applies NMF to the magnitude spectrograms of the mixtures with channels concatenated in time. Each dictionary atom is clustered to either the speech or the noise according to its spatial origin.
- Arberet's method [38]: using the similar local Gaussian model, the algorithm further constrains the intermediate source variances by unsupervised NMF with criterion (15). Such algorithm is implemented by Ozerov *et. al.* in [37]. This method is actually the most relevant prior art to compare with as it falls in the same LGM framework.

The proposed approach with different variants are summarized as:

- GSSM + SV denoising: The proposed GSSM + full-rank spatial covariance approach where the estimated variances of each sources $\mathbf{V}_j$ are further constrained by criterion (22). We submitted results obtained by this method to the SiSEC 2016 BGN task and obtained the best performance over the actual test set in term of SDR [4].

- GSSM + SV separation: The proposed approach with source variance separation by optimizing criterion (24). In order to investigate the benefit of the sparsity constraint, we further report the results obtained by this method when $\lambda = 0$. Finally, to confirm the effectiveness of the GSSM construction by (18), we report the results obtained when the GSSM of the same size is learned jointly by concatenating all example's spectrograms $\mathbf{S}_j^l$ as (20). In this case, only the component sparsity is applied (*i.e.,* $\gamma = 0$) as block does not exist. This setting is named "GSSM'+component sparsity" in Table 1.

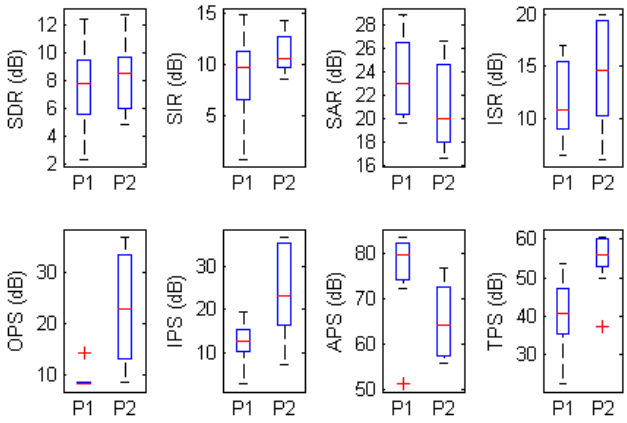


Fig. 4. Boxplot for the speech separation performance obtained by the proposed "GSSM + SV denoising" (P1) and "GSSM + SV separation" (P2) methods.

The separation results obtained by different methods for each noisy environment (Ca, Sq, Su), and the average overall mixtures are summarized in Table 1. The boxplot to illustrate the variance of the results obtained by the two proposed approaches is shown in Fig. 4. It is interesting to see that the results obtained by the proposed approach without sparsity constraint were lower than that of Arberet's method for all noisy environments, even the former used the pre-trained GSSM while the latter is completely unsupervised. It reveals that the GSSM itself is redundant and contains some irrelevant spectral patterns with the actual sources in the mixture. Thus constraining the source variances by the GSSM without a relevant spectral pattern selection guided by the sparsity penalty is even worse than the unsupervised NMF case where the spectral patterns are randomly initialized and then updated by MU rules. The importance of such sparsity penalty is explicitly confirmed by the fact that the results obtained by the proposed approach with sparsity constraint are far better than the setting without the sparsity constraint. Also, it is not surprising to see that the "GSSM + SV denoising" clearly outperforms Arberet's method (except for the ISR and the TPS) in all noisy environments as the former exploits additional information about the types of sources in the mixtures in order to learn the GSSM in advance. The "GSSM + SV separation" offers better separation performance in terms of SDR, SIR, OPS, IPS, on square and subway environments as well as on average compared to the "GSSM + SV denoising" and the "GSSM' + component sparsity". This confirms the effectiveness of the proposed source variance separation criterion (24) and the GSSM construction (18).

When compared to the top-performing state-of-the-art methods in the SiSEC campaigns, the proposed approach performs generally better in terms of the energy-based criteria but worse for the perceptually-motivated ones. Especially in Ca environment the OPS obtained by the proposed approach is far below those offered by other methods. This may be due to the fact that the hyper-parameters were optimized for the SDR, but not the OPS. The "GSSM + SV separation" with sparsity constraint outperforms all other methods, but Wang's approach, in terms of the SDR, the most important energy-based criterion, at all noisy environment. This confirms the effectiveness of the proposed approach where the GSSM is successfully exploited in the LGM framework. It should be noted that Wang's method [54] is based on the frequency-domain ICA so it is not applicable for under-determined mixtures where the number of sources is larger than the number of channels. Also, in this method, an additional post-filtering technique was applied to the separated speech source so as to maximize the denoising capability.

V. CONCLUSION

In this paper, we have presented a novel multichannel audio source separation algorithm weakly guided by some source examples. The considered approach exploits the use of generic source spectral model learned by NMF within the well-established local Gaussian model. In particular, we have proposed a new source variance separation criterion in order to better constrain the intermediate source variances estimated in each EM iteration. Experiments with the benchmark dataset from the SiSEC campaigns have confirmed the effectiveness of the proposed approach compared to the state of the art. Motivated by the effectiveness of the GSSM, future work can be devoted to extending the current approach in order to exploit in addition the use of a *generic spatial covariance model*, which remains to be defined. In addition, the theoretical grounding of the source variance separation criterion needs to be further investigated. Another promising investigation could be extending the idea of source variance separation to DNN-based models inspired by the work of Nugraha *et al.* [24].

VI. ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers of this manuscript and [43] for their helpful and constructive comments that greatly contributed to improving the quality of the paper. We also would like to thank Professor Nobutaka Ono for providing us results of the SiSEC 2013 and the SiSEC 2015.

APPENDIX
DERIVATION OF MU RULE FOR UPDATING THE ACTIVATION MATRIX IN ALGORITHM1

Let $\mathcal{L}(\widetilde{\mathbf{H}})$ denote the minimization criterion (24) with the mixed sparsity constrained $\Omega(\widetilde{\mathbf{H}})$ defined as in (25) and $D(\cdot\|\cdot)$

Methods	Ca1				Sq1				Su1				Average			
	SDR	SIR	SAR	ISR	SDR	SIR	SAR	ISR	SDR	SIR	SAR	ISR	SDR	SIR	SAR	ISR
	OPS	IPS	APS	TPS	OPS	IPS	APS	TPS	OPS	IPS	APS	TPS	OPS	IPS	APS	TPS
Martinez-Munoz*	5.4	15.4	6.1	-	9.6	17.3	10.7	-	1.5	5.8	5.8	-	6.4	14.1	7.9	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Wang* [54]	10.4	21.6	12.8	13.5	10.3	19.1	12.3	15.0	8.1	19.3	10.0	10.7	**9.8**	**20.0**	**12.0**	**13.5**
	41.9	60.4	52.9	58.8	37.4	51.8	53.3	55.7	33.1	52.4	42.3	43.7	**37.9**	**54.8**	**50.7**	**54.1**
Le Magoarou* [9]	9.2	11.6	13.4	19.8	4.0	6.2	8.3	20.4	-5.2	-4.5	2.7	9.7	3.7	5.6	8.8	17.8
	31.3	29.3	52.8	77.9	38.9	45.2	39.9	75.6	22.9	24.6	34.8	28.3	32.8	35.3	43.1	65.9
Bryan* [55]	5.6	18.4	5.9	-	10.2	15.6	12.1	-	4.2	13.6	4.9	-	7.3	16.1	7.6	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Rafii* [56]	8.8	13.0	12.1	13.3	6.2	9.6	8.9	10.7	-2.7	-2.7	4.4	11.0	5.1	8.0	9.0	11.6
	29.2	27.3	58.0	56.2	34.6	38.7	55.8	60.7	23.9	21.6	51.3	50.4	30.4	31.1	55.5	56.9
Ito* [57]	7.2	25.9	7.2	-	8.9	23.7	9.1	-	4.9	15.3	5.6	-	7.4	22.6	7.7	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Liu*	-1.0	4.9	19.7	4.1	-8.5	-2.9	15.1	1.9	-12.8	-8.0	7.6	3.8	-7.0	-1.4	15.0	3.1
	9.5	16.8	77.1	42.3	14.2	18.9	70.2	38.8	21.2	15.7	60.0	49.5	14.2	17.5	70.3	42.3
Wood* [58]	3.0	9.4	5.0	3.7	1.9	2.4	4.0	7.5	0.2	-2.6	1.3	2.5	1.9	3.6	3.7	5.1
	33.7	60.7	39.0	40.5	38.6	60.5	43.3	57.6	25.9	47.6	31.7	24.4	34.1	57.7	39.3	44.5
Arberet [37], [38]	9.1	10.0	16.1	19.5	3.3	3.3	10.4	15.3	-0.2	-1.2	9.5	11.7	4.4	4.6	12.1	**15.9**
	13.3	10.9	70.4	50.5	8.3	10.5	82.3	47.5	10.2	3.7	56.6	23.4	10.4	9.1	72.6	43.2
GSSM + SV denoising	10.5	11.8	27.7	16.2	7.0	8.5	22.0	9.8	5.1	5.6	20.7	8.1	7.7	9.0	23.6	11.6
($\lambda = 10, \gamma = 0.2$)	8.4	12.7	83.0	49.9	8.5	14.7	77.6	39.0	11.3	7.8	61.8	27.6	18.1	12.5	75.9	40.1
GSSM + SV separation	7.9	10.2	20.2	11.2	-1.1	-2.6	17.6	8.0	-1.6	-3.2	20.4	7.6	1.8	1.5	19.1	8.9
(No sparsity constraint)	25.0	19.3	64.1	55.8	32.4	29.4	55.0	60.2	18.7	11.4	56.0	35.7	18.8	22.0	58.2	53.3
GSSM + SV separation	7.3	10.0	19.4	9.7	4.4	6.1	16.0	6.9	2.4	1.8	18.3	8.8	4.9	6.5	17.7	8.3
(GSSM' + component sparsity)	20.7	18.2	66.2	48.1	30.3	28.3	57.9	52.8	21.6	16.5	56.0	43.0	21.3	22.3	60.3	49.1
GSSM + SV separation	10.6	13.5	25.6	19.6	7.8	11.1	19.3	12.3	5.0	7.1	18.7	9.5	**8.1**	**11.0**	**21.3**	**14.1**
($\lambda = 10, \gamma = 0.2$)	11.4	13.0	81.6	61.0	31.6	31.4	62.0	57.4	23.7	27.8	47.3	37.6	**23.1**	**24.5**	**65.2**	**54.2**

TABLE I

SPEECH SEPARATION PERFORMANCE OBTAINED ON THE DEVSET OF THE BGN TASK OF THE SISEC CAMPAIGN. * INDICATES SUBMISSIONS BY THE AUTHORS AND "-" INDICATES MISSING INFORMATION.

being IS divergence. The partial derivative of $\mathcal{L}(\widetilde{H})$ with respect to an entry h_{kn} is

$$\nabla_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) = \sum_{f=1}^{F} u_{fk}\left(\frac{1}{[\mathbf{U}\widetilde{\mathbf{H}}]_{n,f}} - \frac{v(n,f)}{[\mathbf{U}\widetilde{\mathbf{H}}]^2_{n,f}}\right) + \frac{\lambda.\gamma}{\epsilon + \|\mathbf{H}_p\|_1} + \frac{\lambda.(1-\gamma)}{\epsilon + \|\mathbf{h}_k\|_1} \quad (27)$$

This $\nabla_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}})$ can be written as a sum of two nonnegative parts, denoted by $\nabla^+_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) \geq 0$ and $\nabla^-_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) \geq 0$, respectively, as

$$\nabla_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) = \nabla^+_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) - \nabla^-_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) \quad (28)$$

with

$$\begin{aligned}\nabla^+_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) &\triangleq \sum_{f=1}^{F} u_{fk}\frac{1}{[\mathbf{U}\widetilde{\mathbf{H}}]_{n,f}} + \frac{\lambda.\gamma}{\epsilon + \|\mathbf{H}_p\|_1} + \frac{\lambda.(1-\gamma)}{\epsilon + \|\mathbf{h}_k\|_1},\\ \nabla^-_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) &\triangleq \sum_{f=1}^{F} u_{fk}\frac{v(n,f)}{[\mathbf{U}\widetilde{\mathbf{H}}]^2_{n,f}}. \end{aligned} \quad (29)$$

Following a standard approach for MU rule derivation [19], [20]), h_{kn} is updated as

$$h_{kn} \leftarrow h_{kn}\left(\frac{\nabla^-_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}})}{\nabla^+_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}})}\right)^{.\eta}, \quad (30)$$

where $\eta = 0.5$ following the derivation in [47], [59], which was shown to produce an accelerated descent algorithm. Putting (29) into (30) and rewriting it in a matrix form, we obtain the updates of $\widetilde{\mathbf{H}}$ as

$$\widetilde{\mathbf{H}} \leftarrow \widetilde{\mathbf{H}} \odot \left(\frac{\mathbf{U}^\top(\widetilde{\mathbf{V}} \odot \widehat{\mathbf{V}}^{\cdot -2})}{\mathbf{U}^\top(\widehat{\mathbf{V}}^{\cdot -1}) + \lambda(\gamma\mathbf{Y} + (1-\gamma)\mathbf{Z})}\right)^{\cdot\frac{1}{2}}, \quad (31)$$

where $\widehat{\mathbf{V}} = \mathbf{U}\widetilde{\mathbf{H}}$, $\mathbf{Y} = [\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_P^\top]^\top$ with $\mathbf{Y}_p, p = 1, \ldots P$ an uniform matrix of the same size as $\mathbf{H}_p$ whose entries are $\frac{1}{\epsilon + \|\mathbf{H}_p\|_1}$, and $\mathbf{Z} = [\mathbf{z}_1^\top, \ldots, \mathbf{z}_K^\top]^\top$ with $\mathbf{z}_k, k = 1, \ldots K$ a uniform vector of the same size as $\mathbf{h}_k$ whose entries are $\frac{1}{\epsilon + \|\mathbf{h}_k\|_1}$.

REFERENCES

[1] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*, Springer, 2007.

[2] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[3] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The Signal Separation Campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[4] A. Liutkus, F. R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation*, 2017, pp. 323–332.

[5] A. Liutkus, J. L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *Proc. IEEE Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2013, pp. 1–4.

[6] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From Blind to Guided Audio Source Separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[7] N. Souviraà-Labastie, A. Olivero, E. Vincent, and F. Bimbot, "Multi-channel audio source separation using multiple deformed references," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, pp. 1775–1787, 2015.

[8] P. Smaragdis and G. J. Mysore, "Separation by humming: User-guided sound extraction from monophonic mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 69–72.

[9] L. L. Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, pp. 1–5, 2014.

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

[11] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 45–48.

[12] S. Ewert, B. Pardo, M. Mueller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, 2014.

[13] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Perez, and G. Richard, "Motion informed audio source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[14] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 141–145.

[15] D. E. Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly audio source separation - a novel user-friendly framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 261–272, 2017.

[16] H. T. T. Duong, Q. C. Nguyen, C. P. Nguyen, and N. Q. K. Duong, "Single-channel speaker-dependent speech enhancement exploiting generic noise model learned by non-negative matrix factorization," in *Proc. IEEE Int. Conf. on Electronics, Information, and Communications (ICEIC)*, 2016, pp. 1–4.

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

[18] H. T. T. Duong, Q. C. Nguyen, C. P. Nguyen, T. H. Tran, and N. Q. K. Duong, "Speech enhancement based on nonnegative matrix factorization with mixed group sparsity constraint," in *Proc. ACM Int. Sym. on Information and Communication Technology (SoICT)*, 2015, pp. 247–251.

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

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

[21] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[22] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2135–2139.

[23] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[24] A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 9, pp. 1652–1664, 2016.

[25] Z.-Q. Wang, J. L. Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[26] S. Michael, L. Jan, K. Ulrik, and C. Lucas, "A survey of convolutive blind source separation methods," in *Springer Handbook of Speech Processing*. Springer, 2007, pp. 1–34.

[27] S. R. A. Jourjine and O. Yılmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, June 2000, pp. 2985–2988.

[28] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[29] Z. El Chami, D. T. Pham, C. Serviere, and A. Guerin, "A new model based underdetermined source separation," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008, pp. 147–150.

[30] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, article ID 24717, 2007.

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

[32] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[33] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1622–1637, 2016.

[34] N. Q. K. Duong, E. Vincent, and R. 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.

[35] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 727–739, 2014.

[36] M. Fakhry, P. Svaizer, and M. Omologo, "Audio source separation in reverberant environments using beta-divergence based nonnegative factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, 2017.

[37] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[38] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. IEEE Int. Conf. on Information Science, Signal Processing and their Applications (ISSPA)*, 2010, pp. 1–4.

[39] N. Q. K. Duong, H. Tachibana, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "Multichannel harmonic and percussive component separation by joint modeling of spatial and spectral continuity," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 205–208.

[40] A. Liutkus, D. Fitzgerald, and Z. Rafii, "Scalable audio separation with light kernel additive modelling," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 76–80.

[41] S. Leglaive, U. Şimşekli, A. Liutkus, R. Badeau, and G. Richard, "Alpha-stable multichannel audio source separation," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[42] P. Magron, R. Badeau, and A. Liutkus, "Lévy NMF for robust nonnegative source separation," in *Proc. IEEE Int. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[43] H. T. T. Duong, N. Q. K. Duong, Q. C. Nguyen, and C. P. Nguyen, "Multichannel audio source separation exploiting NMF-based generic source spectral model in Gaussian modeling framework," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LCA/ICA)*, 2018.

[44] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010.

[45] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[46] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparse criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066 – 1074, 2007.

[47] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito non-negative matrix factorization with group sparsity," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 21–24.

[48] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition," in *Proc. Interspeech*, 2012, pp. 17–20.

[49] H. Kuttruff, *Room Acoustics*, 4th ed. New York: Spon Press, 2000.

[50] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[51] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2012.

[52] N. Ono, Z. Koldovsk, S. Miyabe, and N. Ito, "The 2013 Signal Separation Evaluation Campaign," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2013, pp. 1–6.

[53] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 Signal Separation Evaluation Campaign," in *Latent Variable Analysis and Signal Separation (LVAICA)*. Springer, 2015, vol. 9237, pp. 387–395.

[54] L. Wang, H. Ding, and F. Yin, "A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures," *Trans. Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 549–557, 2011.

[55] N. Bryan and G. Mysore, "An efficient posterior regularized latent variable model for interactive sound source separation," in *Proc. The 30th International Conference on Machine Learning (ICML)*, 2013, pp. 208–216.

[56] Z. Rafii and B. Pardo, "Online REPET-SIM for real-time speech enhancement," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 848–852.

[57] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013, pp. 3238–3242.

[58] S. U. N. Wood, J. Rouat, S. Dupont, and G. Pironkov, "Blind Speech Separation and Enhancement With GCC-NMF," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 745–755, 2017.

[59] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, Sep. 2011.

B. PAPER 2: GAUSSIAN MODELING-BASED MULTICHANNEL AUDIO SOURCE SEPARATION EXPLOITING GENERIC SOURCE SPECTRAL MODEL

Appendix C

Paper 3: Weakly Supervised Representation Learning for Audio-Visual Scene Analysis

Weakly Supervised Representation Learning for Audio-Visual Scene Analysis

Sanjeel Parekh, Slim Essid, Alexey Ozerov, *Senior Member, IEEE*, Ngoc Q.K. Duong, *Senior Member, IEEE*, Patrick Perez, and Gael Richard, *Fellow, IEEE*

***Abstract*—Audio-visual representation learning is an important task from the perspective of designing machines with the ability to understand complex events. To this end, we propose a novel multimodal framework that instantiates multiple instance learning. We show that the learnt representations are useful for performing several tasks such as event/object classification, audio event detection, audio source separation and visual object localization. The system is trained using only video-level event labels without any timing information. An important feature of our method is its capacity to learn from unsynchronized audio-visual events. We also demonstrate our framework's ability to separate out the audio source of interest through a novel use of nonnegative matrix factorization. State-of-the-art classification results are achieved on DCASE 2017 smart cars challenge data with promising generalization to diverse object types such as musical instruments. Visualizations of localized visual regions and audio segments substantiate our system's efficacy, especially when dealing with noisy situations where modality-specific cues appear asynchronously.**

***Index Terms*—Audio-visual fusion, multimodal deep learning, multiple instance learning, event classification, audio-visual localization, audio source separation**

I. Introduction

We are surrounded by events that can be perceived via distinct audio and visual cues. Be it a ringing phone or a car passing by, we instantly identify the audio-visual (AV) components that characterize these events. This remarkable ability helps us understand and interact with our environment. For building machines with such scene understanding capabilities, it is important to design algorithms for learning audio-visual representations from real-world data.

This work is a step in that direction, where we aim to learn such representations through weak supervision.

Specifically, we are interested in designing a system that simultaneously tackles multiple related scene understanding tasks which include video event classification, spatial-temporal visual object localization and corresponding audio object enhancement and temporal localization. Obtaining precisely annotated data for doing so is an expensive endeavor, made even more challenging by multimodal considerations. The annotation process is not only error prone and time consuming but also subjective to an extent. Often, event boundaries in audio, extent of video objects or even their presence is ambiguous. Thus, we opt for a weakly-supervised learning approach using data with only video-level event labels, that is labels given for whole video documents without timing information.

S. Parekh, S. Essid and G. Richard are with Telecom ParisTech, Paris, France

A. Ozerov and N. Duong are with Technicolor R&I, Cesson Sevigne, France

P. Perez is with Valeo.ai, Paris, France

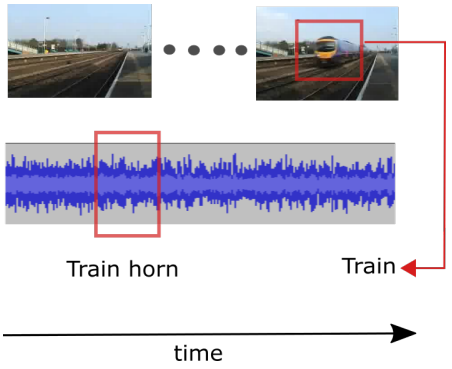


Fig. 1. **Pictorial representation of the problem**: Given a video labeled as "train horn", we would like to: (i) identify the event, (ii) localize both, its visual presence and the temporal segment(s) containing the characteristic sound, and (iii) segregate the characteristic audio cue from the background. Note that the train horn may sound before the train is visible. Our model can deal with such unsynchronized AV events.

A. Problem description

To motivate our tasks and method, consider a video labeled as "train horn", depicted in Fig. 1. Assuming that the train is both visible and audible at some time in the video, in addition to identifying the event, we are interested in learning representations that help us answer the following:

- *Where is the visual object or context that distinguishes the event?* In this case it might be the train (object) or tracks, platform (context) *etc.* We are thus aiming for their spatio-temporal localization in the image sequence.
- *When does the sound event occur?* Here it is the train horn. We thus want to temporally localize the audio event.
- *How to enhance the audio object?* Here we are interested in audio source extraction *i.e.* segregating the source of interest from the background sounds.

The variety of noisy situations that one may encounter in unconstrained environments or videos adds to the difficulty of this very challenging problem. Apart from modality-specific noise such as visual clutter, lighting variations and low audio signal-to-noise ratio, in real-world scenarios the appearance of audio and visual elements characterizing the event are often unsynchronized in time. This is to say that the train horn may sound before or after the train is visible, as in previous

example. In the extreme, not so rare case, the train may not appear at all. The latter is also commonly referred to as "off–screen" audio [1]. We are interested in designing a system to tackle the aforementioned questions and situations.

Prior research has utilized audio and visual modalities for classification and localization tasks in various contexts. Fusing modality-specific hand-crafted or deep features has been a popular approach for problems such as multimedia event detection and video concept classification [2]–[5]. On the other hand, AV correlations have been utilized for localization and representation learning in general, through feature space transformation techniques such as canonical correlation analysis (CCA) [6], [7] or deep networks [8]–[12]. However, a unified multimodal framework for our task, that is learning data representations for simultaneously identifying real world events and extracting the AV cues depicting them has not been extensively studied in previous works.

B. Contributions and outline

In this work, we present a complete AV event understanding framework where the modality-specific modules can be trained jointly to perform multiple tasks such as event/object classification, spatio-temporal visual localization, temporal audio localization and source separation. Key attributes and results of our approach are summarized below:

- We report state-of-the-art event classification performance on DCASE smart cars challenge data [13] and demonstrate usefulness of AV complementatrity. We also show results on an instrument dataset [14] to validate our framework's application to diverse object types.
- To highlight flexibility provided by our modular design, we propose several task-specific instantiations. These include changes to allow detection of synchronously appearing AV cues and capability to enhance the audio source of interest.
- Additionally, we also show encouraging qualitative visual localization results.

Paper outline. We begin by briefly mentioning connections and distinctions with related works in Section II. This is followed by a description of the proposed framework and its instantiations for tackling classification and localization in Section III. Finally, we validate the usefulness of the learnt representations for these tasks with a thorough analysis in Section IV.

II. Related work

To position our work, we briefly discuss some relevant literature that employs weakly supervised learning for visual object localization, audio event detection and source separation. We also delineate several distinctions between the present study and recent multimodal deep learning approaches.

A. Audio scene analysis

Detection and segregation of individual sources in a mixture is central to computational auditory scene analysis [15]. A significant amount of literature exists on supervised audio event detection (AED) [16]–[19]. However, progress with weakly labeled data in the audio domain has been relatively recent. An early work [20] showed the usefulness of MIL techniques to audio using SVM and neural networks.

The introduction of the weakly-labeled audio event detection task in the 2017 DCASE challenge [21][1], a challenge on DCASE, along with the release of Google's AudioSet data[2] [22], has led to accelerated progress in the recent past. AudioSet is a large-scale weakly-labeled dataset of audio events collected from YouTube videos. A subset of this data was used for the DCASE 2017 task on large-scale AED for smart cars.[3] Several submissions to the task utilized sophisticated deep architectures with attention units [23], as well as max and softmax operations [24]. Another recent study introduced a CNN with global segment-level pooling for dealing with weak labels [25]. It is worth noting that the field is growing rapidly. Concurrent and subsequent studies have greatly exploited the MIL and attention-based learning paradigm [26]–[28]. While we share with these works the high-level goal of weakly-supervised learning, apart from our multimodal design, our audio sub-module, as discussed in the next section, is significantly different.

Audio source separation research in weakly supervised regime has followed a similar trend. Recent progress includes several vision–inspired [29] and vision–guided [30]–[32] systems. Use of NMF basis vectors is particularly interesting in [32]. Our proposed separation technique goes in this direction with several key differences discussed in Sec. III-D1.

B. Visual object localization and classification

There is a long history of works in computer vision applying weakly supervised learning for object localization and classification. MIL techniques have been extensively used for this purpose [33]–[39]. Typically, each image is represented as a set of regions. Positive images contain at least one region from the reference class while negative images contain none. Latent structured output methods, *e.g.*, based on support vector machines (SVMs) [40] or conditional random fields (CRFs) [41], address this problem by alternating between object appearance model estimation and region selection. Some works have focused on better initialization and regularization strategies [39], [42], [43] for solving this non-convex optimization problem.

Owing to the exceptional success of convolutional neural networks (CNNs) in computer vision, recently, several approaches have looked to build upon CNN architectures for embedding MIL strategies. These include the introduction of operations such as max pooling over regions [35], global average pooling [38] and their soft versions [44]. Another line of research consists in CNN-based localization over class-agnostic region proposals [36], [37], [45] extracted using a state-of-the-art proposal generation algorithm such as Edge-Boxes [46], Selective Search [47], *etc.* These approaches are supported by the ability to extract fixed size feature maps from

[1] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/
[2] https://research.google.com/audioset/
[3] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-large-scale-sound-event-detection

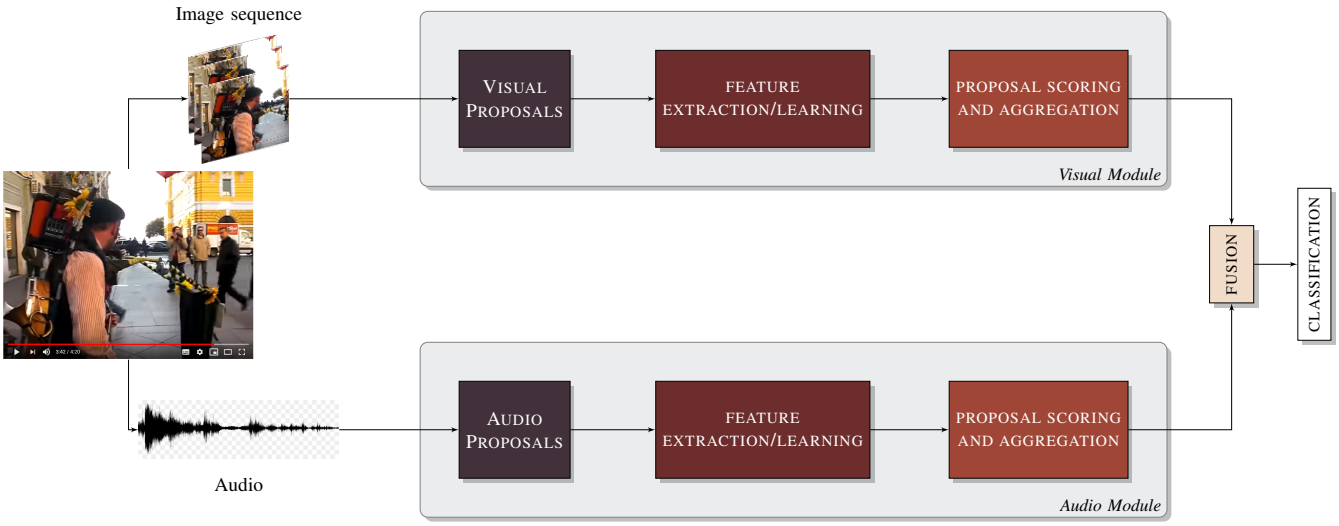


Fig. 2. **High level view of the proposed approach**: Given a video captured using a single microphone and camera, we propose the depicted framework for weakly supervised representation learning.

CNNs using region-of-interest [48] or spatial pyramid pooling [49]. Our work is related to such techniques. We build upon ideas from the two-stream architecture [37] for classification and localization.

State-of-the-art end-to-end object detection networks such as Faster RCNN [50] and its instance segmentation extension Mask RCNN [51] incorporate proposal generation as part of the system (region proposal network) instead of a separate stage. Nonetheless, these approaches require label annotations for different regions. It is also worth mentioning that some works have extended class-agnostic proposal generation from 2D images to video tube proposals for tasks such as action localization [52] and object detection [53]. However, these involve a computationally expensive pipeline preventing large-scale usage.

C. Differences with recent AV deep learning studies

We formulate the problem as a MIL task using class-agnostic proposals from both video frames and audio. This allows us to simultaneously solve the classification and localization problems. Finally, by construction, our framework deals with the difficult case of asynchronous AV events. This is significantly different from recent multimodal deep learning based studies on several counts: Contrary to prior works, where unsupervised representations are learnt through audio–image correlations (temporal co-occurrence), we adopt a weakly-supervised learning approach using event classes. Unlike [8], [9], [11], we focus on localizing discriminative audio and visual components for real-world events.

III. Proposed Framework and its Instantiations

The tasks under consideration can be naturally formulated as MIL problems [54]. MIL is typically applied to cases where labels are available over bags (sets of instances) instead of individual instances. The task then amounts to jointly selecting appropriate instances and estimating classifier parameters. In our case, a video can be seen as a labeled bag, containing a collection of visual and audio proposals. The term *proposal* refers to image or audio "parts" that may potentially constitute the object of interest. This step is at the core of our approach.

The key idea, as illustrated in Fig. 2, is to extract features from generated proposals and transform them for: (1) scoring each according to their relevance for class labels; (2) aggregating these scores in each modality and fusing them for video-level classification. This not only allows us to train both the sub-modules together through weak-supervision but also enables localization using the proposal relevance scores. Moreover, use of both the modalities with appropriate proposals makes the system robust against noisy scenarios. We present different task-specific variants of this general framework.

We now formalize the design of each building block to specifically tackle event classification, visual object and audio event localization. An overview is provided in Fig. 3. We model a video V as a bag of M selected image regions, $\mathcal{R} = \{r_1, r_2, \ldots, r_M\}$, obtained from sub-sampled frames and S audio segments, $\mathcal{A} = \{a_1, a_2, \ldots, a_S\}$. Given L such training examples, $\mathcal{V} = \{V^{(l)}\}_{l=1}^{L}$, organized into C classes, our goal is to learn a representation to jointly classify and localize image regions and audio segments that characterize a class. Each block from proposal generation to classification is discussed below in detail.

A. Generating proposals and extracting features

Visual Proposals. Generating proposals for object containing regions from images is at the heart of various visual object detection algorithms [55], [56]. As our goal is to spatially and temporally localize the most discriminative region pertaining to a class, we choose to apply this technique over sub-sampled video frame sequences. In particular, we sub-sample the extracted frame sequences of each video at a rate of 1 frame per second. This is followed by class-agnostic region proposal generation on the selected frames using EdgeBoxes

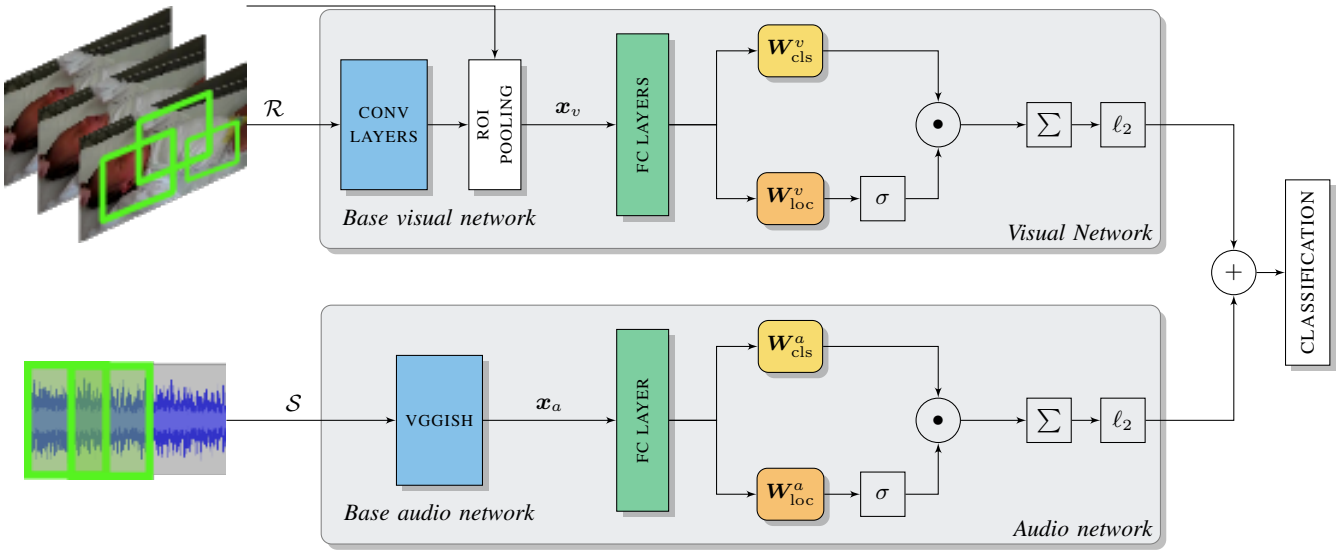


Fig. 3. **Module design**: Given a video, we consider the depicted pipeline for going from audio and visual proposals to localization and classification. Here $\boldsymbol{W}_{\text{cls}}$ and $\boldsymbol{W}_{\text{loc}}$ refer to the fully-connected classification and localization streams respectively; σ denotes softmax operation over proposals for each class, $\odot$ refers to element-wise multiplication; Σ to a summation over proposals and ℓ_2 to a normalization of scores. During training we freeze the weights of blocks denoted in blue.

[46]. This proposal generation method builds upon the insight that the number of contours entirely inside a box is indicative of the likelihood of an object's presence. Its use in our pipeline is motivated by experiments confirming better performance in terms of speed/accuracy tradeoffs over most competing techniques [57]. EdgeBoxes additionally generates a confidence score for each bounding box which reflects the box's "objectness". To reduce the computational load and redundancy, we use this score to select the top M_{img} proposals from each sampled image and use them for feature extraction. Hence, given a 10 second video, the aforementioned procedure would leave us with a list of $M = 10 \times M_{\text{img}}$ region proposals.

A fixed-length feature vector, $\boldsymbol{x}_{vis}(r_m; V) \in \mathbb{R}^{d_v}$ is obtained from each image region proposal, r_m in V, using a convolutional neural network altered with a region-of-interest (RoI) pooling layer. An RoI layer works by computing fixed size feature maps (*e.g.* 6×6 for `caffenet` [58]) from regions of an image using max-pooling [48]. This helps to ensure compatibility between convolutional and fully connected layers of a network when using regions of varying sizes. Moreover, unlike Region-based CNN (RCNN) [56], the shared computation for different regions of the same image using Fast-RCNN implementation [48] leads to faster processing. In Fig. 3 we refer to this feature extractor as the base visual network. In practice, feature vectors $\boldsymbol{x}_{vis}(\cdot)$ are extracted after RoI pooling layer and passed through two fully connected layers, which are fine-tuned during training. Typically, standard CNN architectures pre-trained on ImageNet [59] classification are used for the purpose of initializing network weights.

Audio Temporal Segment Proposals.

We first represent the raw audio waveform as a log-Mel spectrogram [60]. Each proposal is then obtained by sliding a fixed-length window over the obtained spectrogram along the temporal axis. These are the so called audio temporal segment proposals, also referred to as Temporal Segment Proposals (TSPs). The dimensions of this window are chosen to be compatible with the audio feature extractor. For our system we set the proposal window length to 960ms and stride to 480ms.

We use a VGG-style deep network known as `vggish` for base audio feature extraction. Inspired by the success of CNNs in visual object recognition Hershey *et al.* [61] introduced this state-of-the-art audio feature extractor as an audio parallel to networks pre-trained on ImageNet for classification. `vggish` has been pre-trained on a preliminary version of YouTube-8M [62] for audio classification based on video tags. It stacks 4 convolutional and 2 fully connected layers to generate a 128 dimensional embedding, $\boldsymbol{x}_{\text{aud}}(a_s; V) \in \mathbb{R}^{128}$ for each input log-Mel spectrogram segment $a_s \in \mathbb{R}^{96 \times 64}$ with 64 Mel-bands and 96 temporal frames. Prior to proposal scoring, the generated embedding is passed through a fully-connected layer that is learnt from scratch.

B. Proposal scoring network and fusion

So far, we have extracted base features for each proposal in both the modalities and passed them through fully connected layers in their respective modules. Equipped with this transformed representation of each proposal, we use the two-stream architecture proposed by Bilen *et al.* [37] for scoring each of them with respect to the classes. There is one scoring network of the same architecture for each modality as depicted in Fig. 3. Thus, for notational convenience, we generically denote the set of audio or visual proposals for each video by $\mathcal{P}$ and let proposal representations before the scoring network be stacked in a matrix $\mathbf{Z} \in \mathbb{R}^{|\mathcal{P}| \times d}$, where d denotes the dimensionality of the audio/visual proposal representation.

The architecture of this module consists of parallel classification and localization streams. The former classifies each

region by passing $\mathbf{Z}$ through a linear fully connected layer with weights $\boldsymbol{W}_{\text{cls}}$, giving a matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{P}|\times C}$. On the other hand, the localization layer passes the same input through another fully-connected layer with weights $\boldsymbol{W}_{\text{loc}}$. This is followed by a softmax operation over the resulting matrix $\mathbf{B} \in \mathbb{R}^{|\mathcal{P}|\times C}$ in the localization stream. The softmax operation on each element of $\mathbf{B}$ can be written as:

$$[\sigma(\mathbf{B})]_{pc} = \frac{e^{b_{pc}}}{\sum_{p'=1}^{|\mathcal{P}|} e^{b_{p'c}}}, \ \forall (p,c) \in (1,|\mathcal{P}|) \times (1,C). \quad (1)$$

This allows the localization layer to choose the most relevant proposals for each class. Subsequently, the classification stream output is weighted by $\sigma(\mathbf{B})$ through element-wise multiplication: $\mathbf{E} = \mathbf{A} \odot \sigma(\mathbf{B})$. Class scores over the video are obtained by summing the resulting weighted scores in $\mathbf{E}$. Concurrent work by [63] discusses a similar MIL module for audio classification.

After performing the above stated operations for both audio and visual sub-modules, in the final step, the global video-level scores are ℓ_2 normalized and added. In preliminary experiments we found this to work better than addition of unnormalized scores. We hypothesize that the system trains better because ℓ_2 normalization ensures that the scores being added are in the same range.

C. Classification loss and network training

Given a set of L training videos and labels, $\{(V^{(l)}, \boldsymbol{y}^{(l)})\}_{l=1}^{L}$, we solve a multi-label classification problem. Here $\boldsymbol{y} \in \mathcal{Y} = \{-1,+1\}^C$ with the class presence denoted by +1 and absence by -1. To recall, for each video $V^{(l)}$, the network takes as input a set of image regions $\mathcal{R}^{(l)}$ and audio segments $\mathcal{A}^{(l)}$. After performing the described operations on each modality separately, the ℓ_2 normalized scores are added and represented by $\phi(V^{(l)}; \boldsymbol{w}) \in \mathbb{R}^C$, with all network weights and biases denoted by $\boldsymbol{w}$. All the weights, including and following the fully-connected layer processing stage for both the modalities, are included in $\boldsymbol{w}$. Note that both sub-modules are trained jointly.

The network is trained using the multi-label hinge loss on a batch of size B:

$$L(\boldsymbol{w}) = \frac{1}{CB} \sum_{l=1}^{B} \sum_{c=1}^{C} \max\Big(0, 1 - y_c^{(l)} \phi_c(V^{(l)}; \boldsymbol{w})\Big). \quad (2)$$

To summarize, we have discussed a general instantiation of our framework, capable of processing spatio-temporal visual regions, temporal audio segments for event classification and localizing characteristic proposal in each modality. Dealing with each proposal independent of the time at which it occurs allows tackling AV asynchronicity.

D. Variants

In the proposed framework (depiced in Fig. 2) module design can be flexibly modified in a task–specific manner. To demonstrate this, we discuss next two variants that allow performing audio source enhancement and synchronous AV fusion, respectively.

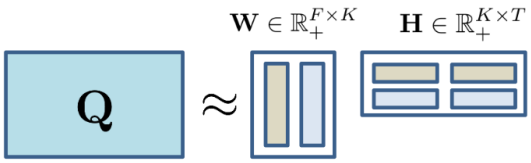


Fig. 4. NMF component proposals depiction where spectral patterns, $\mathbf{w}_k$ and corresponding activation vectors, $\mathbf{h}_k$ are shown in the same colour. Furthermore, each part in $\mathbf{h}_k$ refers to a non-overlapping temporal segment.

1) Source enhancement variant: Here we propose to design novel audio proposals using NMF with the goal of enhancing the audio source of interest. The primary reason for performing such a decomposition is the hope that each of the resulting spectral patterns would represent a part of just one source. Specifically, using NMF we decompose audio magnitude spectrograms $\mathbf{Q} \in \mathbb{R}_+^{F\times N}$ consisting of F frequency bins and N short-time Fourier transform (STFT) frames, such that,

$$\mathbf{Q} \approx \mathbf{W}\mathbf{H}, \quad (3)$$

where $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ are nonnegative matrices that can be interpreted as the characteristic audio spectral patterns $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and their temporal activations $\mathbf{H} \in \mathbb{R}_+^{K\times N}$, respectively. Here K is the total number of spectral patterns.

We then apply soft mask based filtering [64] to an audio recording to decompose it into K tracks (also referred to as NMF components) each obtained from $\mathbf{w}_k, \mathbf{h}_k$ for $k \in [1, K]$, where $\mathbf{w}_k$ and $\mathbf{h}_k$ denote spectral pattern and activation vectors corresponding to the k^{th} component, respectively. This is depicted in Fig. 4.

They can now be considered as proposals that may or may not belong to the class of interest. Specifically, we chunk each NMF component into temporal segments, which we call NMF Component proposals or NCPs. We denote the set of NCPs by $\mathcal{D} = \{d_{k,t}\}$, where each element is indexed by the component, $k \in [1, K]$ and temporal segment $t \in [1, T]$. The same audio network is used for both TSPs and NCPs. Thus, for each NMF component or track we follow the TSP computation procedure. However, this is done with a non-overlapping window for reducing computational load.

Our system scores each NMF component with respect to its relevance for a particular class. These relevance scores can be appropriately aggregated to perform source enhancement. We proceed as follows:

- Denoting by $\beta_{k,t}$ the score for k^{th} component's t^{th} temporal segment, we compute a global score for each component as

$$\alpha_k = \max_{t\in T} \ \beta_{k,t}.$$

  It is worth mentioning that other pooling strategies such as mean or weighted rank pooling [44] could also be considered instead of the max operation. However, in our preliminary experiments we found them to yield similar results.
- Next, we apply min-max scaling between [0,1]:

$$\alpha'_k = \frac{\alpha_k - \alpha^l}{\alpha^u - \alpha^l}, \ \text{where } \alpha^l = \min_{k'}(\alpha'_k), \ \alpha^u = \max_{k'}(\alpha'_k)$$

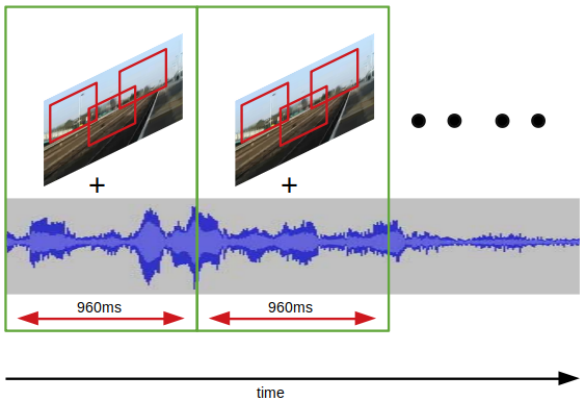


Fig. 5. Synchronized variant - herein audio and visual scores over each temporal segment are aggregated and the best temporal segment is chosen for classification.

- This is followed by soft mask based source and noise spectrogram reconstruction using complex-valued mixture STFT $\mathbf{X}$. Note that we can optionally apply a hard threshold τ on α'_k to choose the top ranked components for the source. This amounts to replacing α'_k by the indicator function $\mathbf{1}[\alpha'_k \geq \tau]$ in the following reconstruction equations:

$$\mathbf{S} = \frac{\sum_{k=1}^{K} \alpha'_k \mathbf{w}_k \mathbf{h}_k}{\mathbf{WH}} \mathbf{X} \tag{4}$$

$$\mathbf{N} = \frac{\sum_{k=1}^{K} (1-\alpha'_k) \mathbf{w}_k \mathbf{h}_k}{\mathbf{WH}} \mathbf{X} \tag{5}$$

  Here $\mathbf{S}$ and $\mathbf{N}$ are the estimates of source of interest and of background noise, respectively. These can be converted back to the time domain using inverse STFT.

It is worth noting two key differences with the approach in [32]: (i) In [32] only the NMF basis vectors are used for training without their corresponding activations. Hence no temporal information is utilized. (ii) Unlike us, they perform a supervised dictionary construction step after training to decompose a test signal.

2) Synchronous fusion variant: Framework instantiation depicted in Fig. 3 constructs the global score vector for each modality by combining scores over all the proposals, regardless of their temporal index. As noted, such a system is capable of dealing with asynchronous appearance of cues in both the modalities. On the other hand, we could envision a synchronized variant, where we only add scores of visual and audio proposals appearing in the same temporal segment. And construct the global score vector by choosing for each class the best scoring temporal segment. This is illustrated in Fig. 5. This essentially allows us to determine temporal segments where AV cues appear simultaneously. We list below specific changes made to the proposal score computation and fusion module:

1) Firstly, in the localization stream the softmax operation is performed over proposals from each temporal window separately. This amounts to replacing $|\mathcal{P}|$ by $|\mathcal{P}_t|$ in equation (1), where the proposals are indexed by the temporal segment they belong to. For the visual branch this corresponds to region proposals from a frame within the t^{th} temporal segment.
2) Secondly, after obtaining $\mathbf{E}$ *i.e.* the output of the two stream classification, we compute a class score vector for each temporal interval by summing up proposal scores separately over $p \in \mathcal{P}_t$. This gives us a matrix with dimensions $C \times T$ in each modality. Their addition gives us a synchronous AV temporal score.
3) Finally, for each class, the best AV temporal segment is chosen through a $\log - \text{sum} - \exp$ operation. This gives us the class score vector ϕ required for weakly–supervised training using multi–label hinge loss (refer to equation (2)).

IV. Experimental Validation

A. Setup

All systems except that of [23], including variants, are implemented in Tensorflow. They were trained for 25K iterations using Adam optimizer [65] with a learning rate of 10^{-5} and a batch size of 24. We use the MATLAB implementation of EdgeBoxes for generating region proposals, obtaining approximately 100 regions per video with $M_{\text{img}} = 10$ and a duration of 10 sec. The implementation is used with default parameter setting. Base visual features, $\boldsymbol{x}_{vis} \in \mathbb{R}^{9216}$ are extracted using `caffenet` [58] with pre-trained ImageNet weights and RoI pooling layer modification [48]. With 6×6 RoI pooling we get a 9216 ($= 256 \times 6 \times 6$) dimensional feature vector. For this, the Fast-RCNN Caffe implementation is used [48]. The fully connected layers, namely fc_6 and fc_7, each with 4096 neurons, are fine-tuned, with 50% dropout during training.

For audio, each recording is resampled to 16 kHz before processing. Log-Mel spectrum over the whole file is computed with a window size of 25ms and 10ms hop length. The resulting spectrum is chunked into segment proposals using a 960–ms window with a 480–ms stride.

For a 10–second recording, this yields 20 segments of size 96×64. We use the official Tensorflow implementation of `vggish`.[4]

B. Datasets

DCASE Smart Cars. We use the recently introduced dataset for the DCASE challenge on large-scale weakly supervised sound event detection for smart cars [21]. This is a subset of Audioset [22] which contains a collection of weakly-annotated unconstrained YouTube videos of vehicle and warning sounds spread over 17 classes. It is categorized as follows (abbreviations used in experiment tables are given in parenthesis that follow each category):

- *Warning sounds*: Train horn (trn-hrn), Air/Truck horn (air-hrn), Car alarm (car-alm), Reversing beeps (rv-bps), Ambulance siren (amb), Police car siren (pol-car), Fire engine/fire truck siren (f-eng), Civil defense siren (civ-def), Screaming (scrm).

[4] https://github.com/tensorflow/models/tree/master/research/audioset

- *Vehicle sounds*: Bicycle (bik), Skateboard (skt), Car (car), Car passing by (car-pby), Bus (bus), Truck (trk), Motorcycle (mbik), Train (trn).

This multi-label dataset contains 51,172 training, 488 validation and 1103 testing samples. Despite our best efforts, due to download issues, we were able to fetch 48,719 training, 462 validation and 1030 testing clips. It is worth mentioning that the training data is highly unbalanced with the number of samples for the classes ranging from 175 to 24K. To mitigate the negative effect of this imbalance on training, we introduce some balance by ensuring that each training batch contains at least one sample from some or all of the under-represented classes. Briefly, each batch is generated by first randomly sampling labels from a specific list, followed by fetching examples corresponding to the number of times each label is sampled. This list is generated by ensuring higher but limited presence of classes with more examples. We use a publicly available implementation for this purpose [23].[5]

Kinetics instruments (KI). We also use a subset of the Kinetics dataset [14] that contains 10-s YouTube videos from 15 music instrument classes. From a total of 10,267 videos, we create training and testing sets that contain 9199 and 1023 videos, respectively. For source enhancement evaluation, we handpicked 45 "clean" instrument recordings, 3 per class. Due to their unconstrained nature, the audio recordings are mostly noisy, *i.e.* videos are either shot with accompanying music/instruments or in acoustic environments containing other background events. In that context, "clean" refers to solo instrument samples with minimal amount of such noise.

In what follows, we thoroughly evaluate the proposed framework's performance on various scene analysis tasks. In particular, we compare the asynchronous and synchronous variants of our system against several strong baselines for event classification on the DCASE smart cars benchmark. Generalization to diverse object types is shown through results on KI. This is followed by results for temporal localization of the audio event on DCASE. For completeness, we also present experiments on segregating the audio source of interest, as discussed in our prior work [66]. This allows us to demonstrate our system's capability to perform good source enhancement while training just for weak label classification. This is done by utilizing NMF-based proposals as described in Sec. III-D1. We conclude this section with a discussion of qualitative visual localization examples that show how we deal with extreme noise, including asynchronous AV cues.

C. Event classification

Baselines. To our best knowledge, there is no prior work on deep architectures that perform the task of weakly supervised classification and localization for unsynchronized AV events. Our task and method are substantially different from recently proposed networks like L3 [10], [11] which are trained using synchronous AV pairs on a large collection of videos in a self-supervised manner. However, we designed several strong baselines for comparison and an ablation study. In particular, we compare against the following networks:

1) AV One-Stream Architecture: Applying MIL in a straight-forward manner, we could proceed only with a single stream. That is, we can use the classification stream followed by a max operation for selecting the highest scoring regions and segments for obtaining global video-level scores. As done in [37], we choose to implement this as a multimodal MIL-based baseline. We replace the $\max$ operation by the $\log-\text{sum}-\exp$ operator, its soft approximation. This has been shown to yield better results [34]. The scores on both the streams are ℓ_2 normalized before addition for classification. This essentially amounts to removing from Fig. 3 the localization branches and replacing the summation over proposals with the soft-maximum operation described above. To avoid any confusion, please note that we use the term 'stream' to refer to classification and localization parts of the scoring network.

2) Visual-Only and Audio-Only Networks: These networks only utilize one of the modalities for classification. However, note that there are still two streams for classification and localization, respectively. For a fair comparison and ablation study we train these networks with ℓ_2 normalization. In addition, for completeness we also implement Bilen *et al.*'s architecture for weakly supervised deep detection networks (WSDDN) with an additional softmax on the classification stream. As the scores are in the range [0,1], we train this particular network with C binary log-loss terms [37]. When discussing results we refer to this system as WSDDN-Type.

3) CVSSP Audio-Only [23]: This state-of-the-art method is the DCASE 2017 challenge winner for the audio event classification sub-task. The system is based on Gated convolutional RNN (CRNN) for better temporal modeling and attention-based localization. They use no external data and training/evaluation is carried out on all the samples. We present results for both their winning fusion system, which combines prediction of various models and Gated-RCNN model trained with log-Mel spectrum.

Results and discussion. We show in Table I the micro-averaged F1 scores for each of the systems described in this paper. In particular, systems (a)-(b) in Table I are the proposed asynchronous and synchronous AV systems respectively and (c)-(f) present variants of (a) which are also treated as baselines, (g)-(h) denote results from CVSSP team [23], winners of the DCASE AED for smart cars audio event tagging task. The proposed systems and their variants are trained with audio temporal segment proposals only. Our proposed two stream multimodal and audio-only systems (a,b,c) outperform all the other approaches by a significant margin. Among the multimodal systems, the two-stream architecture performs much better than the one-stream counter-part, designed with only a classification stream and soft-maximum for region selection.

[5] https://github.com/yongxuUSTC/dcase2017_task4_cvssp/blob/master/data_generator.py

On the other hand, the state-of-the-art CVSSP fusion system, which combines predictions of various models, achieves a better precision than the other methods. It is also worth mentioning that performance of the sync. AV system (b) is lower than the unsynchronized one (a). This is expected as the dataset contains some samples with asynchronously appearing cues. However, the sync. system would still be useful for detecting temporal segments where the AV cues appear together. Several important and interesting observations can be made by looking at these results in conjunction with the class-wise scores reported in Table II.

Most importantly, the results emphasize the complementary role of visual and audio sub-modules for this task. To see this, we could categorize the data into two sets: (i) classes with clearly defined AV elements, for instance car, train, motorcycle; (ii) some warning sounds such as, *e.g.*, reverse beeping, screaming, air horn, where the visual object's presence is ambiguous. The class-wise results of the video only system are a clear indication of this split. Well-defined visual cues enhance the performance of the proposed multimodal system over audio-only approaches, as video frames carry vital information about the object. On the other hand, in the case of warning sounds, video frames alone are insufficient as evidenced by results for the video-only system. In this case, the presence of audio assists the system in arriving at the correct prediction. The expected AV complementarity is clearly established through these results.

Note that for some warning sounds the CVSSP method achieves better results. In this regard, we believe better temporal modeling for our audio system could lead to further improvements. In fact, we currently operate with a coarse temporal window of 960ms, which might not be ideal for all audio events. RNNs could also be used for further improvements. We think such improvements are orthogonal and were not the focus of this study. We also observe that results for under-represented classes in the training data such as air horn and reversing beeps are relatively lower. This can possibly be mitigated through data augmentation strategies.

In Table III we report results for the case where all layers of `vggish` are fine-tuned. For this, we remove the FC adaptation layer from the audio network (refer to Fig. 3). It is also worth noting that for these experiments, we reduced the batch size to one due to memory constraints. For DCASE data, which contains approximately 48K training samples, this results in significantly more number of variable updates. Thus, to avoid overfitting, we run the system for 10 epochs and report results with the model that gives the lowest validation error. As expected, fine-tuning `vggish` results in improved performance as the audio features are better adapted to the dataset. We also see competitive instrument classification performance with KI, where the multimodal system fairs better than audio alone.

D. Audio temporal localization

We show the sound event detection performance on DCASE smart cars data in Table IV. Following DCASE evaluation protocol, here we report segment–wise aggregated F1 score and error rate (ER) for each system. The official metric, ER, computes total number of substitution, deletion and insertion errors by comparing the ground truth and estimated output using one second long sub–segments [13].

TABLE I
RESULTS ON DCASE SMART CARS TASK TEST SET. WE REPORT HERE THE MICRO-AVERAGED F1 SCORE, PRECISION AND RECALL VALUES AND COMPARE WITH STATE-OF-THE-ART. TS IS AN ACRONYM FOR TWO-STREAM.

	System	F1	Precision	Recall
(a)	AV Two Stream	**64.2**	59.7	**69.4**
(b)	Sync. AV Two Stream	62.0	57.2	67.6
(c)	TS Audio-Only	57.3	53.2	62.0
(d)	TS Video-Only	47.3	48.5	46.1
(e)	TS Video-Only WSDDN-Type [37]	48.8	47.6	50.1
(f)	AV One Stream	55.3	50.4	61.2
(g)	CVSSP - Fusion system [23]	55.6	**61.4**	50.8
(h)	CVSSP - Gated-CRNN-logMel [23]	54.2	58.9	50.2

The results for the proposed systems are computed by simply thresholding the two–stream output from the audio sub–module at $\tau = 0$ for the predicted label(s). We note that the results are comparable with the best performing CVSSP system. Note that the winning system for this subtask from Lee *et al.* [67] employs an ensemble method to optimally weigh multiple learned models, using ER as the performance metric to make the final selection. No such fine tuning is performed in our case.

E. Audio source enhancement

Systems. We evaluate audio-visual (V + A) systems with different audio proposal types, namely:

- A (NCP): NMF component proposals,
- A (TSP, NCP): all TSPs and NCPs are put together into the same bag and fed to the audio network.

`vggish` is fine-tuned (as discussed earlier) for the systems listed above to adapt to NCP input.

Baselines. We compare with the following NMF related methods:

- Supervised NMF [68]: We use the class labels to train separate dictionaries of size 100 for each music instrument with stochastic mini-batch updates. At test time, depending on the label, the mixture is projected onto the appropriate dictionary for source reconstruction.
- NMF Mel-Clustering [69]: This blind audio-only method reconstructs source and noise signals by clustering mel-spectra of NMF components. We take help of the example code provided online for implementation in MATLAB [70].

Testing protocol. We corrupt the original audio with background noise corresponding to recordings of environments such as bus, busy street, park, etc. using one audio file per scene from the DCASE 2013 scene classification dataset [71]. The system can be utilized in two modes: *label known* and *label unknown*. For the former, where the source of interest is known, we simply use the proposal ranking given by the

System	Vehicle Sounds								Warning Sounds								
	bik	bus	car	car-pby	mbik	skt	trn	trk	air-hrn	amb	car-alm	civ-def	f-eng	pol-car	rv-bps	scrm	trn-hrn
AV TS	**75.7**	54.9	75.0	**34.6**	**76.2**	78.6	82.0	**61.5**	40.0	64.7	53.9	80.4	64.4	49.2	36.6	81.1	47.1
Sync. AV TS	65.0	**55.6**	**75.7**	25.6	74.0	**80.5**	**85.1**	57.8	28.4	**65.7**	54.1	82.1	61.3	52.6	39.6	70.6	48.8
TS Audio-Only	42.1	38.8	69.8	29.6	68.9	64.9	78.5	44.0	40.4	58.2	53.0	79.6	61.0	51.4	42.9	72.1	46.9
TS Video-Only	72.5	52.0	61.2	15.0	54.1	64.2	73.3	49.7	12.0	33.9	13.5	68.6	46.5	19.8	21.8	44.1	32.1
AV OS	68.2	53.6	74.1	25.6	67.1	74.4	82.8	52.8	28.0	54.7	20.6	76.6	60.4	56.3	18.8	49.4	36.2
CVSSP - FS	40.5	39.7	72.9	27.1	63.5	74.5	79.2	52.3	**63.7**	35.6	**72.9**	**86.4**	**65.7**	**63.8**	**60.3**	**91.2**	**73.6**

TABLE II
CLASS-WISE COMPARISON ON TEST SET USING F1 SCORES. WE USE TS, OS AND FS AS ACRONYMS TO REFER TO TWO-STREAM, ONE-STREAM AND FUSION SYSTEM, RESPECTIVELY. CLASS ABBREVIATIONS ARE DETAILED IN SEC. IV-B

TABLE III
RESULTS ON DCASE AND KI WITH FINE TUNED VGGISH

Systems	DCASE F1	DCASE Precision	DCASE Recall	KI Accuracy
AV TS - VGGISH FT	65.0	64.9	65.0	84.5
AO TS - VGGISH FT	61.7	61.5	61.9	75.3

TABLE IV
F1 SCORE AND ERROR RATE FOR SOUND EVENT DETECTION TASK

System	F1	ER
AV TS	51.0	0.76
AO TS	48.5	0.78
AV TS - VGGISH FT	52.3	0.74
AO TS - VGGISH FT	53.0	0.75
CVSSP - Fusion system [23]	51.8	0.73
CVSSP - Gated-CRNN-logMel [23]	47.5	0.78
SNU - Ensemble method [67]	**55.5**	**0.66**

corresponding classifier for reconstruction. For the latter, the system's classification output is used to infer the source.

Results and discussion. We report, in Table V, average Source to Distortion Ratio (SDR) [72] over 450 audio mixtures created by mixing each of the 45 clean samples from the dataset with 10 noisy audio scenes. The results look promising but not state-of-the-art. This performance gap can be explained by noting that the audio network is trained for the task of audio event detection and thus does not yield optimal performance for source enhancement. The network focuses on discriminative components, failing to separate some source components from the noise by a larger margin, possibly requiring adaptive thresholding for best results. In other words, as the component scores vary for each example, a single threshold for all cases proves to be sub-optimal. It is worth noting that performance for the proposed systems does not degrade when used in "Label Unknown" mode, indicating that despite incorrect classification the system is able to cluster acoustically similar sounds. Performance of supervised NMF seems to suffer due to training on a noisy dataset. Separation results on in-the-wild YouTube videos are made available on our companion website.[6]

[6]http://bit.ly/2HEJbrl

TABLE V
AVERAGE SDR OVER MIXTURES CREATED BY COMBINING CLEAN INSTRUMENT EXAMPLES WITH ENVIRONMENTAL SCENES.

System	Label Known	Label Unknown
Supervised NMF	2.3	–
NMF Mel-Clustering	–	**4.3**
V + A (NCP), soft	3.3	3.3
V + A (NCP), $\tau = 0.1$	**3.8**	3.9
V + A (NCP), $\tau = 0.2$	3.6	3.6
V + A (NCP, TSP), soft	2.1	2.2

F. Qualitative visual localization

In Fig. 6 we present some visual localization results for the 'train' category from DCASE. Localization in extreme asynchronous conditions is also discussed in Fig. 7. In the first case **A**, the sound of a car's engine is heard in the first two seconds followed by music. The normalized audio localization heatmap at the bottom displays the scores assigned to each temporal audio segment, s_t by the car classifier. The video frames placed above are roughly aligned with the audio temporal axis to show the video frame at the instant when the car sounds and the point where the visual network localizes. The localization is displayed through a yellow bounding box. To better understand the system's output, we modulate the opacity of the bounding box according to the system's score for it. Higher the score, more visible the bounding box. As expected, we do not observe any yellow edges in the first frame. Clearly, there exists temporal asynchrony, where the system locks onto the car, much later, when it is completely visible. **B** depicts an example, where due to extreme lighting conditions the visual object is not visible. Here too, we localize the audio object and correctly predict the 'motorcycle' class.

For full videos and more such examples we refer the reader to our companion website.[6]

V. CONCLUSION

Building upon ideas from multiple instance learning, we have proposed a modular deep AV scene understanding framework that can be trained jointly to perform several tasks simultaneously. Exploiting our method's modularity, we investigate several instantiations capable of dealing with unsynchronized AV cue appearance, determining synchronous temporal segments and segregating the audio into constituent

Fig. 6. Visual localization on DCASE test video frames from the 'train' category. The localization results are shown in green bounding boxes. Below each image we display the scaled region proposal (top) and audio segment scores for the class of interest as heatmaps. The visual heatmap is a concatenation of proposals from all the sub-sampled frames, arranged in temporal order. More results on our companion website.[6]

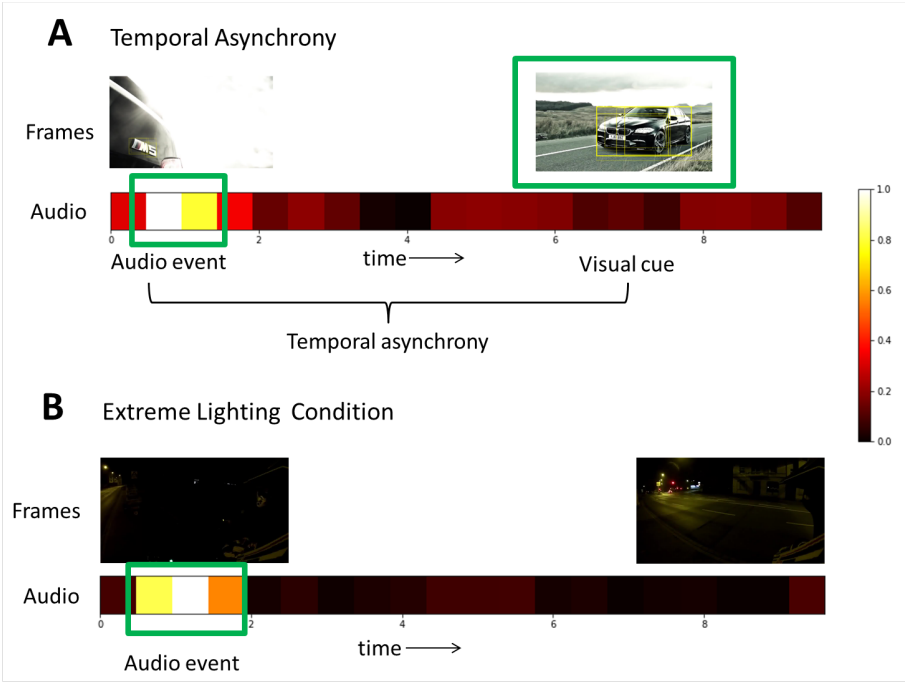


Fig. 7. Qualitative results for unsynchronized AV events. For both the cases A and B, the heatmap at the bottom denotes audio localization over segments for the class under consideration. For heatmap display, the audio localization vector has been scaled to lie between [0,1]. The top row depicts video frames roughly aligned to the audio temporal axis. (A) Top: Here we show a video where the visual object of interest appears after the audio event. This is a 'car' video from the validation split. The video frames show bounding boxes where edge opacity is controlled by the box's detection score. In other words, higher score implies better visibility (B) Bottom: This is a case from the evaluation data where due to lighting conditions, the visual object is not visible. However the system correctly localizes in audio and predicts the 'motorcycle' class.

sources. The latter is made possible through a novel use of NMF decomposition, where, unlike most earlier methods, we only use the given weak labels for training. We report state-of-the-art event classification performance on DCASE 2017 smart cars data along with promising results for spatio-temporal visual localization, audio event detection and source separation. The method generalizes well to diverse object types. Experiments have have also shown that a more accurate audio temporal modeling would be needed to better cope with situations where the visual modality is inefficient. Furthermore, we believe the presented method could benefit from appropriately incorporating several recent developments in feature and modality fusion [73], [74].

References

[1] J. Woodcock, W. J. Davies, T. J. Cox, F. Melchior *et al.*, "Categorization of broadcast audio objects in complex auditory scenes," *Journal of the Audio Engineering Society*, vol. 64, no. 6, pp. 380–394, 2016.

[2] W. Jiang, C. Cotton, S. F. Chang, D. Ellis, and A. Loui, "Short-term audiovisual atoms for generic video concept classification," in *Proceedings of the 17th ACM International Conference on Multimedia*. ACM, 2009, pp. 5–14.

[3] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo, "Large-scale multimodal semantic concept detection for consumer video," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*. ACM, 2007, pp. 255–264.

[4] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 2, pp. 352–364, 2018.

[5] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level

event recognition in unconstrained videos," *International journal of multimedia information retrieval*, vol. 2, no. 2, pp. 73–101, 2013.
[6] H. Izadinia, I. Saleemi, and M. Shah, "Multimodal analysis for identification and segmentation of moving-sounding objects," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 378–390, Feb 2013.
[7] E. Kidron, Y. Schechner, and M. Elad, "Pixels that sound," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, June 2005, pp. 88–95 vol. 1.
[8] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2405–2413.
[9] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *Proc. of European Conference on Computer Vision*. Springer, 2016, pp. 801–816.
[10] R. Arandjelović and A. Zisserman, "Look, listen and learn," in *IEEE International Conference on Computer Vision*, 2017.
[11] ——, "Objects that sound," *CoRR*, vol. abs/1712.06651, 2017.
[12] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. of International Conference on Machine Learning*, 2013, pp. 1247–1255.
[13] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 85–92.
[14] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
[15] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.
[16] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations," in *ICASSP*. IEEE, 2015, pp. 151–155.
[17] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.
[18] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *ICASSP*. IEEE, 2017, pp. 771–775.
[19] V. Bisot, S. Essid, and G. Richard, "Overlapping sound event detection with supervised nonnegative matrix factorization," in *ICASSP*. IEEE, 2017, pp. 31–35.
[20] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1038–1047.
[21] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE2017 challenge setup: Tasks, datasets and baseline system," in *Proc. of Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 85–92.
[22] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 776–780.
[23] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Surrey-CVSSP system for DCASE2017 challenge task4," DCASE2017 Challenge, Tech. Rep., September 2017.
[24] J. Salamon, B. McFee, and P. Li, "DCASE 2017 submission: Multiple instance learning for sound event detection," DCASE2017 Challenge, Tech. Rep., September 2017.
[25] A. Kumar, M. Khadkevich, and C. Fugen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," *arXiv preprint arXiv:1711.01369*, 2017.
[26] Q. Kong, C. Yu, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Weakly labelled audioset classification with attention neural networks," *arXiv preprint arXiv:1903.00765*, 2019.
[27] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.
[28] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 31–35.
[29] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, "Sound event detection and time–frequency segmentation from weakly labelled data," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 27, no. 4, pp. 777–787, 2019.
[30] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *ECCV*, September 2018.
[31] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 112:1–112:11, Jul. 2018. [Online]. Available: http://doi.acm.org/10.1145/3197517.3201357
[32] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in *ECCV*, September 2018.
[33] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Advances in neural information processing systems*, 2006, pp. 1417–1424.
[34] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with posterior regularization," in *Proceedings BMVC 2014*, 2014, pp. 1–12.
[35] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 685–694.
[36] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "Contextlocnet: Context-aware deep network models for weakly supervised localization," in *European Conference on Computer Vision*. Springer, 2016, pp. 350–365.
[37] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *CVPR*, 2016, pp. 2846–2854.
[38] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016, pp. 2921–2929.
[39] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 189–203, 2017.
[40] H. Bilen, V. P. Namboodiri, and L. J. Van Gool, "Object and action classification with latent window parameters," *International Journal of Computer Vision*, vol. 106, no. 3, pp. 237–251, 2014.
[41] T. Deselaers, B. Alexe, and V. Ferrari, "Localizing objects while learning their appearance," in *European conference on computer vision*. Springer, 2010, pp. 452–466.
[42] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Advances in Neural Information Processing Systems*, 2010, pp. 1189–1197.
[43] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell, "Weakly-supervised discovery of visual pattern configurations," in *Advances in Neural Information Processing Systems*, 2014, pp. 1637–1645.
[44] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 695–711.
[45] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with r* cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1080–1088.
[46] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*. Springer, 2014, pp. 391–405.
[47] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.
[48] R. Girshick, "Fast R-CNN," in *ICCV*. IEEE, 2015, pp. 1440–1448.
[49] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.
[50] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
[51] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.
[52] J. C. Van Gemert, M. Jain, E. Gati, C. G. Snoek *et al.*, "Apt: Action localization proposals from dense trajectories." in *Proc. of BMVC*, vol. 2, 2015, p. 4.
[53] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid, "Spatio-temporal object detection proposals," in *European conference on computer vision*. Springer, 2014, pp. 737–752.

[54] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[55] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 17–24.

[56] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[57] J. Hosang, R. Benenson, and B. Schiele, "How good are detection proposals, really?" in *25th British Machine Vision Conference*. BMVA Press, 2014, pp. 1–12.

[58] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[59] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[60] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *Readings in speech recognition*. Elsevier, 1990, pp. 65–74.

[61] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "CNN architectures for large-scale audio classification," in *ICASSP*. IEEE, 2017, pp. 131–135.

[62] S. Abu-El-Haija, N. Kothari, J. Lee, A. P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8M: A large-scale video classification benchmark," in *arXiv:1609.08675*, 2016. [Online]. Available: https://arxiv.org/pdf/1609.08675v1.pdf

[63] C. Yu, K. S. Barsim, Q. Kong, and B. Yang, "Multi-level attention model for weakly supervised audio classification," *arXiv preprint arXiv:1803.02353*, 2018.

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

[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[66] S. Parekh, A. Ozerov, S. Essid, N. Duong, P. Pérez, and G. Richard, "Identify, locate and separate: Audio-visual object extraction in large video collections using weak supervision," *under review at WASPAA 2019*, Draft available at https://arxiv.org/abs/1811.04000.

[67] D. Lee, S. Lee, Y. Han, and K. Lee, "Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input," DCASE2017 Challenge, Tech. Rep., September 2017.

[68] C. Févotte, E. Vincent, and A. Ozerov, "Single-channel audio source separation with NMF: divergences, constraints and algorithms," in *Audio Source Separation*. Springer, 2018, pp. 1–24.

[69] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *in Proceedings of International Conference on Digital Audio Effects DAFx09*, 2009.

[70] *NMF Mel Clustering Code*, http://www.ient.rwth-aachen.de/cms/dafx09/.

[71] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[72] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[73] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances In Neural Information Processing Systems*, 2016, pp. 289–297.

[74] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *ECCV*, September 2018.

Sanjeel Parekh received B. Tech (hons.) degree in electronics and communication engineering from LNM Institute of Information Technology, India in 2014 and M.S. in Sound and Music Computing from Universitat Pompeu Fabra (UPF), Spain in 2015. His Ph.D. thesis titled 'Learning representations for robust audio-visual scene analysis' was completed in collaboration with Technicolor R&D and Télécom ParisTech, France between 2016-19. His research focuses on developing and applying machine learning techniques to problems in audio and visual domains. Currently, he is with LTCI lab at Telecom ParisTech, France.

Slim Essid received his state engineering degree from the Éole Nationale d'Ingénieurs de Tunis, Tunisia, in 2001, his M.Sc. (D.E.A.) degree in digital communication systems from the cole Nationale Supérieure des Télécommunications, Paris, France, in 2002, his Ph.D. degree from the Universit Pierre et Marie Curie (UPMC), Paris, France, in 2005, and his Habilitation à Diriger des Recherches degree from UPMC in 2015. He is a professor in Télécom ParisTechs Department of Images, Data, and Signals and the head of the Audio Data Analysis and Signal Processing team. His research interests are machine learning for audio and multimodal data analysis. He has been involved in various collaborative French and European research projects, among them Quaero, Networks of Excellence FP6-Kspace, FP7-3DLife, FP7-REVERIE, and FP-7 LASIE. He has published over 100 peer-reviewed conference and journal papers, with more than 100 distinct co-authors. On a regular basis, he serves as a reviewer for various machine-learning, signal processing, audio, and multimedia conferences and journals, e.g., a number of IEEE transactions, and as an expert for research funding agencies.

Alexey Ozerov received the M.Sc. degree in mathematics from the Saint-Petersburg State University,Saint Petersburg, Russia, in 1999, the M.Sc. degree in applied mathematics from the University of Bordeaux 1, Bordeaux, France, in 2003, and the Ph.D. degree in signal processing from the University of Rennes 1, Rennes, France. He was working toward the Ph.D. degree from 2003 to 2006 with the labs of France Telecom R&D and in collaboration with the IRISA institute. From 1999 to 2002, he worked with Terayon Communicational Systems (USA) as an R&D software engineer, first in Saint-Petersburg, and then, in Prague, Czech Republic. He was with Sound and Image Processing Lab, KTH Royal Institute of Technology, Stockholm, Sweden, for one year (2007), with the Signal and Image Processing (TSI) Department, TELECOM ParisTech/CNRS LTCI, for one and half year (20082009), and with METISS team, IRISA/INRIA, Rennes,France, for two years (20092011). He is currently with Technicolor Research & Innovation, Rennes, France. His research interests include various aspects of audio and image/video analysis and processing. Since 2016, he has been a Distinguished Member of the Technicolor Fellowship Network and is currently a Member of the IEEE Signal Processing Society Audio and Acoustic Signal Processing Technical Committee. He is currently an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and received the IEEE Signal Processing Society Best Paper Award in 2014.

Ngoc Q.K. Duong received the B.S. degree from the Posts and Telecommunications Institute of Technology, Hanoi City, Vietnam, in 2004, the M.S. degree in electronic engineering from Paichai University, Daejeon, South Korea, in 2008, and the Ph.D. degree in computer science and control with the French National Institute for Research, Rennes, France, in 2011. From 2004 to 2006, he was with Visco JSC as a System Engineer. He was also a Research Engineer for the acoustic echo/noise cancellation system with Emersys Company, Korea, in 2008. He is currently a Senior Scientist with Technicolor R&D, Rennes, France, where he has worked since Nov. 2011. He is the co-author of more than 45 scientific papers and about 30 patent submissions. His research interests include signal processing and machine learning, applied to audio, image, and video. He was the recipient of the several research awards, including the IEEE Signal Processing Society Young Author Best Paper Award, in 2012 and the Bretagne Young Researcher Award,in 2015.

Patrick Pérez is a Scientific Director with Valeo.ai,Paris, France, a Valeo research lab on artificial intelligence for automotive applications. Before joining Valeo, he was a Distinguished Scientist with Technicolor (2009-2018), and a Researcher with Inria (1993-2000, 2004-2009) and with Microsoft Research Cambridge (2000-2004). His research interests include audio/video description, search, and analysis. He is currently on the Editorial Board for the International Journal of Computer Vision.

Gaël Richard received his State Engineering degree from Télécom ParisTech, France, in 1990, his Ph.D. degree from the University of Paris XI, France, in 1994 in speech synthesis, and his Habilitation à Diriger des Recherches degree from the University of Paris XI in 2001. After receiving his Ph.D. degree, he spent two years at Rutgers University, Piscataway, New Jersey, in the Speech Processing Group of Prof. J. Flanagan, where he explored innovative approaches for speech production. In 2001, he joined Télécom ParisTech, where he is now a professor in audio signal processing and the head of the Image, Data, and Signal Department. He is co-author of more than 200 papers, His research interests are mainly in the field of speech and audio signal processing and include topics such as signal representations and signal models, source separation, machine-learning methods for audio/music signals, music information retrieval, and multimodal audio processing. He is a Fellow of the IEEE.

Appendix D

Paper 4: VideoMem: Constructing, Analyzing, Predicting Short-Term and Long-Term Video Memorability

VideoMem: Constructing, Analyzing, Predicting Short-Term and Long-Term Video Memorability

Romain Cohendet
Technicolor, Rennes, France
romain.cohendet@laposte.net

Claire-Hélène Demarty
InterDigital, Rennes, France
claire-helene.demarty@interdigital.com

Ngoc Q. K. Duong
InterDigital, Rennes, France
quang-khanh-ngoc.duong@interdigital.com

Martin Engilberge
InterDigital, Rennes, France
martin.engilberge@interdigital.com

Abstract

Humans share a strong tendency to memorize/forget some of the visual information they encounter. This paper focuses on understanding the intrinsic memorability of visual content. To address this challenge, we introduce a large scale dataset (VideoMem) composed of 10,000 videos with memorability scores. In contrast to previous work on image memorability – where memorability was measured a few minutes after memorization – memory performance is measured twice: a few minutes and again 24-72 hours after memorization. Hence, the dataset comes with short-term and long-term memorability annotations. After an in-depth analysis of the dataset, we investigate various deep neural network-based models for the prediction of video memorability. Our best model using a ranking loss achieves a Spearman's rank correlation of 0.494 *(respectively* 0.256*) for short-term (resp. long-term) memorability prediction, while our model with attention mechanism provides insights of what makes a content memorable. The VideoMem dataset with pre-extracted features is publicly available*[1].

1. Introduction

While some contents have the power to burn themselves into our memories for a long time, others are quickly forgotten [17]. Evolution made our brain efficient to remember only the information relevant for our survival, reproduction, happiness, *etc*. This explains why, as humans, we share a strong tendency to memorize/forget the same images, which translates into a high human consistency in image memorability (IM) [20], and probably also a high consistency for video memorability (VM). Although, like for any other perceptual concept, we can observe individual differences while memorizing content, in this paper we target the capture and prediction of the part of the memorability that is shared by humans, as it can be assessed by averaging individual memory performances. This *shared-across-observers* part of the memorability, and especially long-term memorability, has a very broad application range in various areas including education and learning, content retrieval, search, filtering and summarizing, storytelling, *etc*.

The study of VM from a computer vision point of view is a new field of research, encouraged by the success of IM since the seminal work of Isola *et al.* [17]. In contrast to other cues of video importance, such as aesthetics, interestingness or emotions, memorability has the advantage of being clearly definable and objectively measurable (*i.e.,* using a measure that is not influenced by the observer's personal judgement). This certainly participates to the growing interest for its study. IM has initially been defined as the probability for an image to be recognized a few minutes after a single view, when presented amidst a stream of images [17]. This definition has been widely accepted within subsequent work [24, 21, 3, 20, 23]). The introduction of deep learning to address the challenge of IM prediction causes models to achieve results close to human consistency [20, 1, 34, 18, 31, 12]. As a result of this success, researchers have recently extended this challenge to videos [14, 30, 7, 5]. However, this new research field is nascent. As argued in [7], releasing a large-scale dataset for VM would highly contribute to launch this research field, as it was the case for the two important dataset releases in IM [17, 20]. Such a dataset should try to overcome the weaknesses of the previously released datasets. In particular, previous research on IM focused on the measurement of memory performances only a few minutes after memorization. However, passage of time is a factor well-studied in psychology for its influence on memory, while having been

[1] https://www.technicolor.com/dream/research-innovation/video-memorability-dataset

largely ignored by previous work on IM, probably because of the difficulty to collect long-term annotations at a large scale, in comparison with short-term ones. Measuring a memory performance a few minutes after the encoding step is already a measure a long-term memory, since short-term memory usually lasts less than a minute for unrehearsed information [28]. However, memories continue to change over time: going through a consolidation process (*i.e.,* the time-dependent process that creates our lasting memories), some memories are consolidated and others are not [25]. In other words, short-term memory performances might be poor predictors of longer term memory performances. In the following, we refer to measures of long-term memory a few minutes after memorization as measures of *short-term memorability*, and use the term *long-term memorability* for measures of long-term memory performance after one day. Since long-term memorability is more costly and difficult to collect than short-term memorability, it would nevertheless be interesting to know if the former can be inferred from the latter, which would also push forward our understanding of what makes a video *durably* memorable. A way to achieve this consists in measuring memorability for the same videos at two points of time. These two measures would be particularly interesting if spaced by a time interval in which forgetting is quite significant, to maximize the size of the potentially observable differences depending on the different video features. Observing the different forgetting curves in long-term memory (*e.g*. Ebbinghauss seminal work [9]), one can observe that the drop in long-term memory performance in recall follows an exponential decay and is particularly strong in the first hour, and to a lesser extent in the first day, immediately after the memorization. Measuring long-term memory a few minutes after encoding (as done in studies of IM [17, 20]), and again one day or more after (*i.e.,* to obtain a measure close to very long-term memory), sounds therefore a good trade-off.

The main contributions of this work are fivefold:

- We introduce a new protocol to objectively measure human memory of videos at two points of time (a few minutes and then 24-72 hours after memorization) and release VideoMem, the premier large-scale dataset for VM, composed of 10,000 videos with short-term and long-term memorability scores (Sections 3.1 and 3.2).
- Through an analysis of the dataset, we address the problem of understanding VM, by highlighting some factors involved in VM (Section 4).
- We benchmark several video-based DNN models for VM prediction (Section 5.2) against image-based baseline models (Section 5.1).
- We prove that, similarly to IM, semantics is highly relevant for VM prediction, through the study of a state-of-the-art image-captioning model (Section 5.3). This best model reaches a performance of 0.494 for Spearman's rank correlation on VideoMem for short-term memorability and 0.256 for long-term memorability.
- We propose an extension of the best performing model with an attention mechanism to localize what in an image makes it memorable (Section 5.5).

2. Related work

If long-term memory has been studied for over a century in psychology, since the seminal experimental studies of Ebbinghaus [10], its study from a computer vision point of view started quite recently, with [17]. Images and videos had long been used as material to assess memory performances [32, 2, 13], proving that human possessess an extensive long-term visual memory. The knowledge accumulated in psychology helped to measure memory using classical memory tests (see [29] for an extensive overview) such as recognition tests [17, 20, 14, 7] or textual question-based recall surveys [30]. Several factors are highlighted in the psychological literature for their critical influence on long-term memory, including emotion [19], attention [8], semantics [27], several demographic factors [6], memory re-evocation [26], or passage of time [25], also providing computer vision researchers with insights to craft valuable computational features for IM and VM prediction [24, 16, 7].

Focusing on IM in computer vision, most studies made use of one of the two available large datasets, specifically designed for IM prediction, where IM was measured a few minutes after memorization [17, 20], and consequently focused on predicting a so-called short-term IM [24, 21, 3, 20, 1, 23, 31, 12]. The pioneering work of [17] focused primarily on building computational models to predict IM from low-level visual features [17], and showed that IM can be predicted to a certain extent. Several characteristics have also been found to be relevant for predicting memorability in subsequent work, for example saliency [24], interestingness and aesthetics [16], or emotions [20]. The best results were finally obtained by using fine-tuned or pre-extracted deep features, which outperformed all other features [20, 1, 31, 12], with models achieving a Spearman's rank correlation near human consistency (*i.e.,* .68) when measured for the ground truth collected in [17, 20].

VM study is more recent. To the best of our knowledge, there exist only three previous attempts at measuring it [14, 30, 7]. Inspired by [17], Han *et al.* built a similar but far much heavier protocol to measure VM: the long time span of the experiment makes the generalization of this protocol difficult, in particular if one targets the construction of an extensive dataset. Another approach uses questions instead of a classic visual recognition task to measure VM [30]. As a results, memorability annotations collected for the videos may reflect not only the differences in memory performances but also the differences of complexity between the questions, especially since the authors use

the response time to calculate memorability scores, which might critically depend on the questions' complexity. The most recent attempt at measuring VM, and the only one, to our knowledge, resulting in a publicly available dataset, comes from [7]. The authors introduced a novel protocol to measure memory performance after a significant retention period – *i.e.,* weeks to years after memorization – without needing a longitudinal study. In contrast with previous work, the annotators did not pass through a learning task. It was replaced with a questionnaire designed to collect information about the participants' prior memory of Hollywood-like movies. However, such a protocol implies a limited choice of content: authors needed contents broadly disseminated among the population surveyed, as the participants should have seen some of them before the task (hence the Hollywood-like movies), leading to a number of annotations biased towards most famous content. Furthermore, the absence of control of the memorizing process and the answers of the questionnaire based on subjective judgments make the measure of memory performance not fully objective. To sum up, none of the previous approaches to measure VM is adapted to build a large-scale dataset with a ground truth based on objective measures of memory performance. Results obtained for VM prediction are yet far from those obtained in IM prediction. Han *et al.* proposed a method which combines audio-visual and fMRI-derived features supposedly conveying part of the brain activity when memorizing videos, which in the end enables to predict VM without the use of fMRI scans [14]. However, the method would be difficult to generalize. Shekhar *et al.* investigated several features, including C3D, semantic features obtained from some video captioning process, saliency features, dense trajectories, and color features, before building their memorability predictor [30]. They found that the best feature combination used dense trajectories, captioning, saliency and color features.

3. VideoMem: large-scale video memorability dataset

In Section 3.1, we describe the collection of source videos that compose the VideoMem dataset. We then introduce a new protocol to collect short-term and long-term memorability annotations for videos (Section 3.2), before explaining the computation of VM scores (Section 3.3).

3.1. Video collection

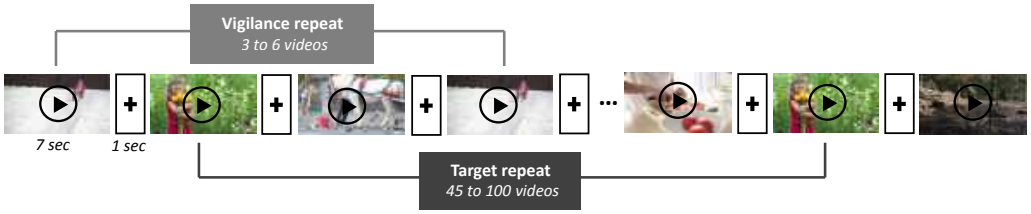
The dataset is composed of 10,000 soundless videos of 7 seconds shared under a license that allows their use and redistribution for research purpose only. In contrast to previous work on VM, where videos came from TRECVID [30, 14] or were extracted from Hollywood-like movies [7], videos in our dataset were extracted from raw footage, mainly from staged settings, dedicated to be further edited by professionals when creating new content, *e.g.* a new motion picture, video clip, television show, advertisements, *etc*. Because such video footage is typically used to save shooting new material, it is usually generic enough to be easily integrated in different sorts of creations. As such, they are context-independent and contain only one semantic scene. By this choice of content, we expect these basic building units to be relevant to train models which generalize on other types of videos. We are also confident that observers never saw the videos before participating in the experiment. Videos are varied and contain different scene types such as animal, food and beverages, nature, people, transportation, *etc*. A few of them contain similarities, *e.g.* same actor, same place but slightly different action, as it is the case in everyday video consumption ($< 1\%$). A small fraction is also slow-motion. Each video comes with its original title, that can often be seen as a list of tags (textual metadata). Example video keyframes are shown in Fig. 1.

The original videos are of high quality (HD or 4k) and of various durations (from seconds to minutes). As it will be described in Section 3.2, our protocol relies on crowdsourcing. For the sake of fluency during the annotation collection and consistency between the videos, we rescaled the videos to HD and re-encoded them in *.webm* format, with a bitrate of 3,000 kbps for 24 fps. To satisfy to the protocol's constraints, *i.e.,* minimal delay before measuring memory performance and maximal duration of the tasks to avoid user fatigue, we also cut the videos to keep only the first 7 seconds. Most videos are short ($<$ a few minutes) and contain one semantic scene. Those 7 seconds should therefore be representative of their content. Videos are soundless, firstly because a large part of the original data came without audio, and secondly, because it is difficult to control the audio modality in crowdsourcing. Accordingly, memorability would be linked only to the visualization of a semantic unit, which sounds a reasonable step forward for VM prediction, without adding a potentially biasing dimension.

Figure 1: Example keyframes from videos of VideoMem, sorted by decreasing long-term memorability (from left to right, and top to bottom).

3.2. Annotation protocol

To collect VM annotations, we introduced a new protocol which enables to measure both human short-term and



(a) Step #1. Interlaced encoding and recognition tasks.

(b) Step #2. Second recognition task after 24 to 72 hours.

Figure 2: Proposed protocol to collect both short-term and long-term video memorability annotations. The second recognition task measures memory of videos viewed as fillers during step #1, to collect long-term memorability annotations.

long-term memory performances. Inspired by what was proposed in [16, 17] for IM, we also used recognition tests for our memorability scores to reflect objective measures of memory performance. However, our protocol differs in several ways, not mentioning the fact that it is dedicated to videos. Firstly, as videos have an inherent duration, we had to revise 1) the delay between the memorization of a video and its recognition test and 2) the number of videos, for the task not be too easy. Secondly, in contrast to previous work on IM, where memorability was measured only a few minutes after memorization, memory performance is measured twice to collect both short-term and long-term memorability annotations: a few minutes after memorization and again (on different items) 24-72 hours later. The retention interval between memorization and measure is not as important as in [7], where it lasts weeks to years. As previously explained, we hope, however, that this measure reflects very-long term memory performance instead of short-term memory, as forgetting happens to a large extent during the first day following the memorization.

Fig. 2 illustrates our protocol, that works in two steps. Step #1, intended to collect short-term annotations, consists of interlaced viewing and recognition tasks. Participants watch a series of videos, some of them – the *targets* – repeated after a few minutes. Their task is to press the space bar whenever they recognize a video. Once the space bar is pressed, the next video is displayed, otherwise current video goes on up to its end. Each participant watches 180 videos, that contain 40 *targets*, repeated once for memory testing, and 80 *fillers* (*i.e.,* non target videos), 20 of which (so-called *vigilance fillers*) are also repeated quickly after their first occurrence to monitor the participant's attention to the task. The 120 videos (not counting the repetitions) that participate to step #1 are randomly selected among the 1000 videos that received less annotations at the time of the selection. Their order of presentation is randomly generated by following the given rule: the repetition of a *target* (respectively a *vigilance filler*) occurs randomly 45 to 100 (resp. 3 to 6) videos after the *target* (resp. *vigilance filler*) first occurrence. In the second step of the experiment, that takes place 24 to 72 hours after step #1, the same participants are proposed another similar recognition task, intended to collect long-term annotations. They watch a new sequence of 120 videos, composed of 80 *fillers* (randomly chosen totally new videos) and 40 *targets*, randomly selected from the *non-vigilance fillers* of step #1. Apart from the vigilance task (step #1 only), we added several controls, settled upon the results on an in-lab test: a minimum correct recognition rate (15%, step #2 only), a maximum false alarm rate (30%, step #1; 40%, step #2) and a false alarm rate lower than the recognition rate (step #2 only). This allows to obtain quality annotations by validating each user's participation; a participant could participate only once to the study. We recruited participants from diverse countries and origins via the Amazon Mechanical Turk (AMT) crowdsourcing platform.

3.3. Memorability score calculation

After a filtering of the participants to keep only those that passed the vigilance controls, we computed the final memorability scores on 9,402 participants for short-term, and 3,246 participants for long-term memorability. On average, a video was viewed as a repeated target 38 times (and at least 30 times) for the short-term task, and 13 times (at least 9 times) for the long-term task (this difference is inherent to the lower number of participants in step #2, as a large part of participants in step#1 did not come back). We assigned a first raw memorability score to each video, defined as the percentage of correct recognitions by participants, for both short-term and long-term memorability.

The short-term raw scores are further refined by applying

a linear transformation that takes into account the memory retention duration to correct the scores. Indeed, in our protocol, the repetition of a video happens after variable time intervals, *i.e.,* after 45 to 100 videos for a *target*. In [16], using a similar approach for images, it has been shown that memorability scores evolve as a function of the time interval between repeats while memorability ranks are largely conserved. We were able to prove the same relation for videos, *i.e.,* memorability decreases linearly when the retention duration increases (see Fig. 3, left). Thus, as in [20], we use this information to apply a linear correction (shown in Fig. 3) to our raw memorability scores to explicitly account for the difference in interval lengths, with the objective for our short-term memorability scores to be the most representative of the typical memory performance after the maximal interval (*i.e.,* 100 videos). Note that the applied correction has nevertheless little effect on the scores both in terms of absolute and relative values. Note also that we did not apply any correction for long-term memorability scores (Fig. 3, right). Indeed, we observed no specific, strong enough relationship between retention duration and long-term memorability. This was somehow expected from what can be found in the literature : according to our protocol, the second measure was carried out 24 to 72 hours after the first measure. After such a long retention duration, it is expected that the memory performance is no more subjected to substantial decrease due to the retention duration. In the end, the average short-term memorability score is 0.859 (instead of 0.875) and the average long-term memorability score is 0.778, all values showing a bias towards high values.

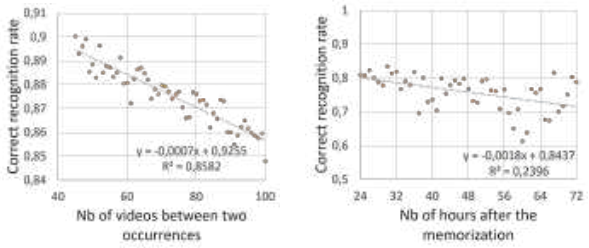
4. Understanding video memorability

4.1. Human consistency *vs.* annotation consistency

Following the method proposed in [16], we measured human consistency when assessing VM. For this purpose, we randomly split our participants into two groups of equal size (4,701 for short-term memorability, 1,623 for long-term memorability), and computed VM scores independently in each group as described in Section 3.3. We then calculated a Spearman's rank correlation between the two groups of scores. Averaging over 25 random half-split trials, an average Spearman's rank correlation, *i.e.,* a global human consistency, of 0.481 is observed for short-term memorability and of 0.192 for long-term memorability.

Such a method divides the number of annotations that is taken into account for the score computation at least by a factor of 2. Moreover, it may end with groups with unbalanced number of annotations per video as the split is randomly applied on the participants, not taking into account which videos they watched. For this reason, we proposed a new metric named *annotation consistency*, more representative of the performance consistency of the users. We reproduced the previous process of human consistency computation but on successive subparts of the dataset by considering for each sub-part only videos which received at least N annotations. Each subpart is then split in two groups of participants while ensuring a balance number of participants per video. By doing so, we obtain the annotation consistency as a function of the number of annotations per video, as presented in Fig. 4. This allows us to interpolate the following values: Annotation consistency reaches 0.616 (respectively 0.364) for the short-term (resp. long-term) task, for a number of annotations of 38 (resp. 13). Both values represent strong (resp. moderate) correlations according to the usual Spearman scale of interpretation. Hence, choosing larger mean number of annotations provides more stable annotations, *i.e.,* 0.616 (resp. 0.364) rather than 0.481 (resp. 0.192) for the short-term (resp. long-term) task.

The value of 0.616 for short-term memorability is to be compared to 0.68 for images as found in [20]. Slightly lower, VideoMem consistency was nevertheless obtained with less annotations than in [20], which is consistent with [7]. The maximum consistency is also slightly higher for VM than for IM (0.81 against 0.75 in [17] and 0.68 in [20]). An explanation is that videos contain more information than images and thus are more easily remembered. However, one should keep in mind that the protocols to collect annotations differ in several ways, making these results not fully comparable. Fig. 4 also shows that long-term and short-term consistencies follow the same evolution.



(a) Step #1. Recognition rate decreases linearly over time.

(b) No significant change in memory performance between 24 and 72 hours after memorization.

Figure 3: Mean correct recognition rate *vs.* the retention interval between the memorization and the measure of memory performance. Blue lines represent linear fitting.

4.2. Memorability consistency over time

In this study, we are interested in assessing how well memorability scores remain consistent over time, *i.e.,* if a video highly memorable after a few minutes of retention remains also highly memorable after 24 to 72 hours. The Spearman's rank correlation coefficient between the long-term and short-term memorability scores for

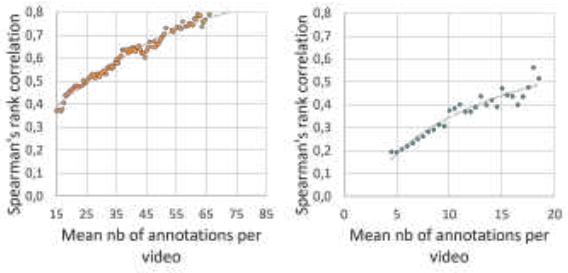


Figure 4: Annotation consistency *vs.* mean number of annotations per video (left: short-term, right: long-term).

the 10,000 videos exhibits a moderate positive correlation ($\rho = 0.305, p < .0001$) between the two variables, as also shown in Fig. 5. To discard a potential bias that would come from the highest number of annotations in step #1 compared to step #2, we computed the correlation for the 500 most annotated videos in the long-term task (that have at least 21 annotations) and then again for the 100 most annotated (at least 28 annotations), observing similar Spearman values of $\rho = 0.333$, $p < .0001$ and $\rho = 0.303$, $p < .0001$, respectively. This result suggests that memory evolves with time and in a non-homogeneous manner depending on the videos: a video highly memorable a few minutes after visualization might not remain highly memorable in long-term memory. This finding is consistent with the hypothesis we proposed in the introductory section, that the information important for a content to be memorized might not be the same for short-term and long-term memorization.

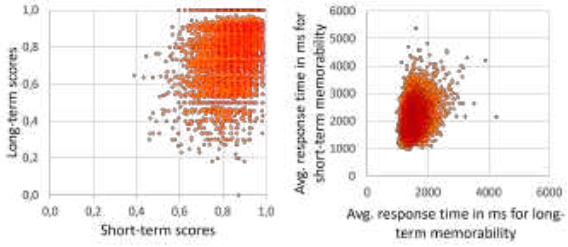


Figure 5: Short-term *vs.* long-term scores (left) and average response times (correct detections only) (right).

4.3. Memorability and response time

We observed negative Pearson correlations between the mean response time to correctly recognize targets and their memorability scores, both for short-term ($r = 0.307$, $p < .0001$) and long-term (0.176, $p < .0001$) memorability, as also illustrated in Fig. 6. This tends to prove that, globally, participants tended to answer more quickly for the most memorable videos than for the less memorable ones. This is consistent with [7], where the authors propose two explanations to this result: either the most memorable videos are also the most accessible in memory, and/or the most memorable videos contain more early recognizable elements than the less memorable ones. As videos in VideoMem consist of semantic units with often one unique shot – with most of the information already present from the beginning – the first explanation sounds more suitable here. This also suggests that participants tend to quickly answer after recognizing a repeated video (even though they did not receive any instruction to do so), maybe afraid of missing the time to answer, or to alleviate their mental charge. This correlation highlights that the average response time might be a useful feature to further infer VM in computational models.

The correlation is, however, lower for long-term memorability. One explanation might be that, after one day, remembering is more difficult. In connection with this explanation, we observed a significant difference between the mean response time to correctly recognize a video during step #1 and during step #2 ($1.43sec.$ *vs.* $3.37sec.$), as showed by a Student's t-test ($t(9999) = -122.59, p < 0001$). Note that the Pearson correlation (0.291) between average response time per video for short-term and long-term memorability is close to the Pearson correlation (0.329) observed between short-term and long-term memorability scores (see Fig. 5, right). Note that the mean response time for a false alarm was $3.17sec.$ for step #1 and $3.53sec.$ for step #2.

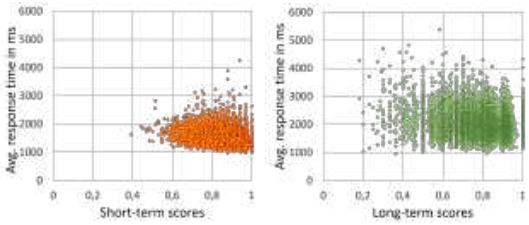


Figure 6: Average response time (correct recognitions only) as a function of memorability scores, for short-term (left) and long-term memorability (right).

5. Predicting video memorability

In this section we focus on predicting VM using various machine learning approaches. We pose the VM score prediction as a standard regression problem. We first benchmark several state-of-art video-based models on our data (Section 5.2), against performances of IM models (Section 5.1). We then focus on assessing how a very recent state-of-the-art image captioning based model, fine-tuned on our data, performs for VM prediction. The aim is here to see if the finding in [31, 7] that semantics highly intervenes in IM prediction still stands for VM prediction. In Section 5.4, we analyze the prediction results of all models and give insights to understand the correlation between IM and VM.

Last, in Section 5.5, we modify the advanced IC model by adding an attention mechanism that helps us better understand what makes a content memorable. Note that, for training (when applied) and evaluating the considered models, we split VideoMem dataset into training (6500 videos), validation (1500), and test (2000) sets, where the test set contains 500 videos with a greater number of annotations. Similarly to previous work in IM and VM, the prediction performance is evaluated in term of the Spearman's rank correlation between the ground truth and the predicted scores.

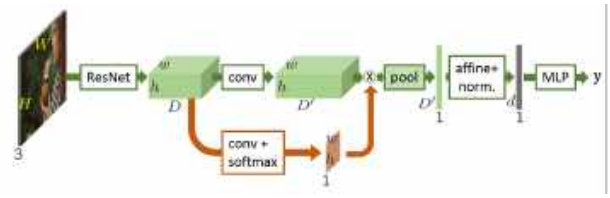


Figure 7: Semantic embedding model without (green pipeline) and with an attention mechanism (full workflow).

5.1. Image memorability-based baselines

In order to investigate the correlation between IM and VM and to build some first baselines on the dataset, we directly used two state-of-the-art models for IM prediction to compute a memorability score for 7 successive frames in the video (one per second): MemNet proposed in [20] and Squalli *et al.* in [31]. The final VM score for one video is obtained by averaging the 7 corresponding frame scores.

5.2. Video-based models

In a first attempt to capture the inherent temporal information of the videos, we investigated the performances of two classic, yet temporal, features: C3D [33] and HMP [4] as input features to some MLP layers. We tested them alone and concatenated, using some grid search for hyperparameter optimisation. Best results were obtained for the features alone, with the parameters: two hidden layers with 10 neurons for HMP and one hidden layer with 100 neurons for C3D, optimizer=IBLGS, activation=tanh, learning rate (lr)=1e-3. Second, instead of using a fix feature extractor, we directly fine-tuned the state-of-the-art ResNet3D model (based on ResNet34) [15]. For this, we replaced the last fully connected layer of ResNet3D by a new one dedicated to our considered regression task. This last layer was first trained alone for 5 epochs (Adam optimizer, batchsize=32, lr=1e-3), then the whole network was re-trained for more epochs (same parameters, but lr=1e-5).

5.3. Semantic embedding-based model

As scene semantic features derived from an image captioning system (IC) [22] have been shown to well characterize the memorability of images [31] and videos [7], we also investigated the use some IC system. Also, following the idea of model fine-tuning, we fine-tuned a state-of-art visual semantic embedding pipeline used for image captioning [11], on top of which a 2-layer MLP is added, to regress the feature space to a single memorability score. The overall architecture is shown in Fig. 7, in the green pipeline. As the model in [11] remains at the image-level, we first predict scores for the same 7 frames as in Section 5.1, then compute the final prediction at video level by averaging those 7 values. It is fine tuned on both VideoMem and LaMem [20] datasets, for short-term memorability only, because LaMem only provides short-term annotations. The training is done using the Adam optimizer and is divided in two steps: in the first 10 epochs only the weights of the MLP are updated while those of the IC feature extractor remain frozen. Later the whole model is fine-tuned. The learning rate is initialized to 0.001 and divided in half every three epochs. It is important to note that the original IC model was trained with a new ranking loss (*i.e.,* Spearman surrogate) proposed in [11]. This new loss has proved to be highly efficient for ranking tasks as claimed in [11]. For the fine-tuning however, the training starts with a ℓ_1 loss as initialization step, before coming back to the ranking loss. The reason is that the original model was indeed trained for scores in [-1;1], while our memorability scores are in [0;1]. Thus the ℓ_1 loss forces the model to adapt to this new range.

5.4. Prediction results

From the results in Table 1, we may draw several conclusions. Additional results are presented in the supplementary material. First, it is possible to achieve already quite good results in VM prediction using models designed for IM prediction. This means that the memorability of a video is correlated to some extent with the memorability of its constituent frames. For both C3D and HMP-based models, it seems that the simple MLP layers put on top of those features did not successfully capture the memorability. This might be explained by the fact that most of the videos contain no or little motion (62%), whereas 11% only contain high motion. However, the comparison between short-term and long-term performances exhibits some interesting information: HMP performs better than C3D for short-term and the inverse is true for long-term, as if direct motion information was more relevant for short-term than for long-term memorability. This is a first finding on what distinguishes the two notions. Also, the two fine-tuned models, dedicated to the task, show significantly higher performances. The fine-tuned ResNet3D, although purely video-based, is exceeded by the fine-tuned semantic embedding-based model. However, for the latter, data augmentation was performed using the LaMem dataset [20], which was not possible for the former as LaMem only contains image memorability information. This indeed biases the comparison between

Models	short-term memorability			long-term memorability		
	validation	test	test (500)	validation	test	test (500)
MemNet (Sec. 5.1)	0.397	0.385	0.426	0.195	0.168	0.213
Squalli *et al.* (Sec. 5.1)	0.401	0.398	0.424	0.201	0.182	0.232
C3D (Sec. 5.2)	0.319	0.322	0.331	0.175	0.154	0.158
HMP (Sec. 5.2)	0.469	0.314	0.398	0.222	0.129	0.134
ResNet3D (Sec. 5.2)	0.508	0.462	0.535	0.23	0.191	0.202
Semantic embedding model (Sec. 5.3)	**0.503**	**0.494**	**0.565**	**0.26**	**0.256**	**0.275**

Table 1: Results in terms of Spearman's rank correlation between predicted and ground truth memorability scores, on the validation and test sets, and on the 500 most annotated videos of the dataset (test (500)) that were placed in the test set.

the two models, but current results still show that, as expected, leveraging both a dedicated fine-tuning and the use of high level semantic information from some image captioning system, gives an already quite high prediction performance. For all models, we note that performances were lower for long-term memorability. One interpretation might be that the memorability scores for long-term are based on a smaller number of annotations than for short-term, so they probably capture a smaller part of the intrinsic memorability. However, it may also highlight the difference between short-term and long-term memorability, the latter being more difficult to predict as it is more subjective, while both being still – though not perfectly – correlated. The performances of our models on the 500 most annotated videos are better. This reveals that our dataset might benefit from a larger number of annotations. Last, compared to annotation consistency values, performances remain lower, showing that there is still room for improvement.

5.5. Intra-memorability visualization

To better understand what makes a video frame memorable, we added an attention mechanism to our best model. It will then learn what regions in each frame contribute more to the prediction. For this purpose, a convolutional layer is added in parallel with the last convolutional layer of the feature extractor part. It outputs a 2D attention map which goes through a softmax layer and is multiplied with the last convolution map of the visual pipeline as shown in Fig. 7 (orange branch). An empirical study of the resulting attention maps tends to separate them in two categories. In the first one, when image frames contain roughly one main object and no or rare information apart from this main object (this might be because the background is dark or uniform), it seems that the model focuses, as expected intuitively, on the main object and even, in the case of large enough faces, on details of the faces, as if trying to remember the specific features of faces. Example results for images in the first category can be found in Fig. 8, first row. In the second category that groups all other frames, with several main and secondary objects, cluttered background, *etc.,* it seems on the contrary that the model focuses on all but the main objects/subjects of the images, as if trying to remember little details that will help it differentiate the image from another similar one. Or said differently, the second category shows results that might be interpreted as a second memorization process, once the first one – focusing on the main object – is already achieved. Examples for the second category can be found in the second row of Fig. 8. More results and insights are given in the supplementary material.

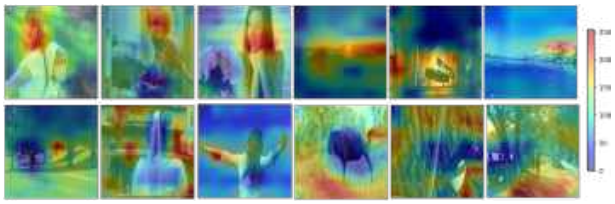


Figure 8: Visualization of the attention mechanism's output. The model focuses either on close enough faces or main objects when the background is dark or uniform (row #1), or it focuses on details outside the main objects (row #2).

6. Conclusions

In this work, we presented a novel memory game based protocol to build VideoMem, a premier large-scale VM dataset. Through an in-depth analysis of the dataset, we highlighted several important factors concerning the understanding of VM: human *vs.* annotation consistency, memorability over time, and memorability *vs.* response time. We then investigated various models for VM prediction. Our proposed model with *spatial* attention mechanism allows to visualize, and thus better understand what type of visual content is more memorable. Future work would be devoted to further study the differences between short-term and long-term memorability, and improve prediction results with a particular focus on temporal aspects of the video, *e.g.* by adding *temporal* attention model and recurrent neural network blocks to the workflow.

References

[1] Yoann Baveye, Romain Cohendet, Matthieu Perreira Da Silva, and Patrick Le Callet. Deep learning for image memorability prediction: the emotional bias. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 491–495, 2016.

[2] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008.

[3] Bora Celikkale, Aykut Erdem, and Erkut Erdem. Visual attention-driven spatial pooling for image memorability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 976–983, 2013.

[4] Arridhana Ciptadi, Matthew S Goodwin, and James M Rehg. Movement pattern histogram for action recognition and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 695–710. Springer, 2014.

[5] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q. K. Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. Mediaeval 2018: Predicting media memorability task. In *Proceedings of the MediaEval Workshop*, 2018.

[6] Romain Cohendet, Anne-Laure Gilet, Matthieu Perreira Da Silva, and Patrick Le Callet. Using individual data to characterize emotional user experience and its memorability: Focus on gender factor. In *Proceedings of the IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2016.

[7] Romain Cohendet, Karthik Yadati, Ngoc Q. K. Duong, and Claire-Hélène Demarty. Annotating, understanding, and predicting long-term video memorability. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, pages 11–14, 2018.

[8] Nelson Cowan. *Attention and memory: An integrated framework*. Oxford University Press, 1998.

[9] Hermann Ebbinghaus. *Memory; a contribution to experimental psychology*. New York city, Teachers college, Columbia university, 1913.

[10] Hermann Ebbinghaus. *Memory: a contribution to experimental psychology*. Number 3. University Microfilms, 1913.

[11] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Sodeep: a sorting deep net to learn ranking loss surrogates. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019.

[12] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6363–6372, 2018.

[13] Orit Furman, Nimrod Dorfman, Uri Hasson, Lila Davachi, and Yadin Dudai. They saw a movie: long-term memory for an extended audiovisual narrative. *Learning & memory*, 14(6):457–467, 2007.

[14] Junwei Han, Changyuan Chen, Ling Shao, Xintao Hu, Jungong Han, and Tianming Liu. Learning computational models of video memorability from fmri brain imaging. *IEEE Transactions on Cybernetics*, 45(8):1692–1703, 2015.

[15] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? *arXiv preprint*, arXiv:1711.09577, 2017.

[16] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1469–1482, 2014.

[17] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 145–152. IEEE, 2011.

[18] Peiguang Jing, Yuting Su, Liqiang Nie, and Huimin Gu. Predicting image memorability through adaptive transfer learning from external sources. *IEEE Transactions on Multimedia*, 19(5):1050–1062, 2017.

[19] Elizabeth A Kensinger and Daniel L Schacter. Memory and emotion. *Handbook of emotions*, 3:601–617, 2008.

[20] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2390–2398, 2015.

[21] Jongpil Kim, Sejong Yoon, and Vladimir Pavlovic. Relative spatial features for image memorability. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pages 761–764, 2013.

[22] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

[23] Souad Lahrache, Rajae El Ouazzani, and Abderrahim El Qadi. Bag-of-features for image memorability evaluation. *IET Computer Vision*, 10(6):577–584, 2016.

[24] Matei Mancas and Olivier Le Meur. Memorability of natural scenes: The role of attention. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 196–200, 2013.

[25] James L McGaugh. Memory–a century of consolidation. *Science*, 287(5451):248–251, 2000.

[26] Lynn Nadel and Morris Moscovitch. Memory consolidation, retrograde amnesia and the hippocampal complex. *Current opinion in neurobiology*, 7(2):217–227, 1997.

[27] M Ross Quillan. Semantic memory. Technical report, Bolt Beranek and Newman Inc Cambridge MA, 1966.

[28] Russell Revlin. *Cognition: Theory and Practice*. Palgrave Macmillan, July 2012.

[29] Alan Richardson-Klavehn and Robert A Bjork. Measures of memory. *Annual review of psychology*, 39(1):475–543, 1988.

[30] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. Show and recall: Learning what makes videos memorable. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2730–2739, 2017.

[31] Hammad Squalli-Houssaini, Ngoc Q. K. Duong, Gwenaëlle Marquant, and Claire-Hélène Demarty. Deep learning for predicting image memorability. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2371–2375, 2018.

[32] Lionel Standing. Learning 10000 pictures. *Quarterly Journal of Experimental Psychology*, 25(2):207–222, 1973.

[33] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the International Conference on Computer Vision (CVPR)*, pages 4489–4497, 2015.

[34] Soodabeh Zarezadeh, Mehdi Rezaeian, and Mohammad Taghi Sadeghi. Image memorability prediction using deep features. In *Proceedings of the IEEE International Conference on Electrical Engineering (ICEE)*, pages 2176–2181, 2017.

D. PAPER 4: VIDEOMEM: CONSTRUCTING, ANALYZING, PREDICTING SHORT-TERM AND LONG-TERM VIDEO MEMORABILITY

Appendix E

Curriculum Vitae

Curriculum Vitae

Rémi, Quang-Khanh-Ngoc DUONG,
Senior Scientist, InterDigital R&D France
Vietnamese and French citizen
Languages: Vietnamese (native), French and English (fluent), Korean (beginner)
Website: Google Scholar page
Email: Quang-Khanh-Ngoc.Duong@interdigital.com

DEGREES

PhD in Signal and Telecommunications November 2008 – October 2011
University of Rennes 1 and INRIA, France
Supervisors: Rémi Gribonval and Emmanuel Vincent

Master of Engineering March 2006 – January 2008
Dept. of Electronic Engineering, Paichai University, Korea
GPA: ~4.39/4.5; Highest Honor MSc Degree

Bachelor of Engineering September 2000 – December 2004
Posts and Telecommunications Institute of Technology, Vietnam
GPA: 8.12/10; Highest Honor BSc Degree

POSITIONS HELD

Senior Scientist at InterDigital R&D France June 2019 – present

Senior Scientist at Technicolor R&D France June 2015 – May 2019

Researcher at Technicolor R&D France May 2013 – May 2015

Postdoc at Technicolor R&D France November 2011 – April 2013

Preparation of PhD at INRIA, Rennes, France November 2008 – October 2011

Research Engineer at Emersys Corp., Korea March 2008 – September 2008

Research Assistance at Media Communication March 2006 – January 2008
& Signal Processing Lab, Paichai University, Korea

System Engineer at Visco JSC, Vietnam August 2004 – February 2006

MAJOR SCIENTIFIC ACHIEVEMENTS

- (Co)author of **2 book chapters**, **9 journal papers** and more than **40 international conference/workshop papers** with over 1400 citations according Google Scholar.
- (Co)inventor of about **30 patent applications** (granted and pending) since 2012.
- **Senior Member** of IEEE since 2017.

- **Associate Editor** of the IEICE Transaction on Information and Systems (Japan) since 2016.
- **Associate Member** of the Technicolor's Fellowship Network ("*Fellowship Network members are involved in multiple activities: evaluate inventions and patents; contribute to technology related strategic decisions; leverage Technicolor scientific and technical excellence; help develop a technology and science culture, offering seminars, training and mentoring to more junior technologists; promote the Technicolor leadership in science and technology"*).

DISTINCTIONS

- **Technicolor's inventor recognition**, 2016 (for the contributions in the Invention's porfolio).
- **Bretagne Young Researcher Award (mention spéciale)**, 2015 (for the contribution in the Science & Technology for Social Development - catégorie Évolutions sociales et sociétales).
- **IEEE SPS Young Author Best Paper Award**, awarded by the IEEE Signal Processing Society, Vancouver, Canada - May 2013 ("*honors the author(s) of an especially meritorious paper dealing with a subject related to the Society's technical scope and appearing in one of the Society's solely owned transactions or the Journal of Selected Topics in Signal Processing and who, upon the date of submission of the paper, is less than 30 years of age"*).
- **2nd PhD thesis prize**, awarded by the Fondation Rennes 1, 2012 ("*Les prix de thèse seront décernés aux travaux présentant les plus forts potentiels d'innovation et/ou de transfert de technologie et auront été évalués par un jury composé d'universitaires et de responsables d'entreprises"*).

PUBLIC RESPONSIBILITIES

- **Technical program committee** of major signal processing conferences for years: ICASSP (2012-2020), EUSIPCO (2011-2019), ICMR 2018, ICCE-Berlin (2012-2015), WASPAA (2011-2017).
- **Co-founder and co-organizer** of the MediaEval benchmark Predicting Video Memorability task (running in 2018 and 2019).
- **Co-founder and co-organizer** of the MediaEval Predicting Media Interestingness task (running in 2016 and 2017).
- **Session chair, special session organizer, and panelist at several conferences:** LVA/ICA, ICMR 2017, ICMR 2018, ICASSP 2018.
- **Organizing member of the scientific seminars** at Technicolor/InterDigital since 2013.
- **Member of the organizing commitee**: 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), 2010, France; and Second international community-based Signal Separation Evaluation Campaign (SiSEC 2010).
- **Journal reviews**: IEEE/ACM Transactions on Audio Speech and Language Processing, IEEE Transactions on Signal Processing, IEEE Transaction on Multimedia, Speech Communication (Elsevier), Signal Processing (Springer), IEEE Singnal Processing Letter.

COLLABORATIVE PROJECTS

- **MAD** (French ANR project on Missing Audio Data Inpainting): developed new audio inpainting concepts and algorithms relying on signal processing and machine learning (completed).
- **VERSAMU**S (INRIA Associate Team with the SOno lab, University of Tokyo): integrated probabilistic music representation for versatile music content processing (completed).

- **i3DMUSIC** (EUREKA Eurostars project with Audionamix SA, Sonic Emotion AG and EPFL): 3D music rendering and remixing platform (completed).
- **AI4Media** (EU H2020 project): Artificial Intelligence for the Society and the Media Industry, started Sept. 2020.

RESEARCH SUPERVISION

Postdoc:

1. **Romain Cohendet** (2017-2018), "Constructing, analyzing, and predicting video memorability," (co-advised with Claire-Hélène Demarty).
2. **Suresh Kirthi Kumaraswamy** (since Aug. 2020), "Flexible and distributed DNN models," (co-advised with A. Ozerov).

PhD students:

1. **Sanjeel Parekh** (2016 - 2019), "Learning Representations for Robust Audio-Visual Scene Analysis," (co-advised with S. Essid, A. Ozerov, P. Pérez, and G. Richard).
2. **Hien-Thanh Duong** (2015-2019), "Audio source separation exploiting NMF-based generic source spectral model" (informally co-advised with C. Nguyen and P. Nguyen).
3. **Karthik Yadati** (2017), three months PhD internship at Technicolor (co-advised with C.-H. Demarty).
4. **Deniz Engin** (since Sept. 2020), « Semantic Multimodal Question Answering", (co-advised with F. Schnitzler and Y. Avrithis)

Master students:

1. **Luc Le Magoarou** (2013) "Text-informed audio source separation" (co-advised with A. Ozerov) Technicolor Best Internship Award (annual evaluation among 30 to 40 interns).
2. **Dalia El Badawy** (2014) "On-the-fly audio source separation" (co-advised with A. Ozerov) Technicolor Best Internship Award (annual evaluation among 30 to 40 interns).
3. **Pierre Prablanc** (2015) "Voice conversion for speech inpainting" (co-advised with A. Ozerov and P. Pérez) running for Technicolor Best Internship Award (top 3).
4. **Gustavo Sena Mafra** (2015) "Acoustic scene classification with deep neural network" (co-advised with A. Ozerov and P. Pérez).
5. **Swann Leduc** (2016) "Cross-lingual voice-conversion" (co-advised with A. Ozerov and P. Pérez).
6. **Yuesong Shen** (2016) "Deep learning for predicting media interestingness" (co-advised with C.-H. Demarty) Ecole Polytechnique Best Internship Award.
7. **Eric Grinstein** (2017) "Audio manipulation with deep representations" (co-advised with A. Ozerov and P. Pérez).
8. **Eloise Berson** (2017) "Understanding and predicting socially-driven media interestingness" (co-advised with C.-H. Demarty).
9. **Hammad Squalli-Houssaini** (2017) "Understanding and predicting image memorability" (co-advised with C.-H. Demarty).
10. **Van-Huy Vo** (2017) "Audio inpainting" (informally co-advised with P. Pérez) Technicolor Best Internship Award (annual evaluation among 30 to 40 interns).

11. **Ha-Quang Le** (2018) "Audio attributes modification with deep representations" (co-advised with A. Ozerov and G. Puy).
12. **Antoine Caillon** (2019) "Speech transformations using deep generative models" (co-advised with A. Ozerov and G. Puy).
13. **Valentin Bilot** (2019) "Audio event classification via multiple instance learning" (co-advised with A. Ozerov).
14. **Viet-Dao Nguyen** (2020) "Network distillation for flexible machine learning models" (co-advised with A. Ozerov).

PUBLICATIONS

BOOK CHAPTER:

[2] S. Essid, S. Parekh, N. Q. K. Duong, R. Serizel, A. Ozerov, F. Antonacci, and A. Sarti, "*Multiview approaches to event detection and scene analysis*," In Computational Analysis of Sound Scenes and Events, Springer, 2018

[1] C-H. Demarty, M. Sjöberg, G. Constantin, N. Q. K. Duong, B. Ionescu, T. Do, and H. Wang, "*Predicting Interestingness of Visual Content*," In Visual Content Indexing and Retrieval with Psycho-Visual Models, Springer, 2017

JOURNAL:

[9] M. Costantin, L. Stefan, B. Ionescu, N. Q. K. Duong, C-H. Demarty, and M. Sjoberg "*Visual Interestingness Prediction: A Benchmark Framework and Literature Review*," Int. Journal in Computer Vision (IJCV), 2020, in revision.

[8] S. Parekh, S. Essid, A. Ozerovy, N. Q. K. Duong, P. Perez, and G. Richard, "*Weakly Supervised Representation Learning for Audio-Visual Scene Analysis*," IEEE/ACM Transaction on Audio, Speech and Language Processing, 2019.

[7] H-T. Duong, N. Q. K. Duong, and C. Nguyen, *"Gaussian modeling-based multichannel audio source separation exploiting generic source spectral model,"* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019

[6] D. Badawy, N. Q. K. Duong, and A. Ozerov, "*On-the-fly audio source separation - a greatly simplified user-guided approach*," IEEE/ACM Transaction on Audio, Speech and Language Processing, 2017, (Technicolor Best Internship Award).

[5] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong, **"***Text-informed audio source separation. Example-based approach using non-negative matrix partial co-factorization*," Springer Journal of Signal Processing Systems, 2014 (Technicolor Best Internship Award).

[4] N. Q. K. Duong**,** E. Vincent and R. Gribonval, *Spatial location priors for Gaussian model-based reverberant audio source separation*, EURASIP journal on Advanced Signal Processing, 2013.

[3] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutte, N. Q. K. Duong, *The Signal Separation Campaign (2007-2010): chievements and Remaining Challenges*, Signal Processing, Elsevier, 2012.

[2] N. Q. K. Duong, E. Vincent and R. Gribonval, *Under-determined reverberant audio source separation using a full-rank spatial covariance model*, IEEE Transactions on Audio, Speech and Language Processing, Special Issue on Processing Reverberant Speech, Vol. 18, No. 7, pp. 1830-1840, Sep. 2010. (IEEE Signal Processing Society's Young Author Best Paper Award)

[1] N. Q. K. Duong, C Park, and SH Nam, "*Application of block on-line blind source separation to acoustic echo cancellation*," *The Journal of the Acoustical Society of Korea*, No. 27, Vol 1E, pp. 17-24, 2008.

INTERNATIONAL CONFERENCES & WORKSHOPS:

[43] H. Phan, L. Pham, P. Koch, N. Q. K. Duong, I. McLoughlin, A. Mertin, "Audio Event Detection and Localization with Multitask Regression Network", Proc. DCASE, 2020

[42] R. Cohendet, C.H. Demarty, and N. Q. K. Duong, "*VideoMem: Constructing, Analyzing, Predicting Short-term and Long-term Video Memorability," Proc. I*nt. Conf. On Computer Vision (ICCV), 2019

[41] Cotatin et al., "*The Predicting Media Memorability Task at MediaEval 2019*", Proc. MediaEval Benchmarking Initiative for Multimedia Evaluation, 2019 [orale presentation].

[40] T. H. Le, P. Gilberton, and N. Q. K. Duong, "*Discriminate natural versus loudspeaker emitted speech," Proc. ICASSP, 2019*

[39] S. Parekh, A. Ozerovy, S. Essid, N. Q. K. Duong, P. Perez, and G. Richard, "*Identify, locate and separate: Audio-visual object extraction in large video collections using weak supervision*," Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2019 [orale presentation].

[38] H. Vo, N. Q. K. Duong, and P. Perez, "*Structural inpainting," Proc. ACM Multimedia* conference, 2018, (Technicolor Best Internship Award) [orale presentation].

[37] R. Cohendet, K. Yadati, N. Q. K. Duong, and C-H Demarty, "*Annotating, Understanding, and Predicting Long-term Video Memorability*," Proc. ACM on International Conference on Multimedia Retrieval (ICMR), 2018 [orale presentation].

[36] R. Cohendet, C-H Demarty, and N. Q. K. Duong, "*Transfer learning for video memorability prediction*", Proc. MediaEval Benchmarking Initiative for Multimedia Evaluation, 2018

[35] Cohendet et al., "*MediaEval 2017 predicting video memorability task*," Proc. MediaEval Benchmarking Initiative for Multimedia Evaluation, 2018 [orale presentation].

[34] E. Grinstein, N. Q. K. Duong, A. Ozerov, and P. Perez, "*Audio style transfer," Proc. IEEE* International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018

[33] T. Duong, N. Q. K. Duong, and C. Nguyen, "*Multichannel speech enhancement exploiting NMF-based generic source spectral model," Proc. LVA/ICA 2018*

[32] H. Squalli-Houssaini, N. Q. K. Duong, G. Marquant , and C.-H. Demarty, "*Deep learning for predicting image memorability*," Proc. ICASSP 2018 [orale presentation].

[31] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Perez, and G. Richard, "*Weakly supervised representation learning for unsynchronized audio-visual events*," Proc.CVPR Workshop on Sight and Sound, 2019.

[30] C-H Demarty et al., "*MediaEval 2017 predicting media interestingness task*," Proc. MediaEval Benchmarking Initiative for Multimedia Evaluation, Ireland, 2017.

[29] E. Berson, C-H. Demarty, and N .Q. K. Duong, "*Multimodality and Deep Learning when predicting Media Interestingness*," Proc. MediaEval Benchmarking Initiative for Multimedia Evaluation, Ireland, 2017 [orale presentation].

[28] Y. Shen, C-H. Demarty, and N. Q. K. Duong, "*Deep learning for multimodal-based video interestingness prediction*", Proc. IEEE Int. Conf. On Multimedia and Expo (ICME), Hong Kong, 2017**.** (Ecole Polytechnique Best Internship Award)

[27] G. Puy, A. Ozerov, N. Q. K. Duong, and P. Perez, "*Informed source separation by compressive graph signal processing*," Proc. 41st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), USA, 2017 [orale presentation].

[26] S. Parekhy, S. Essid, A. Ozerovy, N. Q. K. Duong, P. Perez, and G. Richard, "*Motion informed audio source separation*," Proc. 41st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), USA, 2017 [orale presentation].

[25] P. Prablanc, A. Ozerov, N. Q. K. Duong, and P. Pérez, "*Text-informed speech inpainting via voice conversion*," Proc. European Signal Processing Conference (EUSIPCO), Budapest, Hungary, 2016, running for Technicolor Best Internship Award (top 3) [orale presentation].

[24] G. Mafra, N. Q. K. Duong, A. Ozerov, and P. Pérez, "*Acoustic scene classification: An evaluation of an extremely compact feature representation*", Proc. Int. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), Budapest, Hungary, 2016

[23] Y. Shen, C-H. Demarty, and N .Q. K. Duong, "*Technicolor@ MediaEval 2016 Predicting Media Interestingness Task*", Proc. MediaEval Benchmarking Initiative for Multimedia Evaluation, Neitherland, 2016.

[22] C-H Demarty et al., "*Mediaeval 2016 predicting media interestingness task*," Proc. MediaEval Benchmarking Initiative for Multimedia Evaluation, Neitherland, 2016 [orale presentation].

[21] H. T. Duong, C. Nguyen, P. Nguyen, and N. Q. K. Duong, "*Speech enhancement based on nonnegative matrix factorization with mixed group sparsity constraint*," Proc. 15th IEEE Int. Conf. on Electronics, Information, and Communications (ICEIC), 2016 [orale presentation].

[20] H. T. Duong, C. Nguyen, P. Nguyen, H. Tran, and N. Q. K. Duong, "*Speech enhancement based on nonnegative matrix factorization with mixed group sparsity constraint*," Proc. 6th ACM Int. Conf. on Information and Communication Technology (SoICT), 2015 [orale presentation].

[19] N. Q. K. Duong and H. T. Duong, "*A Review of Audio Features and Statistical Models Exploited for Voice Pattern Design*," Proc. 7th Int. Conf. on Pervasive Patterns and Applications (PATTERNS), Nice, France, 2015 [orale presentation].

[18] D. Badawy, A. Ozerov, N. Q. K. Duong, "*Relative group sparsity for non-negative matrix factorization with application to on-the-fly audio source separation*," Proc. 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Shanghai, China, 2015

[17] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot, "*An interactive audio source separation framework based on non-negative matrix factorization*," Proc. 37th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Italy, 2014 [orale presentation].

[16] N. Q. K. Duong, A. Ozerov, L. Chevallier, "*Temporal annotation-based audio source separation using weighted nonnegative matrix factorization*," Proc. IEEE Int. Conf. on Consumer Electronics (ICCE-Berlin), Germany, Sept. 2014 [orale presentation]**.**

[15] D.Badawy, N. Q. K. Duong, and A. Ozerov, "*On-the-fly audio source separation*," Proc. 24th IEEE International Workshop on Machine Learning for Signal Processing (MLSP), France, Sept. 2014 [orale presentation]**.**

[14] A. Ozerov, N. Q. K. Duong, L. Chevallier, "*On monotonicity of multiplicative update rules for weighted nonnegative tensor factorization*," Proc. Int. Symp. on Nonlinear Theory and its Applications (NOLTA), Switzerland, Sept. 2014 [orale presentation]**.**

[13] L Le Magoarou, A Ozerov, N. Q. K Duong, "*Text-informed audio source separation using nonnegative matrix partial co-factorization*," IEEE International Workshop on Machine Learning for Signal Processing (MLSP), UK, 2013 [orale presentation].

[12] N. Q. K. Duong and F. Thudor, "*Movie synchronization by audio landmark matching*," Proc. 37th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada, 2013.

[11] N. Q. K. Duong**,** C. Howson, and Y. Legallais, "*Fast second screen TV synchronization combining audio fingerprint technique and generalized cross correlation*," Proc. IEEE Int. Conf. on Consumer Electronics (ICCE-Berlin), 2012 [orale presentation].

[10] N. Q. K. Duong, E. Vincent, and R. Gribonval, "*An acoustically-motivated spatial prior for under-determined reverberant source separation*," Proc. 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Budapest, Hungary, pp. 9-12, 2011 [orale presentation].

[9] N. Q. K. Duong, H. Tachibana, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "*Multichannel harmonic and percussive component separation by joint modeling of spatial and spectral continuity*," Proc. 36th IEEE ICASSP, Budapest, Hungary, pp. 205-208, 2011.

[8] N. Q. K. Duong, E. Vincent and R. Gribonval, "*Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation*," Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA), Vol. 6365/2010, pp. 73-80, Springer, 2010.

[7] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Q. K. Duong, "*The 2010 signal separation campaign (SiSEC2010) - Audio source separation*," Proc. Int. Conf. on Latent Variabe Analysis and Signal Separation (LVA/ICA), Vol. 6365/2010, pp. 114-122, Springer, Sep. 2010 [invited paper].

[6] S. Araki, F. Theis, G. Nolte, D. Lutter, A. Ozerov, V. Gowreesunker, H. Sawada, and N. Q. K. Duong, "*The 2010 signal separation campaign (SiSEC2010) - Biomedical source separation*," Proc. Int. Conf. on Latent Variabe Analysis and Signal Separation (LVA/ICA), Vol. 6365/2010, pp. 123-130, Springer, Sep. 2010 [invited paper].

[5] N. Q. K. Duong, E. Vincent and R. Gribonval, "*Under-determined convolutive blind source separation using spatial covariance models*," Proc. 35th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Dallas-USA, pp. 9-12, Mar. 2010 [orale presentation].

[4] S. Arberet, A. Ozerov, N. Q. K Duong**,** E. Vincent, R Gribonval, F. Bimbot and P Vandergheynst, "*Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation,"* Proc. IEEE International Conference on Information Science, Signal Processing and their Applications (ISSPA), 2010 [orale presentation].

[3] N. Q. K. Duong, E. Vincent and R. Gribonval, "*Spatial covariance models for under-determined reverberant audio source separation*," Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), NewYork-USA, pp. 129-132, 2009 [orale presentation].

[2] N. Q. K. Duong, C. Park and S. H. Nam, "*An acoustic echo canceller combined with blind source separation,"* Proc. Conf. of the Acoustical Society of Korea, Vol.24, No.1, Aug. 2007.

[1] N. Q. K. Duong and S. H. Nam, "*Implementation of a basic acoustic echo cancelation*," Proc. of the Korea 19th Joint Signal Processing conf., Sep. 2006.

References I co-authored

[1] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-informed audio source separation using nonnegative matrix partial co-factorization," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2013, pp. 1–6.

[2] L. L. Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, pp. 1–5, 2014.

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

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

[5] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot, "An interactive audio source separation framework based on nonnegative matrix factorization," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 1586–1590.

[6] D. El Badawy, A. Ozerov, and N. Q. K. Duong, "Relative group sparsity for non-negative matrix factorization with application to on-the-fly audio source separation," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[7] H. T. T. Duong, Q. C. Nguyen, C. P. Nguyen, T. H. Tran, and N. Q. K. Duong, "Speech enhancement based on nonnegative matrix factorization with mixed group sparsity constraint," in *Proc. ACM Int. Sym. on Information and Communication Technology (SoICT)*, 2015, pp. 247–251.

[8] D. El Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly audio source separation-a novel user-friendly framework," *IEEE/ACM Transaction on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 261–272, 2017.

[9] G. Puy, A. Ozerov, N. Q. K. Duong, and P. Pérez, "Informed source separation via compressive graph signal sampling," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[10] H. T. T. Duong, N. Q. K. Duong, Q. C. Nguyen, and C. P. Nguyen, "Multichannel audio source separation exploiting NMF-based generic source spectral model in Gaussian modeling framework," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LCA/ICA)*, 2018.

[11] ——, "Gaussian modeling-based multichannel audio source separation exploiting generic source spectral model," *IEEE/ACM Transaction on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 32–43, 2019.

[12] N. Q. K. Duong, C. Howson, and Y. Legallais, "Fast second screen TV synchronization combining audio fingerprint technique and generalized cross correlation," in *IEEE Int. Conf on Consumer Electronics (ICCE-Berlin)*, 2014.

[13] N. Q. K. Duong and F. Thudor, "Movie synchronization by audio landmark matching," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3632–3636.

[14] N. Q. K. Duong, P. Berthet, S. Zabre, M. Kerdranvat, A. Ozerov, and L. Chevallier, "Audio zoom for smartphones based on multiple adaptive beamformers," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2017.

[15] E. Grinstein, N. Q. K. Duong, A. Ozerov, and P. Pérez, "Audio style transfer," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[16] P. Prablanc, A. Ozerov, N. Q. K. Duong, and P. Pérez, "Text-informed speech inpainting via voice conversion," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2016.

[17] H. V. Vo, N. Q. K. Duong, and P. Pérez, "Structural inpainting," in *Proc. ACM Int. Conf. on Multimedia (ACMM)*, 2018, pp. 1948–1956.

[18] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Pérez, and G. Richard, "Motion informed audio source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[19] ——, "Guiding audio source separation by video object information," in *Proc. IEEE Int. Workshop and Applications of Signal Processing to Audio and Acoutics (WASPAA)*, 2017.

[20] ——, "Weakly supervised representation learning for unsynchronized audio-visual events," in *Proc. CVPR Workshop*, 2019.

[21] S. Parekh, A. Ozerov, S. Essid, N. Q. K. Duong, P. Pérez, and G. Richard, "Identify, locate and separate: Audio-visual object extraction in large video collections using weak supervision," in *Proc. IEEE Int. Workshop and Applications of Signal Processing to Audio and Acoutics (WASPAA)*, 2019.

[22] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Pérez, and G. Richard, "Weakly supervised representation learning for audio-visual scene analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[23] Y. Shen, C.-H. Demarty, and N. Q. K. Duong, "Deep learning for multimodal-based video interestingness prediction," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 1003–1008.

[24] E. Berson, C. Demarty, and N. Q. K. Duong, "Multimodality and deep learning when predicting media interestingness," in *MediaEval Workshop, Dublin, Ireland, September 13-15.*, vol. 1984, 2017.

[25] C.-H. Demarty, M. Sjöberg, M. G. Constantin, N. Q. K. Duong, B. Ionescu, T.-T. Do, and H. Wang, "Predicting interestingness of visual content," in *Visual Content Indexing and Retrieval with Psycho-Visual Models*. Cham: Springer, 2017, pp. 233–265.

[26] E. Berson, N. Q. K. Duong, and C.-H. Demarty, "Collecting, analyzing and predicting socially-driven image interestingness," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2019.

[27] H. Squalli-Houssaini, N. Q. K. Duong, M. Gwenaëlle, and C.-H. Demarty, "Deep learning for predicting image memorability," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2371–2375.

[28] R. Cohendet, K. Yadati, N. Q. K. Duong, and C.-H. Demarty, "Annotating, understanding, and predicting long-term video memorability," in *Proc. ACM Int. Conf. on Multimedia Retrieval (ICMR)*, 2018, pp. 11–14.

[29] R. Cohendet, N. Q. K. Duong, and C.-H. Demarty, "Videomem: Constructing, analyzing, predicting short-term and long-term video memorability," in *Proc. International Conference on Computer Vision (ICCV)*, 2019.

[30] C.-H. Demarty, M. Sjöberg, B. Ionescu, T.-T. Do, H. Wang, N. Q. K. Duong, and F. Lefebvre, "Mediaeval 2016 predicting media interestingness task," in *Proc. of the MediaEval Workshop*, Hilversum, Netherlands, 2016.

[31] C.-H. Demarty, M. Sjöberg, B. Ionescu, T.-T. Do, M. Gygli, and N. Q. K. Duong, "Mediaeval 2017 predicting media interestingness task," in *Proc. of the MediaEval Workshop, Dublin, Ireland, September 13-15.*, vol. 1984, 2017.

[32] R. Cohendet, C. Demarty, N. Q. K. Duong, M. Sjöberg, B. Ionescu, and T. Do, "Mediaeval 2018: Predicting media memorability task," in *Proc. of the MediaEval Workshop*, 2018.

[33] M. G. Constantin, B. Ionescu, C.-H. Demarty, N. Q. K. Duong, X. Alameda-Pineda, and M. Sjöberg, "Predicting media memorability task at Mediaeval 2019," in *Proc. of the MediaEval Workshop*, 2019.

[34] M. Constantin, L. Stefan, B. Ionescu, N. Q. K. Duong, C.-H. Demarty, and M. Sjöberg, "Visual interestingness prediction: A benchmark framework and literature review," *International Journal of Computer Vision (IJCV), in revision*, 2020.

[35] N. Q. K. Duong, C. Howson, and Y. Legallais, "Second screen TV synchronization combining audio fingerprint technique and generalized cross correlation," in *Proc. IEEE Int. Conf. on Consumer Electronics - Berlin (ICCE-Berlin)*, 2012, pp. 241–244.

[36] G. Mafra, N. Q. K. Duong, A. Ozerov, and P. Pérez, "Acoustic scene classification: An evaluation of an extremely compact feature representation," in *Proc. IEEE Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[37] H. Le, P. Gilberton, and N. Q. K. Duong, "Discriminate natural versus loudspeaker emitted speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[38] V. Bilot, N. Q. K. Duong, and A. Ozerov, "Acoustic scene classification with multiple instance learning and fusion," in *Proc. IEEE Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) - Technical report*, 2019.

[39] N. Q. K. Duong, E. Vincent, and R. 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.

[40] H. T. T. Duong, Q. C. Nguyen, C. P. Nguyen, and N. Q. K. Duong, "Single-channel speaker-dependent speech enhancement exploiting generic noise model learned by non-negative matrix factorization," in *Proc. IEEE Int. Conf. on Electronics, Information, and Communications (ICEIC)*, 2016, pp. 1–4.

[41] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. IEEE Int. Conf. on Information Science, Signal Processing and their Applications (ISSPA)*, 2010, pp. 1–4.

[42] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for gaussian model based reverberant audio source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 1, pp. 1–11, 2013.

[43] ——, "Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010.

[44] Y. Shen, C.-H. Demarty, and N. Q. K. Duong, "Technicolor@MediaEval 2016 Predicting Media Interestingness Task," in *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, 2016.

Other references

[AEHKL+16] S. Abu-El-Haija, N. Kothari, J. Lee, A. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8M: A large-scale video classification benchmark. In *arXiv:1609.08675*, 2016.

[AFCS11] P. Arias, G. Facciolo, V. Caselles, and G. Sapiro. A variational framework for exemplar-based image inpainting. *Int. J. Computer Vision*, 93(3):319–347, 2011.

[AJY00] S. Rickard A. Jourjine and Ö. Yılmaz. Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2985–2988, June 2000.

[AZ17] R. Arandjelović and A. Zisserman. Objects that sound. *CoRR*, abs/1712.06651, 2017.

[BDTL15] P. Buyssens, M. Daisy, D. Tschumperlé, and O. Lézoray. Exemplar-based inpainting: Technical review and new heuristics for better geometric reconstructions. *IEEE Trans. Image Processing*, 24(6):1809–1824, 2015.

[Ber60] D.E. Berlyne. *Conflict, arousal and curiosity.* Mc-Graw-Hill, 1960.

[BMC05] J. Benesty, S. Makino, and J. Chen. *Speech Enhancement.* Springer, 2005.

[BPT14] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *Proceedings BMVC 2014*, pages 1–12, 2014.

[BS97] M.S. Brandstein and H.F. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *Proc. IEEE Int. Conf.*

on Acoustics, Speech, and Signal Processing (ICASSP), pages 375–378, 1997.

[BS10] J. Bitzer and K. U. Simmer. Superdirective microphone arrays. In *Microphone Arrays*, chapter 2, pages 19–38. Springer Verlag, 2010.

[BSCB00] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proc. SIGGRAPH*, 2000.

[BSFG09] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics*, 28(3):24–1, 2009.

[BSS10] S. Bhattacharya, R. Sukthankar, and M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proc. of ACM International Conference on Multimedia (MM), Florence, IT*, pages 271–280, 2010.

[BV16] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854, 2016.

[CDP01] A. Chen, P. W. Darst, and R. P. Pangrazi. An examination of situational interest and its sources. *British Journal of Educational Psychology*, 71(3):383–400, 2001.

[CEE13] B. Celikkale, A. Erdem, and E. Erdem. Visual attention-driven spatial pooling for image memorability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 976–983, 2013.

[CGDSLC16] R. Cohendet, A.-L. Gilet, M. Da Silva, and P. Le Callet. Using individual data to characterize emotional user experience and its memorability: Focus on gender factor. In *Proc. IEEE Int. Conf. on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2016.

[CGR14] Arridhana Ciptadi, Matthew S Goodwin, and James M Rehg. Movement pattern histogram for action recognition and retrieval. In *Proc. European Conference on Computer Vision (ECCV)*, pages 695–710. Springer, 2014.

[CMS15] S. Cappallo, T. Mensink, and C. G. M. Snoek. Latent factors of visual popularity prediction. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (ICMR)*, pages 195–202, 2015.

[Cow98] Nelson Cowan. *Attention and memory: An integrated framework*. Oxford University Press, 1998.

[CPT04] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Processing*, 13(9):1200–1212, 2004.

[CRZI19] M. Constantin, M. Redi, G. Zen, and B. Ionescu. Computational understanding of visual interestingness beyond semantics: literature survey and analysis of covariates. *ACM Computing Surveys*, 2019.

[CWSB19] J. Cramer, H.-H. Wu, J. Salamon, and J.-P. Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856. IEEE, 2019.

[CZ12] K. Chatfield and A. Zisserman. Visor: Towards on-the-fly large-scale object category retrieval. In *Asian Conference on Computer Vision*, Lecture Notes in Computer Science, pages 432–446. Springer, 2012.

[DOB11] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1657–1664, 2011.

[DSC+17] C.-H. Demarty, M. Sjöberg, G. Constantin, N. Q. K. Duong, B. Ionescu, T. T. Do, and H. Wang. Predicting interestingness of visual content. In *Visual Content Indexing and Retrieval with Psycho-Visual Models*. Springer, 2017.

[ECPC19] M. Engilberge, L. Chevallier, P. Pérez, and M. Cord. Sodeep: a sorting deep net to learn ranking loss surrogates. In *Proc. CVPR*, 2019.

[EI08] L. Elazary and L. Itti. Interesting objects are visually salient. *Journal of vision*, 8(3):3–3, 2008.

[EML+18] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37(4):112:1–112:11, July 2018.

[EPMP14] S. Ewert, B. Pardo, M. Müller, and M. D. Plumbley. Score-informed source separation for musical audio recordings: An overview. *IEEE Signal Processing Magazine*, 31(3):116–124, May 2014.

[EVC20] W. Ellahi, T. Vigier, and P. Le Callet. Can visual scanpath reveal personal image memorability? investigation of hmm tools for gaze patterns analysis. In *Proc. International Conference on Quality of Multimedia Experience (QoMEX)*, 2020.

[EVHH12] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2046–2057, 2012.

[FAMR18] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6363–6372, 2018.

[FBD09] C. Févotte, N. Bertin, and J. L. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[FRC10] S. Frintrop, E. Rome, and H. I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans. Appl. Percept.*, 7(1):1–39, January 2010.

[FSO17] M. Fakhry, P. Svaizer, and M. Omologo. Audio source separation in reverberant environments using beta-divergence based nonnegative factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(7), 2017.

[GFG18] R. Gao, R. Feris, and K. Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, September 2018.

[GGR+13] M. Gygli, H. Grabner, H. Riemenschneider, F. Fabian, and L. Van Gool. The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 1633–1640, 2013.

[Gir15] Ross Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448. IEEE, 2015.

[GKSS+17] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *arXiv:1612.00837v3*, 2017.

[HCE+17a] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. CNN architectures for large-scale audio classification. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[HCE+17b] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN architectures for large-scale audio classification. In *ICASSP*, pages 131–135. IEEE, 2017.

[HCL+18] G. Huang, D. Chen, T. Li, F. Wu, L. Maaten, and K. Weinberger. Multi-scale dense networks for resource efficient image classification. In *Proc. International Conference on Learning Representation (ICLR)*, 2018.

[HCRW16] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[HCS+15] J. Han, C. Chen, L. Shao, X. Hu, J. Han, and T. Liu. Learning computational models of video memorability from fMRI brain imaging. *IEEE Transactions on Cybernetics*, 45(8):1692–1703, 2015.

[HKHS15] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 23(12):2136–2147, 2015.

[HKS17] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? *arXiv preprint*, arXiv:1711.09577, 2017.

[HO00] Aapo Hyvarinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13:411–430, 2000.

[HX05] A. Hanjalic and L. Xu. Video affective content analysis: A survey of state-of-the-art methods. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[IAN13] Nobutaka Ito, Shoko Araki, and Tomohiro Nakatani. Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3238–3242, May 2013.

[ISS13] H. Izadinia, I. Saleemi, and M. Shah. Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Transactions on Multimedia*, 15(2):378–390, Feb 2013.

[IXP+14] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1469–1482, 2014.

[IXTO11] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 145–152. IEEE, 2011.

[KCS+17] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[KMK+19] J. Kim, M. Ma, K. Kim, S. Kim, and C.-D. Yoo. Progressive attention memory network for movie story question answering. In *Proc. CVPR*, 2019.

[KOCL16] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *European Conference on Computer Vision*, pages 350–365. Springer, 2016.

[KRTO15] A. Khosla, A. Raju, A. Torralba, and A. Oliva. Understanding and predicting image memorability at a large scale. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 2390–2398, 2015.

[KS08] Elizabeth A Kensinger and Daniel L Schacter. Memory and emotion. *Handbook of emotions*, 3:601–617, 2008.

[KSE05] E. Kidron, Y.Y. Schechner, and M. Elad. Pixels that sound. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 88–95 vol. 1, June 2005.

[KSH13] A. Khosla, A. Sarma, and R. Hamid. What makes an image popular? In *Proceedings of the International conference on World Wide Web*, pages 867–876, 2013.

[KSZ14] R. Kiros, R. Salakhutdinov, and R. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

[KWS+20] Q. Kong, Y. Wang, X. Song, Y. Cao, W. Wang, and M. D. Plumbley. Source separation with weakly labelled data: An approach to computational auditory scene analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 101–105, 2020.

[KYI+19] Q. Kong, C. Yu, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley. Weakly labelled audioset classification with attention neural networks. *arXiv preprint arXiv:1903.00765*, 2019.

[KYP13] J. Kim, S. Yoon, and V. Pavlovic. Relative spatial features for image memorability. In *Proc. ACM Int. Conf. on Multimedia (ACMM)*, pages 761–764, 2013.

[LAPGH20] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud. A recurrent variational autoencoder for speech enhancement. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 371–375, 2020.

[LBF11] A. Lefèvre, F. Bach, and C. Févotte. Itakura-Saito non-negative matrix factorization with group sparsity. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 21–24, 2011.

[LDDR13] A. Liutkus, J. L. Durrieu, L. Daudet, and G. Richard. An overview of informed audio source separation. In *Proc. IEEE Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4, 2013.

[LEOEQ16] S. Lahrache, R. El Ouazzani, and A. El Qadi. Bag-of-features for image memorability evaluation. *IET Computer Vision*, 10(6):577–584, 2016.

[LS01] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural and Information Processing Systems 13*, pages 556–562, 2001.

[LSR+17] A Liutkus, F. R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave. The 2016 signal separation evaluation campaign. In *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation*, pages 323–332, 2017.

[MCCD13] T. Mikolov, K. Chen, G.S. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, 2013.

[McG00] James L McGaugh. Memory–a century of consolidation. *Science*, 287(5451):248–251, 2000.

[MFL+19] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa†, and H. Ghasemzadeh. Improved knowledge distillation via teacher assistant. *arXiv:1902.03393v2*, 2019.

[MHD+17a] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen. DCASE 2017 challenge setup: Tasks, datasets and baseline system. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pages 85–92, November 2017.

[MHD+17b] Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. DCASE2017 challenge setup: Tasks, datasets and baseline system. In *Proc. of Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pages 85–92, November 2017.

[MHDV15] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen. Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations. In *ICASSP*, pages 151–155. IEEE, 2015.

[ML03] X. Mestre and M. Lagunas. On diagonal loading for minimum variance beamformers. In *Proc. IEEE Int. Symp. on Signal Processing and Information Technology (ISSPIT)*, pages 459–462, 2003.

[MLM13] Matei Mancas and Olivier Le Meur. Memorability of natural scenes: The role of attention. In *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, pages 196–200, 2013.

[MOD14] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong. Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization. *Journal of Signal Processing Systems*, pages 1–5, 2014.

[OBLS15] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015.

[OIM+16] A. Owens, P. Isola, J. McDermott, A. Torralba, E.-H. Adelson, and W.-T. Freeman. Visually indicated sounds. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2405–2413, 2016.

[OVB12] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1118–1133, 2012.

[OWM+16] A. Owens, J. Wu, J.-H. McDermott, W.-T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *Proc. of European Conference on Computer Vision*, pages 801–816. Springer, 2016.

[PKD+16] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. In *Proc. Conf. Comp. Vision Pattern Rec.*, 2016.

[PVZ12] O. M. Parkhi, A. Vedaldi, and A. Zisserman. On-the-fly specific person retrieval. In *13th Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4, 2012.

[Qui66] M Ross Quillan. Semantic memory. Technical report, Bolt Beranek and Newman Inc Cambridge MA, 1966.

[RDP+11] U. Rimmele, L. Davachi, R. Petrov, S. Dougal, and E. A. Phelps. Emotion enhances the subjective feeling of remembering, despite lower accuracy for contextual details. *Psychology Association*, 2011.

[RP13] Zafar Rafii and Bryan Pardo. Online REPET-SIM for real-time speech enhancement. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 848–852, 2013.

[RQD] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19–41.

[SCKP09] Mohammad Soleymani, Guillaume Chanel, Joep J. M. Kierkels, and Thierry Pun. Affective characterization of movie scenes based on content analysis and physiological changes. *Int. J. Semantic Computing*, 3:235–254, 2009.

[Sil05] Paul J Silvia. What is interesting? exploring the appraisal structure of interest. *Emotion*, 5(1):89, 2005.

[SM13] D. L. Sun and G. J. Mysore. Universal speech models for speaker independent single channel source separation. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 141–145, 2013.

[SSS+17] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. Show and recall: Learning what makes videos memorable. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2730–2739, 2017.

[SZ14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[SZM+17a] A. Siarohin, G. Zen, C. Majtanovic, X. Alameda-Pineda, E. Ricci, and N. Sebe. How to make an image more memorable?: A deep style transfer approach. In *Proceedings of the ACM on International Conference on Multimedia Retrieval (ICMR)*, pages 322–329, 2017.

[SZM+17b] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. How to make an image more memo-

rable? a deep style transfer approach. In *ACM International Conference on Multimedia Retrieval*, 2017.

[SZM+19] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. Increasing image memorability with neural style transfer. *ACM Transactions on Multimedia Computing Communications and Applications*, 2019.

[TBF+15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proc. Int. Conf. on Computer Vision (CVPR)*, pages 4489–4497, 2015.

[VBGB14] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot. From Blind to Guided Audio Source Separation: How models and side information can improve the separation of sound. *IEEE Signal Processing Magazine*, 31(3):107–115, 2014.

[VGF06] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.

[Vir07] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparse criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066 – 1074, 2007.

[Wan03] A. L-C. Wang. An industrial-strength audio search algorithm. In *Proc. Int. Sym. on Music Information Retrieval (ISMIR)*, pages 1–4, 2003.

[WLM19] Y. Wang, J. Li, and F. Metze. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2019.

[WR16] S. Wood and J. Rouat. Blind speech separation with GCC-NMF. In *Proc. Interspeech*, pages 3329–3333, 2016.

[XKWP17] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley. Surrey-CVSSP system for DCASE2017 challenge task4. Technical report, DCASE2017 Challenge, September 2017.

[YAT16] C. Vondrick Y. Aytar and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proc. Neural Information Processing Systems (NIPS)*, 2016.

[YH] J. Yu and T.-S. Huang. Universally slimmable networks and improved training techniques. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1803–1811.

[YLL+16] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *Proc. Conf. Comp. Vision Pattern Rec.*, 2016.

[YSY+17] Y. Yang, Y. Song, Y. Yu, Y. Kim, and G. Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *Proc. CVPR*, 2017.

[ZBFC18] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. From recognition to cognition: Visual commonsense reasoning. *arXiv:1811.10830*, 2018.

[ZD14] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, pages 391–405. Springer, 2014.

[ZdJSJ16] G. Zen, P. de Juan, Y. Song, and A. Jaimes. Mouse activity as an indicator of interestingness in video. In *Proceedings of the ACM on International Conference on Multimedia Retrieval (ICMR)*, pages 47–54, 2016.

[ZGR+18] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *ECCV*, September 2018.

[ZPV06] Cha Zhang, John C Platt, and Paul A Viola. Multiple instance boosting for object detection. In *Advances in neural information processing systems*, pages 1417–1424, 2006.

[ZZHJH10] X. Zhuang, X. Zhou, M. Hasegawa-Johnson, and T.-S. Huang. Real-world acoustic event detection. *Pattern Recognition Letters*, 31(12):1543–1551, 2010.