EEG-based Decoding of Auditory Attention to a Target Instrument for Neuro-steered Music Source Separation

Giorgia Cantisani, Slim Essid, Gael Richard

EEG-based Decoding of Auditory Attention to a Target Instrument for Neuro-steered Music Source Separation

Giorgia Cantisani, Slim Essid, Gaël Richard

***Abstract*—This paper describes a novel *neuro-steered music source separation* framework and conducts an extensive evaluation of the proposed system on MAD-EEG, a dataset composed of EEG recordings of subjects attending to a particular in duo and trio music excerpts. We propose an unsupervised non-negative matrix factorisation (NMF) variant, named Contrastive-NMF, that separates a target instrument, guided by the user's selective auditory attention to that instrument, which is tracked in his/her electroencephalographic (EEG) response to music. We analyse the impact of multiple aspects of the musical stimuli, such as the number and type of instruments in the mixture, the spatial rendering and the music genre, obtaining encouraging results, especially in difficult cases where non-informed models struggle. We believe that this unsupervised NMF variant is advantageous for neuro-steered music source separation as it allows us to incorporate additional information in a principled optimisation fashion and does not need training data, which is particularly difficult to acquire for applications involving EEG recording.**

***Index Terms*—Audio source separation, Auditory attention decoding, Polyphonic music, EEG, Matrix factorisation, Multimodal processing**

I. Introduction

MUSIC source separation aims to isolate individual sources, such as singing voice, guitar, drums, cello, etc., mixed in an audio recording of a musical piece. More precisely, such individual voices can be referred to as *stems*, *i.e.* recordings of individual instruments that are arranged together and mastered into the final audio mix. Considering the case of single-channel recordings, one can assume that the mixture signal $x(t)$ at sample t is a linear mixture of J sources $s_j(t)$ such as:

$$x(t) = \sum_{j=1}^{J} s_j(t). \quad (1)$$

Given only $x(t)$, the goal of a source separation system is to recover one or more sources $s_j(t)$, where $j \in \{1, ..., J\}$.

Recovering such stems is a very challenging problem, and a source separation system can be either directly exploited by the end-user (e.g. a musician or a sound engineer) or be an intermediate step that significantly helps other downstream tasks such as automatic music transcription, instrument classification, score following, lyrics alignment and many others. Nowadays, most state-of-the-art music source separation systems are based on supervised deep learning [1]–[4], where an extensive collection of mixtures and corresponding isolated sources are needed during a training phase. Despite the release of dedicated datasets for this task [5], [6], it is still hard for those models to generalize to unseen test data with significant timbral variation compared to training.

A possible solution to mitigate this issue is to inform the source separation system with additional information one may have about the test data. In this case, the approach is referred to as *informed audio source separation* [7] and, if this additional information comes from another modality than the audio itself, as *multimodal source separation*. Examples of such additional information include the score [8], the pitch contour [9], the lyrics [10], the motion of the sound sources and visual cues [11], or the user feedback [12], [13]. In this context, the user can be considered as a rich source of information about the sources of interest. Beyond manual annotations, our body's reaction to external auditory stimuli manifests itself through many observable physiological phenomena. Reaction to music can be seen in the heartbeat variability [14], in the body movements [15], as well as in the neural activity [16] to mention a few. Such kind of information would help the separation process and make it also *interactive* and *real-time*, allowing for a number of futuristic applications.

Among those physiological responses to music stimuli, we are particularly interested in looking at the neural response, focusing on the concept of *selective auditory attention*. Humans' auditory system is naturally able to process concurrent sounds in a complex auditory scene to isolate the ones of interest. This problem is known as the cocktail party problem [17]–[19] and has been studied mostly for what concerns the perception of speech sources in noisy or multi-speaker settings. Even if the cognitive mechanism behind this capability is not yet fully understood, selective auditory attention has been proven to have a determinant role in it [20]. The attended source's neural encoding appears to be substantially stronger than the one of the unattended sources left in the mixture, and therefore, the attended source can be tracked in the neural activity [21]. This fact makes it possible to decode the auditory attention, *i.e.* determining which sound source a person is "focusing on" by just observing the listener's brain response. This task is known as *auditory attention decoding (AAD)*, and typical applications are intelligent hearing aids where a neuro-steered enhancement of the attended speaker is desired [22]–[24].

When dealing with polyphonic music, one can recast the problem as one of decoding the attention to a particular in-

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068 (MIP-Frontiers).

All three authors are with the Information Processing and Communication Laboratory (LTCI), Télécom Paris, Institut Polytechnique de Paris, 19 place Marguerite Perey, 91120 Palaiseau, France. E-mail: {giorgia.cantisani, slim.essid, gael.richard}@telecom-paris.fr.

strument playing in the ensemble. However, this transposition is not straightforward as, unlike in the cocktail party problem where there is one source of interest to separate from unrelated background noise or speakers, music consists of multiple voices playing together in a coordinated way. Thus, the sources are generally highly correlated, making the decoding problem even more difficult, but still possible [25], [26].

In our previous works [26], [27], we explored how the neural activity reflects information about the attended instrument and how we can use it to inform a source separation system and adapt it to the specific stimulus. We were particularly interested in electroencephalographic signals (EEG), which allow for non-invasive neural activity acquisition with high temporal resolution. Since the topic had not yet been explored, mainly due to the lack of data, we began by acquiring a dataset, namely, the MAD-EEG [28] dataset, which consists of a set of musical stimuli and corresponding EEG responses, where the participants attend to a particular instrument in the stimulus. First, in [26], we studied the problem of EEG-based AAD to a target instrument in polyphonic music, showing that the EEG tracks musically-relevant features which are highly correlated with the time-frequency representation of the attended source and only weakly correlated with the unattended one. Second, in [27], we leveraged this "contrast" to inform an unsupervised source separation model based on a novel non-negative matrix factorization (NMF) variant, named Contrastive-NMF (C-NMF) and automatically separate the attended source.

In this work, we present a consolidated view of our *neuro-steered music source separation* paradigm, further validating and extending the preliminary evidence obtained in [27], where only a subset of pop mixtures with at most two instruments was analyzed. Key contributions of this paper are the following:

- the C-NMF algorithm is generalised so as to make it valid for mixtures containing more than two instruments;
- the AAD problem is reformulated in a "blind" way, *i.e.* without access to the ground truth sources;
- a new experimental setting with a substantial extension of previous experiments is provided, where all the mixtures in the MAD-EEG dataset are considered. All pop mixtures are tested, also considering trios. Classical mixtures are also taken into account, increasing the number of analysed instruments from four to nine;
- an extensive analysis of the algorithm's behaviour w.r.t. to its hyperparameters was conducted, followed by a discussion on the option of adapting them to specific instruments and songs.

The remainder of the paper is organized as follows. In Section II, we describe works related to EEG-based AAD and informed music source separation, while in Section III, we describe the MAD-EEG data. In Section IV and V, we describe the work conducted on EEG-based auditory attention decoding and neuro-steered music source separation, respectively and how they relate to each other. Finally, we conduct a reflection about future research directions and limitations in Section VI.

II. Related Work

A. EEG-based Auditory Attention Decoding

EEG-based Auditory attention decoding (AAD) aims at determining which sound source a person is "focusing on" by analysing the listener's brain response. Most of the literature in the field focuses on decoding auditory attention to naturalistic speech in multi-speaker or noisy scenarios from the brain's electric activity measured on the scalp [29], [30]. Indeed, the topic is raising more and more interest thanks to the multitude of promising applications, especially concerning hearing aids and cochlear implants [22]–[24], [31]–[33].

First studies on AAD based on continuous electrocorticographic (ECoG) [21], [34], [35] and electroencephalographic (EEG) [29], [30], [36] responses have shown that changes in the audio stimulus can be tracked in the neural activity. They evidenced how the attended source's neural encoding is substantially stronger than the one of the other sources left in the mixture, allowing for a successful decoding of selective attention to a speaker. Similarly to Treder et al. [25], in our previous work [26], we recast the AAD problem in the music domain as one of decoding attention to a specific musical instrument playing in a musical ensemble.

The decoding procedure is usually two-fold [29]: first, a feature representation of the attended audio source is reconstructed from the neural response. Secondly, the reconstruction is correlated with the ground truth sources to determine the attended source. The stimulus reconstruction is referred to as the *backward* problem, as one goes from the brain response back to the stimulus. The mapping is usually done using linear models: a multichannel Wiener filter maps the neural activity back to a stimulus feature representation [36]. Such a filter is known in the field as *temporal response function* (TRF) and is learned on a training set using a minimum mean squared error (MMSE) criterion [36].

The majority of works represented the speech merely by its broadband temporal envelope [29], [30], [37], [38]. Other works obtained promising results using the speech spectrogram [21], [34], [35], phonemes [39], or semantic features [40]. In our previous work [27], we compared multiple acoustic representations of the music stimulus in terms of AAD performance, namely the broadband amplitude envelope, the spectrogram and the Mel-spectrogram.

B. EEG-informed Source Separation

The AAD task is naturally related to audio source separation. As previously explained in Section II-A, the decoding paradigm requires access to the ground truth sources to correlate them to the neural data. However, this situation is never met in realistic scenarios such as hearing aids and cochlear implants, where only the mixture of the sound scene recorded by their microphones is available. In such scenarios, an additional *audio source separation* step is needed to extract the reference sources needed for the decoding. Typically, the separation and the decoding tasks are tackled sequentially: a separation system provides the reference sources for the decoding, and the decoding system selects the source which needs to be enhanced.

Most of the studies that relate speech source enhancement and AAD have been working in this direction. Many of them focused on the multi-channel audio scenario using beamforming [24], [41], [42] and multi-channel Wiener filtering [23], [31], [43]. Both techniques estimate spatial filters that return the target speech when applied to the mixture while suppressing the background noise and interfering sources. The former uses only spatial information such as the directions of arrival while the latter also requires information about the target activity to compute the second-order statistics of the noise and interferers.

Other works focus on the single-channel scenario using deep-learning-based approaches. O'Sullivan et al. [44] were the first along this line. However, their paradigm requires prior training on the target speakers, which is a substantial limitation in real scenarios. The problem is tackled by Han et al. [22] with a speaker-independent source separation system able to generalize to unseen speakers. Such a system relies on a deep attractor network, which projects the mixture's time-frequency representation in a high-dimensional space where the speakers are separable [45], [46]. Ceolini et al. [33], instead, informed a speech separation neural network with the decoded attended speech envelope, leading to the extraction of the attended source. However, the training of the source separation model and that of the AAD model are still decoupled due to the lack of large datasets collected for AAD.

In general, performing the source separation and AAD steps independently is sub-optimal. In their work [32], Pu et al. propose a unified model for joint AAD and binaural beamforming. An adaptive beamformer is learned thanks to an objective which minimizes noise and interference but, at the same time, controls the target speaker distortion and maximizes the Pearson correlation between the envelope of the beamformer output and the decoded EEG. In a later work [47], the same authors showed that their algorithm is robust to attention switching, which can be tracked in real time thanks to the joint approach.

Nevertheless, none of these works considers music audio signals. We aim to pursue the joint approach, which we believe is the most promising one for a *neuro-steered music source separation*. In [27], we propose to adapt an NMF-based source separation model to a specific mixture using a weak signal decoded from the EEG using an AAD model. The AAD model is not fixed and is also updated during the optimization. Our work differs from those by Pu at al. [32] as our aim is not to maximize the Pearson correlation between the envelope of the beamformer output and the decoded EEG. Since the decoded output can be significantly deteriorated [26], we leverage instead the fact that the attended instrument's neural encoding is substantially stronger than the one of the unattended sources left in the mixture. This "contrast" is maximized when solving our separation model estimation problem.

III. Data

We start by recalling the main features of the dataset we have assembled for this work to allow the reader to have a clear understanding of the recording protocol and how the proposed algorithm is applied to this data. For more details, the interested reader can refer to [28].

A few publicly available music-related EEG datasets exist [48]–[52] which contain EEG responses to naturalistic music stimuli. Nevertheless, in those cases, the participants were asked to focus on the entire stimulus, making them not relevant for studying the AAD problem. The only music-related EEG dataset where participants were asked to attend to a target instrument in the music mixture is the `music BCI` dataset collected by Treder et al. [25]. However, it was explicitly designed for studying ERP-based AAD using a multi-streamed oddball paradigm, which does not hold in real-world scenarios. On the contrary, when considering speech-related EEG datasets, one can find several of them specifically designed to study the AAD problem using a single-trial approach, but only a few of them are accessible [38], [53].

Taking inspiration from the speech-related EEG datasets, we acquired our EEG dataset from subjects listening to realistic polyphonic music and attending to a particular instrument in the mixture. Our dataset represents the first EEG dataset specifically designed for studying auditory attention decoding applied to music using single-trial techniques.

A. Participants and neural recording

Eight volunteers (7 males and one female, all but one right-handed, aged between 23 and 54 years, mean age of 28) took part in the study. All of them were healthy and reported no history of hearing impairments or neurological disorders. The study conforms with the Declaration of Helsinki [54]. Moreover, all participants signed a consent that informed them about the experiment's modalities and purposes.

They were all non-professional musicians with varying years of musical experience (from 7 to 30 years, mean 13.5). Five out of them play the guitar, one the bass, one the drums, and one is a multi-instrumentalist playing the drums, guitar and bass. They all studied music theory (from 1 to 2 hours per week, mean 1.75) and practised regularly with their instrument (from 2 to 14 hours per week, mean 6.25). All of them were familiar with the modern instruments in the dataset (drums, guitar, bass and singing voice), while for specific classical instruments (bassoon, French horn and oboe), not all of them were equally confident. Thus, they were trained to recognize them before the experiment using excerpts not used as stimuli.

B. Stimuli

The stimuli consist of realistic polyphonic music mixtures containing two to three instruments played concurrently in an ensemble. The chosen mixtures reproduce a realistic setting. In particular, we chose to use real music composition for pop pieces for which we had access to the isolated instrument tracks. For Classical music pieces, instead, we linearly mixed a selection of excerpts played by single instruments as follows: $x(t) = \sum_{j=1}^{J} g_j s_j(t)$, where $s_j(t)$ is the mono-channel audio track of the single instrument j, g_j is the corresponding gain, T its number of samples, and J is the number of instruments. Finally, the sound volume was normalized to avoid bias due to the loudness of the audio.

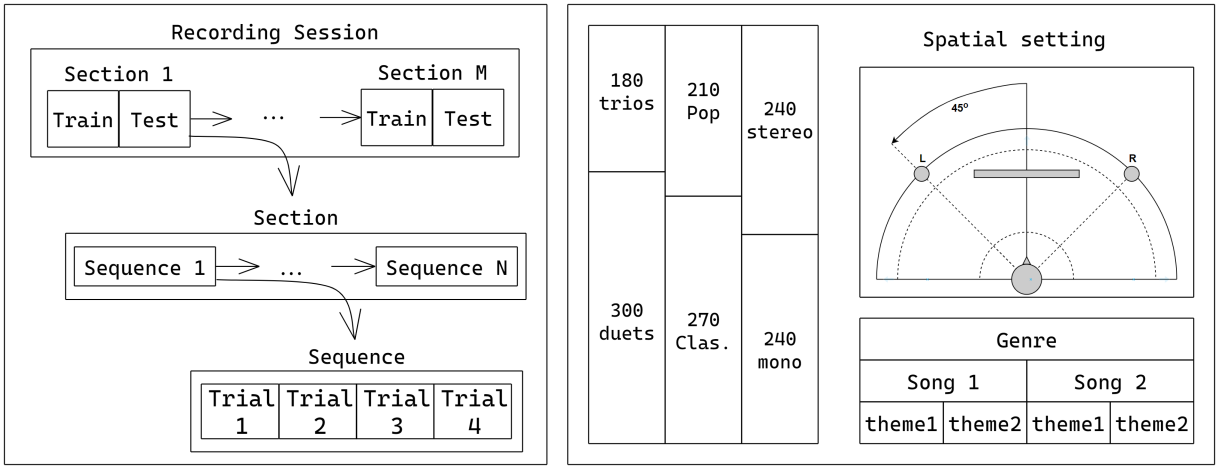

Fig. 1. On the left, an illustration of the recording session for one subject. A recording session is divided into sections. Each section is associated with a given musical piece in the dataset and consists of a training and a test phase, where a series of stimuli sequences is played. Each stimulus sequence consists of 4 trials where the same stimulus is listened to repetitively. On the right details about the mixtures and how they are spatially rendered.

In order to test the influence of certain factors on the attention decoding performance, we considered different configurations in the choice of the musical stimuli (see Figure 1):

- Two musical *genres*: pop and Classical music.
- Two musical *pieces* per genre and two *themes* per musical piece. That is, for the same piece, two different excerpts , corresponding to different parts of the score.
- Two *ensemble types*: duets and trios.
- Two *spatial rendering configurations*: monophonic and stereo. The speakers were situated $\pm 45^o$ along the azimuth direction relative to the listener (see Figure 1). The stereo spatial rendering was implemented by merely using conventional stereo panning where we have one instrument mostly on the right and the other one mostly on the left for duets, while for trios, the third instrument is in the centre. The target instrument is never in the same position across different stimuli.
- *Musical instruments* present in the mixture: different combinations of flute, oboe, French horn, bassoon and cello for Classical pieces, along with singing voice, guitar, bass and drums for pop excerpts.

C. Recording protocol

Each stimulus duration had to be long enough to allow the study of attention decoding on a single-trial basis while targeting realistic music excerpts. On the other hand, the experiment's duration had to remain reasonably short to control the subject's cognitive load and avoid an unsatisfactory level of concentration throughout the session. Consequently, we limited the duration of a stimulus to around 6 seconds. Then, during the experiment, each stimulus was heard by the subject four consecutive times, referred to as *trials*, corresponding to around 24 seconds of EEG recordings. Since each subject listened to 78 stimuli, this corresponds to approximately 30-32 minutes of recordings.

For each subject, the *recording session* was divided into *sections* (see Figure 1). In each section, a series of stimuli *sequences* is played. Each section is actually composed of a *training* and a *test* phase. During the training phase, single instrument tracks of a given piece are played separately as solos in random order. Then, during the test phase, all the corresponding duo and trio variants of the same piece are played, also in a random order, but with a potentially different spatial rendering and considering a different theme of the same musical piece. A section is presented to the user through a slide-show video showing instructions, displayed as white text on a black background, asking the participant to attend to a particular instrument and visually fix a cross at the centre of the screen. A "beep" precedes each stimulus launch.

D. Data Acquisition and Preprocessing

A B-Alert X24[1] headset was used to record the surface EEG, EOG (Electrooculogram), EMG (Electromyogram) and ECG (Electrocardiogram) of the participants, as well as their head motion acceleration, thanks to an integrated inertial measurement unit, all at a sampling frequency $f_s = 256Hz$. The headset consists of a wireless digital acquisition unit connected to an electrode strip. The strip used has electrodes F1, F2, F3, F4, Fz, C1, C2, C3, C4, Cz, CPz, P1, P2, P3, P4, Pz, POz, O1, O2 and Oz, placed according to the 10-20 montage system. Active electrodes were referenced to the left mastoid in a unipolar setting. The acquired EEG data was synchronized with each stimulus, the 50 Hz power-line interference was removed using a notch filter, and EOG/ECG artefacts were detected and removed using independent component analysis (ICA). The frequencies below 1 Hz were filtered out using a Butterworth zero-phase filter with order 2. Each channel was normalized to ensure zero mean and unit variance.

IV. EEG-based Auditory Attention Decoding

The goal is to determine the attended instrument in a single-trial fashion based on 24-second-long EEG excerpts aligned to

[1]https://www.advancedbrainmonitoring.com/xseries/x24/

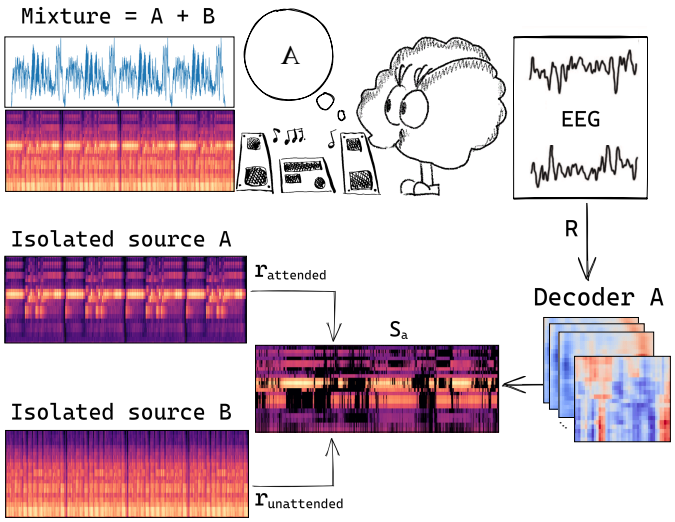

Fig. 2. AAD paradigm: a subject-specific TRF model is used to predict a representation of the attended instrument from the EEG response to the musical stimulus. Then, the reconstruction is correlated with the ground truth sources to determine the attended source.

corresponding audio stimuli. The procedure is two-fold and is similar to the one commonly used for decoding the attention to speech [21], [29], [34]–[36] (see Figure 2). First, a feature representation of the attended audio source is reconstructed from the neural response, exploiting a decoder previously trained on solos of that instrument. Second, the reconstruction is correlated with the ground truth sources to determine the attended source. The attended instrument is recognized as the one that has the highest Pearson's correlation coefficient.

A. Audio Feature Extraction

Choosing the audio representation is a crucial point of the AAD paradigm, as this choice includes a hypothesis about the neural coding of the stimulus and can significantly impact the reconstruction quality and the decoding performance. We studied three different audio representations, one in the time domain and two in the time-frequency (TF) domain: the time domain amplitude envelope computed using the Hilbert transform (AE), the magnitude spectrogram (MAG), and the Mel spectrogram (MEL), a perceptually-scaled representation commonly used for music analysis.

The AE is one of the most used feature representations for AAD with speech stimuli as the EEG was shown to track slowly varying changes in the audio stimulus [29], [55]. The assumption is that the EEG is linearly related to the broadband energy envelope of the stimulus. However, frequency modulations, *i.e.* envelope fluctuations at specific frequencies, can give a more complete view of the audio signal. In fact, the spectrogram envelope of natural sounds fluctuates across both frequency and time, and this was shown, for instance, to be important for the intelligibility of speech [35]. The same can be said for music, where the modulations' complexity is much higher than in speech. In practice, the spectrogram can be seen as a time-varying representation of the amplitude envelope at each frequency bin [35]. Thus, we will assume that the neural responses are linearly related to the spectrogram envelope.

B. Temporal Response Function (TRF)

A feature representation of the attended source $\hat{\mathbf{S}} \in \mathbb{R}^{K \times N}$ where K is the number of features and N is the number of time samples, is reconstructed from the EEG using the TRF backward model commonly used in the AAD framework [36]. This filter can be seen as a *spatio-temporal decoder* which linearly maps the neural activity back to the audio feature representation, as a weighted sum of activity at each electrode in a given temporal context, as follows: $\hat{\mathbf{S}} = \mathbf{g}^T\mathbf{R}$, where $\mathbf{g} = [\mathbf{g}_1, ..., \mathbf{g}_K] \in \mathbb{R}^{(C \times L) \times K}$ is a tensor composed by the set of multi-channel Wiener filters, and $\mathbf{R} \in \mathbb{R}^{(C \times L) \times N}$ is the Toepliz matrix of the neural response. C represents the number of EEG channels and L the number of time lags, *i.e.* the temporal context where we assume to see the EEG response to the stimulus which ranges between 0 and τ_{max}. In practice, each $k-th$ feature of $\hat{\mathbf{S}}$ is reconstructed independently from the others using a multi-channel Wiener filter $\mathbf{g}_k \in \mathbb{R}^{C \times L}$, which is learned through an MMSE criterion on a training set of solos of the same instrument. Each filter is estimated independently as the normalized reverse correlation: $\mathbf{g}_k = \mathbf{C}_{\mathbf{RR}}^{-1}\mathbf{C}_{\mathbf{RS}}$, where $\mathbf{C}_{\mathbf{RR}} = \mathbf{R}\mathbf{R}^T$ is the auto-correlation of the EEG data and $\mathbf{C}_{\mathbf{RS}} = \mathbf{R}\mathbf{S}_k^T$ is the cross-correlation of the stimulus and EEG data across all electrodes and time-lags for the k^{th} feature. Since EEG signals are high-dimensional, autocorrelated, noisy data with high trial-to-trial variability, the estimate of the covariance matrices can be imprecise and subject to overfitting due to the high number of parameters to estimate [56]. Thus, a Ridge regularization is used to constrain the model coefficients as follows: $\mathbf{g}_k = (\mathbf{C}_{\mathbf{RR}} + \gamma\mathbf{I})^{-1}\mathbf{C}_{\mathbf{RS}}$, where $\mathbf{I} \in \mathbb{R}^{C \times L}$ is the identity matrix and $\gamma \in [0, 1]$ is the regularization parameter.

C. Evaluation

We evaluate the TRF reconstruction capabilities through the Pearson's correlation coefficient of the reconstructed stimulus representation with the attended instrument $r_{attended}$, the unattended instrument $r_{unattended}$ and the mixture $r_{mixture}$.

Beside the reconstruction capabilities, we also evaluate the decoding performance in terms of accuracy on the AAD task. Their statistical significance was assessed using an adaptation of the computationally-intensive randomization test [57], a non-parametric hypothesis test, comparing to chance, which does not make any assumption on the score distribution [58]. The considered significance levels are 5%, 1%, 0.1% and 0.01%, and the tests were performed over 10^4 iterations. This was done by implementing the following procedure: first, we considered a random classifier, that, given a test mixture, chooses the attended instrument randomly among the instruments in the given mixture. Then, the performances were computed over the random predictions on the complete test set. This procedure was repeated 10000 times, which resulted in a distribution of the performances. This empirical distribution was then approximated with a theoretical distribution which could be a normal or a t-distribution (the one that fits better).

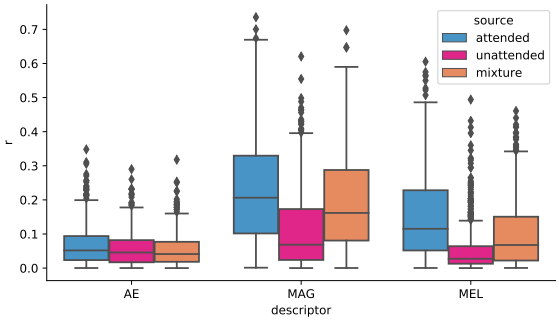

Fig. 3. Pearson's correlation coefficients of the reconstructed stimulus with the attended source (blue), the unattended one (pink) and the mixture (orange) for the three audio descriptors.

Accuracy (%)	All	Duets	Trios
AE	52 ***	59 **	40*
MAG	75 ****	78 ****	69****
MEL	75 ****	76 ****	74 ****

TABLE I
DECODING ACCURACY FOR DIFFERENT SUBSETS OF THE TEST SET WITH $\gamma = 0.1$ AND $\tau_{max} = 250ms$. "****" DENOTES VERY HIGH ($p < 0.0001$), "***" HIGH ($p < 0.001$), "**" GOOD ($p < 0.01$), "*" MARGINAL ($p < 0.05$) AND "N.S." NO ($p > 0.05$) STATISTICAL SIGNIFICANCE FOR A NON-PARAMETRIC RANDOMIZATION TEST.

Then we evaluated how likely our model's actual performances were to be produced by this artificial distribution of performances obtaining the P-value.

D. Experimental Results

In Table I one can see the decoding accuracy with respect to the three audio descriptors and number of instruments in the mixture ($\gamma = 0.1$, $\tau_{max} = 250ms$). All the scores are significantly above the chance level, which is 50% for duets, around 33% for trios, and around 44% for all the test mixtures together. TF representations are clearly beneficial for the decoding indicating that envelope fluctuations at specific frequencies can give a complete view of the music audio signal. The two spectrograms, especially the MEL spectrogram, also proven to be more robust to the mixture's number of instruments. Nevertheless, even if the accuracy scores obtained with the AE are drastically below those obtained with the other two descriptors, they are still statistically significant.

In Figure 3 one can see the Pearson's correlation coefficients of the reconstructed stimulus with the attended source (blue), the unattended one (pink) and the mixture (orange) for the three audio descriptors. The correlation scores are very low, indicating that the reconstructions are highly deteriorated. Nevertheless, the "contrast" between $r_{attended}$ and $r_{unattended}$ is evident, especially for the two TF descriptors, confirming the decoding results of Table I.

The lowest $r_{attended}$ Pearson's coefficients are those related to the AE but are still comparable to those obtained by O'Sullivan et al. in [29] for speech with the same audio descriptor. However, since the contrast between $r_{attended}$ and $r_{unattended}$ is only marginal, the decoding accuracy is much lower than the one obtained by the same authors. The broadband envelope is probably enough for discriminating between attended and unattended speakers but is not enough when dealing with music. Music present complex modulations both in time and frequency, for which the energy envelope is not enough representative. From the same plot, we can observe that the correlations obtained with the MAG spectrogram are marginally higher than the ones obtained with the MEL one (median $r = 0.215$ for MAG, median $r = 0.119$ for MEL). However, the "contrast" between $r_{attended}$ and $r_{unattended}$ is higher for MEL, which is reflected in the decoding accuracy. The MEL spectrogram is a perceptually scaled and compact version of the linear spectrogram (MAG). A non-linear transformation of the frequency scale based on the perception of pitches (Mel scale) is applied to the linear spectrogram so that two pairs of frequencies that are equidistant in the Mel scale are perceived as being equidistant by humans. We observed that a lower number of features K, or MEL bands, is beneficial for the performance during the experiments. In particular, we tested values $\in [12, 60]$, and the results we show are relative to 24 Mel bands. Probably, the MAG representation has a too high number of features K, as it corresponds to the number of frequency bins (in our experiments 512), which might be too complex for the AAD task.

The outcomes are very positive: using a simple linear regression model, we obtain a reconstructed representation that is more correlated with the attended instrument than with the unattended one. This contrast is particularly significant when using TF audio representations, highlighting amplitude modulations in different frequency bands. Among the two TF representations, the more compact and perceptually scaled representation given by the MEL spectrograms appears to be more robust to highlight the contrast.

V. EEG-INFORMED SOURCE SEPARATION

The goal is now to separate a target instrument from a given music mixture. Along with the audio signal, we have access to the EEG recorded from the subject while she/he was listening to the given mixture and attending to the target instrument.

From the experiments presented in the previous Section, we know that the reconstruction of the audio modulations we can get from the EEG is more correlated with those of the attended instrument than with those of the unattended one. We observed that this reconstruction is highly deteriorated but still "good enough" to discriminate among the attended and unattended sources. These two facts can be naturally exploited in an informed NMF-based sound source separation system, where the sources are decomposed into spectral patterns and corresponding activations. Our proposal is then to reconstruct the attended source's activations from the EEG using a linear regression model like the one used in the AAD experiments. Indeed, the NMF activations can be seen as modulations across time of specific spectral patterns found by the factorisation. Thus, they will represent a rough approximation of the TF representations used in our previous AAD experiments.

One advantage over other source separation models is that NMF makes it possible to incorporate additional information about the sources directly in his optimisation cost without

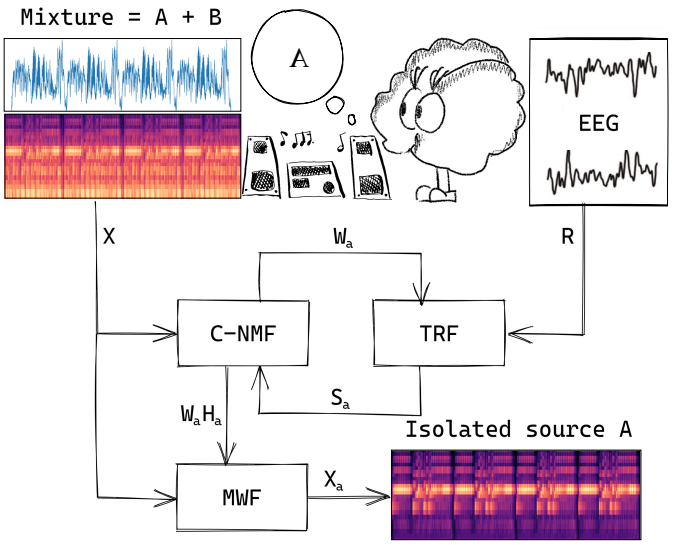

Fig. 4. Proposed scheme: the target instrument's activations are reconstructed from the listener's multi-channel EEG using a pre-trained TRF backward model. They are then used to guide the mixture's factorisation and cluster the components into the respective sources (C-NMF). At the same time, the decoding model is updated every certain number of C-NMF iterations to adapt to the observed signal. After convergence, the dictionary and the activations related to the attended source are used to obtain the Wiener filter soft-mask.

requiring a data intensive training phase. The additional information at our disposal is represented by the attended source's temporal activations for a given set of spectral patterns representing that source reconstructed from the EEG. Since those reconstructed activations are significantly deteriorated, it is hard to use them directly. Nevertheless, these reconstructions are good enough to discriminate the attended instrument from the unattended one. In the proposed Contrastive-NMF, this "contrast" is used to guide the separation. The factorisation and the decoding are learnt jointly. The target instrument's activations are reconstructed from the multi-channel EEG at first using a pre-trained TRF backward model. Then they are used to guide the mixture's factorisation and cluster the components into the respective sources. At the same time, the decoding model is updated every certain number of NMF iterations to adapt to the observed signal. A good initialisation of the TRF can be learned from a small training set of solos and corresponding EEG recordings from the same subject.

A. NMF-based audio source separation

The proposed Contrastive-NMF (C-NMF) is a novel variant of *Non-negative matrix factorization (NMF)*, a technique for data decomposition which has been very popular in many audio inverse problems such as source separation, enhancement or transcription as it is able to unmix superimposed spectral components [59]. Among other factorization techniques (e.g. Principal Component Analysis (PCA), Independent Component Analysis (ICA)), NMF distinguishes itself because of its nonnegativity constraints which lead to a part-based representation of the data that is *interpretable* [60].
In the case of single-channel audio source separation, one can assume that an audio signal $x(t)$ at time sample t is given by the linear mixture of J sources $s_j(t)$:

$$x(t) = \sum_{j=1}^{J} s_j(t). \tag{2}$$

Observing the mixture $x(t)$, a source separation system aims to recover one or more sources $s_j(t)$ of interest. Such a mixture can be represented in matrix form through its magnitude spectrogram $\mathbf{X} \in \mathbb{R}_+^{M\times N}$, where M represents the number of frequency bins and N the number short-time Fourier transform (STFT) frames. $\mathbf{X}$ can be factorized into two unknown matrices $\mathbf{W}$ and $\mathbf{H}$ such that $\mathbf{X} \approx \mathbf{WH}$, where the columns of $\mathbf{W} \in \mathbb{R}_+^{M\times K}$ are interpreted as non-negative audio spectral patterns, expected to correspond to different sources and the rows of $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ as their temporal activations. Usually, one refers to $\mathbf{W}$ as the *dictionary* and to $\mathbf{H}$ as the *activation matrix*. When K, namely the rank of the factorization, is much smaller than M, $\mathbf{WH}$ represents a *low-rank approximation* of the data matrix $\mathbf{X}$ [59]. The factorisation can be achieved by minimizing a cost function as the following:

$$\begin{cases} C(\mathbf{W},\mathbf{H}) = \underbrace{D(\mathbf{X}|\mathbf{WH})}_{\text{audio factorization}} + \underbrace{\mu\|\mathbf{H}\|_1 + \beta\|\mathbf{W}\|_1}_{\text{sparsity}} \\ \mathbf{W},\mathbf{H} \geq 0. \end{cases} \tag{3}$$

Usually, for the mixture reconstruction β-divergences are used, which have been very popular for audio inverse problems. It is also common to impose a sparsity constrain on both $\mathbf{W}$ and $\mathbf{H}$ using an ℓ_1 regularization controlled by the hyperparameters μ and β, respectively, to improve the source modelling. In fact, music is often given by a repetition of a few audio patterns, thus we can easily assume that the activations are sparse [61]. The same can be assumed for the spectral patterns as there is only a low probability that two given sources are highly activated in the same set of frequency bins [62].
At this point, the separation problem reduces to the assignment of each component to the corresponding source j. Then, the complex-valued spectrogram $\mathbf{S}_j$ of each source can be estimated by Wiener filtering as:

$$\mathbf{S}_j = \frac{\mathbf{W}_j\mathbf{H}_j}{\mathbf{WH}} \otimes \tilde{\mathbf{X}}, \tag{4}$$

where the element-wise division $(\mathbf{W}_j\mathbf{H}_j)/(\mathbf{WH})$ is the soft mask associated to source j and $\tilde{\mathbf{X}}$ is the complex spectrogram of the mixture. $\otimes$ denotes an element-wise multiplication. Through an inverse STFT one can recover the corresponding audio signal in the time domain.
What we have described so far is the so-called *unsupervised NMF*, *i.e.* a blind signal decomposition where both the dictionary and the activations are estimated from the mixture [59]. However, in real music compositions a source plays several notes with different pitches and it might be hard to represent it with a single component. Moreover, two sources may be represented by similar components as they might overlap and be highly correlated. Therefore, the component assignment might be hard and requires specific classification or clustering techniques. In such a complex situation, the factorization needs to be "guided" by incorporating prior information about the sources to return a meaningful representation [63].

Starting from the unsupervised formulation, one can incorporate prior knowledge directly in the optimisation cost, e.g., through hard or soft constrains, specific regularizers, pretrained dictionaries, or forcing the elements of $\mathbf{W}$ and/or $\mathbf{H}$ to follow a given distribution [61]. Particularly interesting is the multimodal scenario, where one has access to multiple views of the same phenomenon (e.g., video, motion capture data, score) which are synchronized with the audio. Seichepine et al. [64], for instance, propose to impose the equality (hard constraint) or the similarity (soft constraint) of the source activations in the two modalities. This is not applicable in our case as the time activations we can reconstruct from the EEG are very deteriorated, making it hard to use them directly. Nevertheless, these reconstructions are "good enough" to discriminate the attended instrument from the unattended one, leading to a "contrast" that can guide the separation.

B. A novel NMF variant: Contrastive-NMF (C-NMF)

The general idea of discriminating sources according to some criterion for NMF-based audio source separation was already explored in the past but most of the proposals refer to fully supervised or semi-supervised scenarios, where the basis functions are learned in a training phase. Weninger et al. [65] and Kitamura at al. [66] propose to learn basis matrices that are as much discriminative as possible to have unique spectral templates for each source. Grais et al. [67] propose to minimize the cross-coherence between dictionaries belonging to different sources, while Chung at al. [68] to learn a factorization so that each basis is classified into one source. Kumar et al. [69] propose a max-margin framework, where the projections are learned to maximize an SVM classifier's discriminative ability. Within this work, instead, the projections are learned by an unsupervised NMF to maximize the discrimination ability of a TRF model. Specifically, the proposed cost aims at decomposing the audio spectrogram while maximizing the similarity of the EEG-derived activations with the audio-derived ones for the target source and minimizing it for the interference sources. Thanks to this formulation, the components resulting from the decomposition should already be clustered into the target and interference sources.

Let us analyze the novel cost function. Considering a mixture $x(t)$ given by the linear mixing of the attended source $s_a(t)$ and some interferers $s_u(t)$, let $\mathbf{W}_a$ be a sub-dictionary of $\mathbf{W}$ containing a set of bases representing source $s_a(t)$ and $\mathbf{H}_a$ be their activations. $\mathbf{H}_a$ can be roughly approximated by $\mathbf{S}_a$ reconstructed from the time-lagged EEG response $\mathbf{R}$, the assumption being that it is likely to be more correlated with the NMF-derived activations of the attended source $\mathbf{H}_a$ than with the ones of the interferers $\mathbf{H}_u$. This contrast can be integrated in the unsupervised NMF cost function as follows:

$$\begin{cases} C(\mathbf{W},\mathbf{H}) = \underbrace{D(\mathbf{X}|\mathbf{W}\mathbf{H})}_{\text{audio factorization}} + \underbrace{\mu\|\mathbf{H}\|_1 + \beta\|\mathbf{W}\|_1}_{\text{sparsity}} + \\ \qquad - \underbrace{\delta(\|\mathbf{H}_a\mathbf{S}_a^T\|_F^2 - \|\mathbf{H}_u\mathbf{S}_a^T\|_F^2)}_{\text{contrast}} \\ \mathbf{W},\mathbf{H},\mathbf{S}_a \geq 0 \\ \|\mathbf{h}_{k:}\|_2 = 1, \|\mathbf{s}_{k:}\|_2 = 1. \end{cases} \tag{5}$$

$\mathbf{h}_{k:}$ and $\mathbf{s}_{k:}$ represent the rows of $\mathbf{H}$ and $\mathbf{S}_a$ respectively and are normalized to have unit ℓ_2 norm in order to minimize the effect of a scale mismatch between the modalities.

We derived the update rules for $\mathbf{H}$ and $\mathbf{W}$ using multiplicative update heuristics. They are given on line (10) and (12) of Algorithm 1 respectively.[2] This pseudo-code provides all the details of the algorithm: $\otimes$ and exponents denote element-wise operations, $\mathbf{1}$ is a matrix of ones whose size is given by context and $\mathbf{P}^-, \mathbf{P}^+ \in \mathbb{R}^{K\times N}$ denote the negative and positive parts of the auxiliary matrix $\mathbf{P}$ respectively (for more details see the derivation in the supplementary material). The inputs are the magnitude spectrogram of the mixture $\mathbf{X}$ and the time-lagged EEG data matrix $\mathbf{R}$ while the outputs are the two matrices $\mathbf{W}_a$ and $\mathbf{H}_a$ associated to the attended source. The hyperparameters to be tuned are μ, β and δ.

input : $\mathbf{X}, \mathbf{R}, \mu \geq 0, \beta \geq 0, \delta \geq 0, \gamma \in [0,1]$
output: $\mathbf{W}_a$, $\mathbf{H}_a$

1 $\mathbf{W}$, $\mathbf{H}$, $\mathbf{g}$ *initialization*
2 $\mathbf{H} \leftarrow \text{diag}(\|\mathbf{h}_{1:}\|^{-1}, ..., \|\mathbf{h}_{K:}\|^{-1})\mathbf{H}$ ▷ normalization
3 $\mathbf{W} \leftarrow \mathbf{W}\ \text{diag}(\|\mathbf{h}_{1:}\|, ..., \|\mathbf{h}_{K:}\|)$ ▷ re-scaling
4 $\Lambda = \mathbf{W}\mathbf{H}$
5 **repeat**
6 $\mathbf{S}_a \leftarrow \mathbf{g}^T\mathbf{R}$
7 $\mathbf{S}_a \leftarrow \text{diag}(\|\mathbf{s}_{1:}\|^{-1}, ..., \|\mathbf{s}_{K:}\|^{-1})\mathbf{S}_a$
8 **repeat**
9 $\mathbf{P} \leftarrow [-\mathbf{H}_a\mathbf{S}_a^T\mathbf{S}_a, \mathbf{H}_u\mathbf{S}_a^T\mathbf{S}_a]^T$
10 $\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T(\mathbf{X}\otimes\Lambda^{-1})+\delta\mathbf{P}^-}{\mathbf{W}^T\mathbf{1}+\mu+\delta\mathbf{P}^+}$
11 $\mathbf{H} \leftarrow \text{diag}(\|\mathbf{h}_{1:}\|^{-1}, ..., \|\mathbf{h}_{K:}\|^{-1})\mathbf{H}$
12 $\mathbf{W} \leftarrow \mathbf{W}\ \text{diag}(\|\mathbf{h}_{1:}\|, ..., \|\mathbf{h}_{K:}\|)$
13 $\Lambda = \mathbf{W}\mathbf{H}$
14 $\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{(\Lambda^{-1}\otimes\mathbf{X})\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T+\beta}$
15 $\Lambda = \mathbf{W}\mathbf{H}$
16 **until** *convergence*;
17 update $\mathbf{g}$
18 **until** *convergence*;
19 **return** $\mathbf{W}_a$, $\mathbf{H}_a$

Algorithm 1: Contrastive NMF pseudo-code

C. Experiments

The experiments are designed to evaluate weather the EEG information helps the separation process. However, to verify that the improvement is due to the EEG and not to the cost function's discriminative capacity, it was not enough to have the blind NMF as the only baseline. Therefore, we built a second baseline which consists of the Contrastive NMF to which meaningless side information is given. The meaningless side information consists of random activations sampled from a Gaussian distribution. To summarise, we tested three models:

1) Blind NMF (NMF);
2) Contrastive NMF + Random side activations (C-NMF-r);
3) Contrastive NMF + EEG-derived activations (C-NMF-e).

[2]For the detailed derivation, please refer to https://hal.telecom-paris.fr/hal-02978978/file/Update-rule-C-NMF.pdf

	Pop								Classical									
SDR [dB]	Guitar		Vocals		Drums		Bass		Oboe		Flute		Horn		Cello		Bassoon	
	Duo	Trio	Duo	Trio	Duo	Trio	Duo	Trio	Duo	Trio	Duo	Trio	Duo	Trio	Duo	Trio	Duo	Trio
NMF	3.4	1.9	2.3	5.4	-2.0	**7.8**	0.6	-12.5	4.4	**5.3**	**6.3**	**3.7**	**5.9**	**5.3**	5.5	6.3	**4.7**	-2.9
C-NMF-r	1.0	2.8	3.2	**5.6**	0.4	0.9	0.4	-14.9	3.9	-1.7	1.2	1.6	3.7	2.2	**7.3**	**6.6**	4.6	1.8
C-NMF-e	**4.4**	**3.4**	**3.8**	5.1	**5.6**	2.0	**5.2**	**3.9**	**5.4**	1.4	3.0	1.7	2.1	1.6	4.5	3.6	3.6	**3.7**
Mono	3.4	**3.5**	3.6	**5.2**	5.8	1.7	5.2	3.7	**5.5**	**4.8**	2.9	**2.1**	**2.3**	1.6	**4.9**	2.9	3.6	3.7
Stereo	**4.5**	3.4	**4.0**	3.2	**5.4**	**2.5**	**9.0**	**4.0**	4.9	-3.9	**3.0**	1.4	2.0	**2.3**	4.5	**4.1**	**4.5**	**3.9**

TABLE II

SDR SEPARATION RESULTS FOR DIFFERENT MODELS, ENSEMBLE TYPES AND INSTRUMENTS. THE METRICS ARE SHOWN IN DB AND ALL VALUES ARE MEDIANS OVER THE CORRESPONDING SUBSET OF THE TEST SET. IN THE LAST TWO ROWS, THE SDR RESULTS OF THE PROPOSED METHOD C-NMF-E ARE SPLIT FOR STEREO AND MONO LISTENING TESTS.

As the models are entirely unsupervised, the factorised components need to be assigned to each source before applying the multi-channel Wiener filter. In the two baselines, the components are clustered according to their Mel-frequency cepstral coefficient (MFCC) similarity. In the case of the C-NMF-e, the EEG information automatically identifies and gathers the target instrument components. Thanks to this we can reformulate the AAD problem exposed in Section IV, where we had access to the ground truth sources, differently. This time, the instrument which is predicted as being the attended one is the one that is automatically separated by the proposed source separation system. Specifically for our formulation, the attended instrument is the one represented by the $\mathbf{W}_a$ dictionary ad $\mathbf{H}_a$ activations.

For each method, NMF is run for 400 iterations while the TRF model is updated every 100 iterations of the C-NMF-e. For each method, the initialization of $\mathbf{W}$ and $\mathbf{H}$ is obtained by applying a blind NMF to the mixture for 200 iterations. For a given mixture, the initialization of the three models is the same to guarantee a fair comparison. As a cost function, we chose the Kullback-Leibler divergence. We learned a good initialization of the TRF model from a training set of solos (different from the ones used in the test mixtures) and corresponding EEG recordings for each subject and instrument. The Ridge parameter is set to be $\gamma = 0.1$ and the considered temporal context is $[0, 250]$ms post-stimulus as done in the experiments of Section IV. Note that in our preliminary work [27], only pop duets associated with the musical piece "mixtape" were analyzed. Here we consider all the pop and classical mixtures in the MAD-EEG dataset including also trios. Indeed, we generalized the C-NMF implementation so that it is valid for mixtures containing more than two instruments.

D. Evaluation

The models are evaluated using a standard metric in music source separation, *i.e.* the Signal-to-Distortion Ratio (SDR) expressed in dB and computed using BSSEval v4 [70], [71]. The metric is computed over the whole length of each music excerpt (around 24 seconds). In the tables below are reported median values. To assert the statistical significance of our model's improvement over the baselines, we opted for a non-parametric Wilcoxon test on the metrics' linear values. The considered significance levels are 5%, 1%, 0.1% and 0.01%.

It is worth noting that given the user-driven nature of the EEG-driven separation system, the performance not only depends on the algorithm but also on the subject's ability to properly attend to the target instrument.

E. Experimental results

a) Separation quality: In Table II, one can see the median SDR values for different methods, instruments and spatial rendering. As far as spatial rendering is concerned, it is important to keep in mind that the audio signal processed by the source separation system is always mono (*i.e.* the task is single-channel audio source separation). The "mono" and "stereo" results relate to the way the stimuli were played to the subjects which differently affects their EEG response.

It is immediate to see that the contrast derived from the EEG can improve the separation quality for all the pop instruments, especially when separated from duets. Particularly significant is the improvement over the blind baseline (NMF) for the drums (more than 7 dB).

It is also clear that the proposed model needs to be fed with meaningful side information and that the activations reconstructed with the TRF model are indeed meaningful. In fact, the same model informed with the random side information (C-NMF-r) performs significantly worse than the one fed with the EEG-derived contrast (drums and bass $p < 0.0001$, guitar $p < 0.01$, singing voice $p < 0.05$, Wilcoxon test). In general, the C-NMF-r model introduces lots of artefacts, even without removing the interferers. Moreover, the random side information can even fool the factorization leading to a degradation of the performance w.r.t. the blind NMF. Only in some rare cases (e.g. Vocals, drums, and cello), even with the random information, the proposed paradigm "guides" the separation indirectly by imposing that the $\mathbf{H}_a$ and $\mathbf{H}_u$ activations are different, leading to a little improvement over the blind NMF. Those results confirm the preliminary results obtained in [27], where only pop duets associated with the song "mixtape" were tested. Now, let us analyze the result related to trios and Classical mixtures.

The situation is different for Classical music instruments, where the improvement over the baselines is statistically significant only for the oboe's separation from duets and the bassoon's separation from trios. The main reason is that the blind NMF is already obtaining a good separation, as the Classical music mixtures of the MAD-EEG dataset can be too

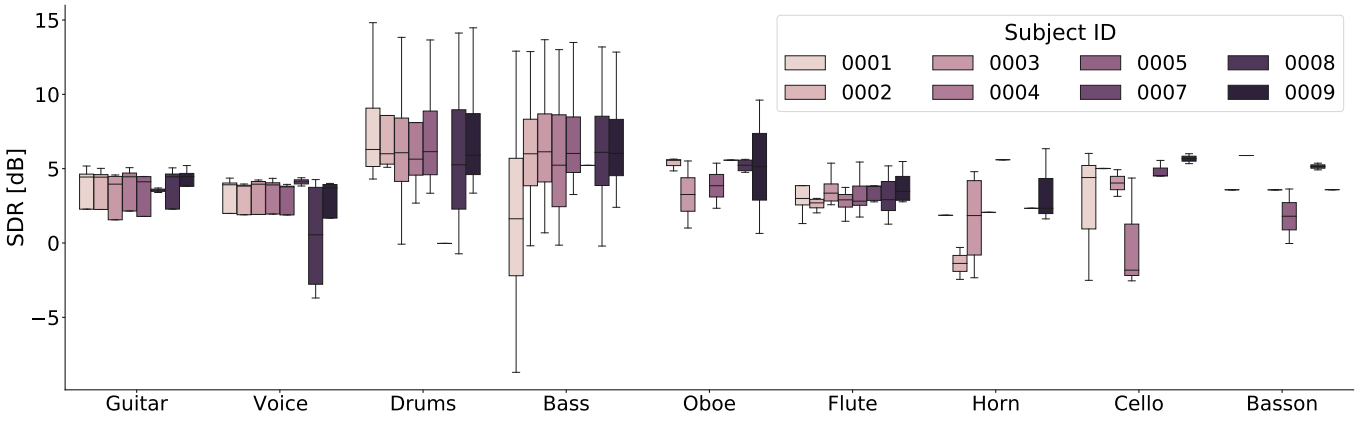

Fig. 5. Inter and intra-subject variability in duets: the SDR results are expressed in dB and different nuances of pink indicate different subjects.

easy to separate, and the EEG information helps especially in difficult cases (e.g. bass separation where the baselines suffer from the task's complexity). In such cases, it is hard to see the beneficial effects of the additional information.

Regarding trios, the SDR improvement over the baseline is statistically significant only for guitar, bass and the oboe. In general, separating trios is more challenging as the input SDR is much lower. Some previous works on AAD applied to speech [38], [72] showed that the attention task is more challenging for the listener with an increasing number of sources and noise levels. In practice, high noise levels can impact the listener's ability to segregate the source of interest leading to poor decoding quality. In music, this problem can be related to how much the attended instrument is in the foreground, *i.e.* its predominance. The input SDR of a given source w.r.t. the mixture can give us a rough idea of one source predominance, and for trios it is lower than for duets.

We remark that the results in Table II were obtained with $K = 16$, $\mu = \beta = 10$, and $\delta = 10^4$, set of values which was found to give good overall results. However, we observed that specific instruments and mixtures would need a specific hyperparameter tuning to maximize the performance. To give an example, by only reducing the value of μ from 10 to 1 when separating the oboe from trios, one can improve the SDR by more than 4 dB. This data-dependent behaviour of NMF scheme's hyperparameters was previously observed [11] and can be mitigated by allowing a user of the system to adjust the hyperparameter values typically through a knob/slider.

b) Spatial rendering: The stimuli were played to the subjects with two possible spatial renderings: one where both instruments are in the centre denoted as *mono* modality, and one where the instruments are spatialized, denoted as *stereo*. The last two rows of Table II show the results for these two different cases for all the instruments in the dataset. The results are differentiated w.r.t. the number of instruments in the mixture, and all values are medians over the test set. Intuitively, the stereo setting should help the subject in focusing on the target instrument as it makes it easier to localize it, leading to a better reconstruction of its activations and finally giving a better separation. We observed statistically significant improvement only for the pop instruments when listened to in duets (guitar $p < 0.01$, singing voice $p < 0.001$, drums and bass $p < 0.05$, Wilcoxon test).

c) Inter and intra-subject variability: Part of the high variance in the SDR performances is because different mixtures in the dataset can be more or less difficult for the separation system. However, most of the variance comes from the very high *inter* and *intra-subject variability*. The attention task may be more or less difficult for different subjects (inter-subject variability), which may depend on factors such as musical training and attention capacity [73]. Simultaneously, one single subject may perform differently throughout the experiment (intra-subject variability), maybe due to stress and fatigue that affect the attention level. These effects are evident in Figure 5, where the SDR results for duets are differentiated according to the participants involved in the experiment and the target instrument. Looking at Figure 5, one can realise that for a given instrument different subjects may behave very differently while for other ones they behave similarly. Moreover, for single instruments, subject's performance may span a wide SDR range. For example, regarding Classical instruments, one can observe that the intra-subject variability is generally lower while sometimes there is a clear inter-subject variability. This may be due to the subjects' unfamiliarity with some instruments like the French horn and the bassoon.

Another factor is that some instruments can be more difficult than others to follow. For instance, instruments like the bass and the drums, which usually guide the rhythm and tempo, are notably more difficult to be tracked, especially for non-professional musicians and this is reflected in the very high inter and intra-subject variability.

d) Attention decoding performances: Even if the SDR improvement is not systematic for all the instruments, the main advantage of the C-NMF-e model is that it gives an automatic clustering of the components and automatically enhances the attended source. Therefore, the instrument that is automatically separated by the proposed source separation system, *i.e.* the one represented by the $\mathbf{W}_a$ and $\mathbf{H}_a$, is predicted as being the attended one. It is an asset w.r.t. the baselines, which need an additional step to cluster the components and cannot auto-

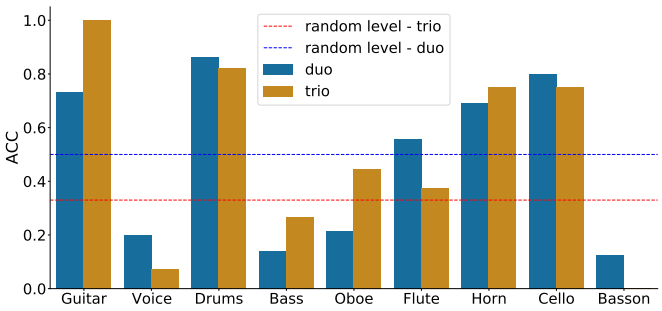

Fig. 6. Decoding accuracy for different instruments and ensemble types compared with the chance level for duets and trios respectively.

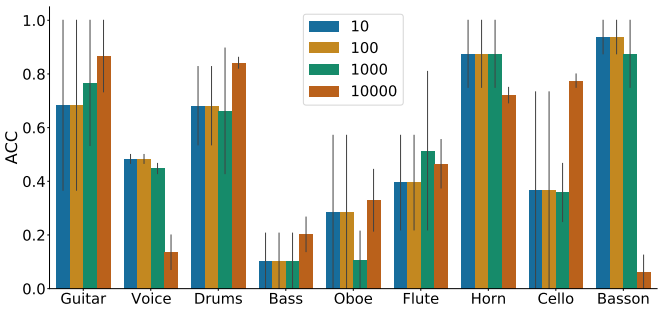

Fig. 7. Decoding accuracy for different instruments and values of the hyperparameter δ that weights the contrastive term.

matically identify the target source. In Figure 6, we report the AAD accuracy values for different instruments and ensemble types. The blue and the red lines represent the chance level for the duets and the trios. The accuracy is satisfactory and statistically above chance for four instruments: guitar, drums, French horn, and cello. For some other instruments (singing voice, bass, bassoon, and oboe), the accuracy is much below chance indicating that the contrastive term is always forcing them not to be represented by $\mathbf{W}_a$ and $\mathbf{H}_a$. The reason for this behaviour lies in a non-customized tuning of the δ parameter. We observed, for instance, that $\delta = 10^4$ causes a drop of the performances for the singing voice and the bassoon, which instead were much above chance with $\delta < 10^4$. As we said previously, this can be easily solved by a customized fine-tuning of the hyperparameters by the user.

e) Effect of hyperparameters: We first analyze the number of NMF components necessary to describe each instrument testing 4 values ($\{4, 8, 16, 32\}$). We observe that an increasing number of components improves the separation performance as it allows a more accurate description of the sources. As for the impact of the sparsity constraints imposed on $\mathbf{H}$ and $\mathbf{W}$ by μ and β, respectively, which in our experiments are set to be equal, we tested 4 values ($\{0, 0.1, 1, 10\}$), observing that higher μ and β improve the separation quality as it allows a better source modelling.
Lastly, we tested four reasonable values for δ ($\{10^1, 10^2, 10^3, 10^4\}$), which weights the contrastive term in the C-NMF cost function. We observed that increasing values of δ lead to significantly higher SDR for all the tested instruments except for the French horn, for which there is no significant difference ($p > 0.05$, Wilcoxon test). However, one has to be careful not to chose a too high value of δ, which may push to a trivial solution where the activations of the interferers $\mathbf{H}_u$ are set to zero and all the sources in the mixture are represented by the $\mathbf{W}_a$ and $\mathbf{H}_a$. This effect is reflected in the AAD accuracy reported in Figure 7, where the performance drops for $\delta = 10^4$ for the vocals and the bassoon. However, this effect is strictly instrument-dependent as for other instruments like the cello, the decoding accuracy becomes statistically better than chance only with $\delta = 10^4$ ($p < 0.0001$, randomization test).

VI. Conclusions

This paper describes a novel *neuro-steered music source separation* framework and conducts an extensive evaluation of the proposed system on the MAD-EEG dataset. We have analysed the impact of various aspects of the musical stimuli, such as the number and type of instruments in the mixture, the spatial rendering and the music genre, obtaining encouraging results. The results support the thesis that the EEG can guide and help a source separation system, especially in difficult cases where non-informed models struggle. Our ablation study, where the proposed model is informed with random side information, shows that the C-NMF formulation is not enough by itself but needs to be informed with meaningful side information and that the activations reconstructed with the TRF model are indeed meaningful. The main advantage of the C-NMF formulation is that it allows us to reformulate the AAD problem without access to the ground truth sources, paving the way for real-life applications. Moreover, it can be generalised and used with temporal activations derived from other modalities than the EEG (e.g., video, score, motion capture data) or from a manual annotation provided by the user (e.g. a sound engineer that annotates when the instrument of interest is active).
One limitation of the C-NMF model is that it needs customised fine-tuning of the hyperparameters for each test mixture to perform optimally. However, as the number of hyperparameters is limited, this can be easily mitigated by allowing users to adjust their values through a set of knobs/sidebars. Further, the EEG-driven C-NMF system has the intrinsic limitation of the subject-related variability: if the level of attention of the subject is not sufficient, this will inevitably impact the performance. Another factor that needs to be considered is musical expertise and training, which may help the subject while attending to an instrument.
We believe that this NMF variant is advantageous for neuro-steered music source separation. Indeed the available music-related EEG datasets are still costly and time-expensive to acquire, precluding the possibility to tackle the problem with data-driven approaches. Unsupervised NMF represents a powerful approach in such applications where there is no or a limited amount of training data. Moreover, additional information can be easily incorporated into the model cost function directly at test time.

REFERENCES

[1] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020.

[2] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint:1911.13254*, 2019.

[3] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix - a reference implementation for music source separation," *Journal of Open Source Software*, 2019.

[4] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *IEEE Int. Conf. on acoustics, speech and signal processing (ICASSP)*, 2017.

[5] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," 2017.

[6] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research." in *International Society for Music Information Retrieval Conf. (ISMIR)*, 2014.

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

[8] S. Ewert, B. Pardo, M. Müller, 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.

[9] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music." in *SAPA@ INTERSPEECH*, 2008.

[10] K. Schulze-Forster, C. Doire, G. Richard, and R. Badeau, "Weakly informed audio source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

[11] S. Parekh, S. Essid, A. Ozerov, N. Q. Duong, P. Pérez, and G. Richard, "Guiding audio source separation by video object information," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[12] M.-Q. Bui, V.-H. Duong, S.-P. Tseng, Z.-Z. Hong, B.-C. Chen, Z.-W. Zhong, and J.-C. Wang, "NMF/NTF-based methods applied for user-guided audio source separation: An overview," in *IEEE Int. Conf. on Orange Technologies (ICOT)*, 2016.

[13] T. Nakano, Y. Koyama, M. Hamasaki, and M. Goto, "Interactive deep singing-voice separation based on human-in-the-loop adaptation," in *Proc. 25th Int. Conf. on Intelligent User Interfaces (IUI)*, 2020.

[14] E. Chew, P. Taggart, and P. Lambiase, "Cardiac response to live music performance: Computing techniques for feature extraction and analysis," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[15] M. Müller, *Information retrieval for music and motion*. Springer, 2007, vol. 2.

[16] I. Sturm, "Analyzing the perception of natural music with EEG and ECoG," *Ph.D. thesis*, 2016.

[17] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[18] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in cognitive sciences*, vol. 12, no. 5, pp. 182–186, 2008.

[19] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[20] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention—focusing the searchlight on sound," *Current opinion in neurobiology*, vol. 17, no. 4, pp. 437–455, 2007.

[21] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex," *Journal of neurophysiology*, 2009.

[22] C. Han, J. O'Sullivan, Y. Luo, J. Herrero, A. D. Mehta, and N. Mesgarani, "Speaker-independent auditory attention decoding without access to clean speech sources," *Science advances*, vol. 5, no. 5, p. eaav6134, 2019.

[23] N. Das, J. Zegers, T. Francart, A. Bertrand *et al.*, "EEG-informed speaker extraction from noisy recordings in neuro-steered hearing aids: linear versus deep learning methods," *BioRxiv*, 2020.

[24] A. Aroudi and S. Doclo, "Cognitive-driven binaural beamforming using EEG-based auditory attention decoding," *IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP)*, vol. 28, pp. 862–875, 2020.

[25] M. S. Treder, H. Purwins, D. Miklody, I. Sturm, and B. Blankertz, "Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification," *Journal of neural engineering*, vol. 11, no. 2, p. 026009, 2014.

[26] G. Cantisani, S. Essid, and G. Richard, "EEG-based decoding of auditory attention to a target instrument in polyphonic music," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

[27] ——, "Neuro-steered music source separation with EEG-based auditory attention decoding and contrastive-NMF," in *IEEE Int. Conf. on acoustics, speech and signal processing (ICASSP)*, 2021.

[28] G. Cantisani, G. Trégoat, S. Essid, and G. Richard, "MAD-EEG: an EEG dataset for decoding auditory attention to a target instrument in polyphonic music," in *Proc. Workshop on Speech, Music and Mind (SMM19)*, 2019, pp. 51–55.

[29] J. A. O'sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2014.

[30] J. A. O'Sullivan, R. B. Reilly, and E. C. Lalor, "Improved decoding of attentional selection in a cocktail party environment with eeg via automatic selection of relevant independent components," in *37th Ann. Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015.

[31] S. Van Eyndhoven, T. Francart, and A. Bertrand, "EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses." *IEEE Trans. Biomed. Engineering*, vol. 64, no. 5, pp. 1045–1056, 2017.

[32] W. Pu, J. Xiao, T. Zhang, and Z.-Q. Luo, "A joint auditory attention decoding and adaptive binaural beamforming algorithm for hearing devices," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[33] E. Ceolini, J. Hjortkjær, D. D. Wong, J. O'Sullivan, V. S. Raghavan, J. Herrero, A. D. Mehta, S.-C. Liu, and N. Mesgarani, "Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception," *NeuroImage*, 2020.

[34] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, p. 233, 2012.

[35] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex," *PLoS biology*, vol. 10, no. 1, p. e1001251, 2012.

[36] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli," *Frontiers in human neuroscience*, vol. 10, p. 604, 2016.

[37] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *European journal of neuroscience*, vol. 31, no. 1, pp. 189–193, 2010.

[38] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *Neuroimage*, vol. 156, pp. 435–444, 2017.

[39] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, 2015.

[40] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor, "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech," *Current Biology*, vol. 28, no. 5, pp. 803–809, 2018.

[41] A. Aroudi, D. Marquardt, and S. Daclo, "EEG-based auditory attention decoding using steerable binaural superdirective beamformer," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[42] A. Aroudi and S. Doclo, "Cognitive-driven binaural LCMV beamformer using EEG-based auditory attention decoding," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[43] N. Das, S. Van Eyndhoven, T. Francart, and A. Bertrand, "Eeg-based attention-driven speech enhancement for noisy speech mixtures using n-fold multi-channel wiener filters," in *25th European Signal Processing Conf. (EUSIPCO)*, 2017.

[44] J. O'Sullivan, Z. Chen, S. A. Sheth, G. McKhann, A. D. Mehta, and N. Mesgarani, "Neural decoding of attentional selection in multi-speaker environments without access to separated sources," in *39th Ann. Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017.

[45] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[46] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Trans. on Audio,*

Speech and Language Processing (TASLP), vol. 26, no. 4, pp. 787–796, 2018.
[47] W. Pu, P. Zan, J. Xiao, T. Zhang, and Z.-Q. Luo, "Evaluation of joint auditory attention decoding and adaptive binaural beamforming approach for hearing devices with attention switching," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
[48] B. Kaneshiro, D. T. Nguyen, J. P. Dmochowski, A. M. Norcia, and J. Berger, "Naturalistic music EEG dataset - hindi (NMED-H)," 2016.
[49] S. Losorelli, D. T. Nguyen, J. P. Dmochowski, and B. Kaneshiro, "NMED-T: A tempo-focused dataset of cortical and behavioral responses to naturalistic music," 2017.
[50] J. Appaji and B. Kaneshiro, "Neural tracking of simple and complex rhythms: Pilot study and dataset," 2018.
[51] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Towards music imagery information retrieval: Introducing the openmiir dataset of EEG recordings from music perception and imagination." in *International Society for Music Information Retrieval Conf. (ISMIR)*, 2015.
[52] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Trans. on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.
[53] N. Das, T. Francart, and A. Bertrand, "Auditory attention detection dataset kuleuven," 2020.
[54] World Medical Association, "World medical association declaration of helsinki: ethical principles for medical research involving human subjects," *JAMA*, vol. 310, no. 20, pp. 2191–2194, 2013.
[55] E. M. Z. Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, J. Z. Simon *et al.*, "Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party"," *Neuron*, vol. 77, no. 5, pp. 980–991, 2013.
[56] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components - a tutorial," *NeuroImage*, vol. 56, no. 2, pp. 814–825, 2011.
[57] E. W. Noreen, *Computer-intensive methods for testing hypotheses*. Wiley New York, 1989.
[58] A. Yeh, "More accurate tests for the statistical significance of result differences," in *Proc. of the 18th Conf. on Computational linguistics*. Association for Computational Linguistics, 2000.
[59] C. Févotte, E. Vincent, and A. Ozerov, "Single-channel audio source separation with NMF: divergences, constraints and algorithms," *Audio Source Separation*, pp. 1–24, 2018.
[60] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.
[61] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.
[62] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on signal processing*, vol. 52, no. 7, 2004.
[63] 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.
[64] N. Seichepine, S. Essid, C. Févotte, and O. Cappé, "Soft nonnegative matrix co-factorization." *IEEE Trans. Signal Processing*, vol. 62, no. 22, pp. 5940–5949, 2014.
[65] F. Weninger, J. L. Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," in *15th Ann. Conf. of the Int. Speech Communication Association*, 2014.
[66] D. Kitamura, N. Ono, H. Saruwatari, Y. Takahashi, and K. Kondo, "Discriminative and reconstructive basis training for audio source separation with semi-supervised nonnegative matrix factorization," in *IEEE Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016.
[67] E. M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation." in *Interspeech*, 2013.
[68] H. Chung, E. Plourde, and B. Champagne, "Discriminative training of NMF model based on class probabilities for speech enhancement," *IEEE Signal Processing Letters*, vol. 23, no. 4, 2016.
[69] B. V. Kumar, I. Kotsia, and I. Patras, "Max-margin non-negative matrix factorization," *Image and Vision Computing*, vol. 30, no. 4-5, 2012.
[70] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP)*, vol. 14, no. 4, 2006.
[71] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Int. Conf. on Latent Variable Analysis and Signal Separation*. Springer, 2018.
[72] N. Das, A. Bertrand, and T. Francart, "EEG-based auditory attention detection: boundary conditions for background noise and speaker positions," *Journal of neural engineering*, vol. 15, no. 6, p. 066017, 2018.
[73] G. M. Di Liberto, C. Pelofi, S. Shamma, and A. de Cheveigné, "Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 361–364, 2020.

Giorgia Cantisani graduated with a Master's Degree in Biomedical Engineering from the Polytechnic University of Turin. Currently a Marie Skłodowska-Curie fellow within the MIP-Frontiers training network, she is pursuing a Ph.D. at Télécom Paris on the topic of multimodal music source separation. Her research interests range from music information retrieval (MIR) to neural signal processing. In particular, she is interested in the analysis of brain responses to music and how these can be used to guide and inform MIR tasks.

Slim Essid received his state engineering degree from the Ecole Nationale d'Ingénieurs de Tunis, Tunisia, in 2001, his M.Sc. (D.E.A.) degree in digital communication systems from the Ecole Nationale Supérieure des Télécommunications, Paris, France, in 2002, his Ph.D. degree from the Universite Pierre et Marie Curie (UPMC), Paris, France, in 2005, and his Habilitation a Diriger des Recherches degree from UPMC in 2015. He is a professor in Télécom Paris's 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.

Gaël Richard (SM'06, F'17) received the State Engineering degree from Télécom Paris, France in 1990, and the Ph.D. degree from University of Paris-Sud, in 1994. He then spent two years at the CAIP Center, Rutgers University, Piscataway, NJ, in the Speech Processing Group of Prof. J. Flanagan. From 1997 to 2001, he successively worked for Matra-Nortel, France, and for Philips, France. He then joined Télécom Paris, where he is now a Full Professor and Head of the Image, Data, Signal department. Co-author of over 250 papers and inventor in 10 patents, his research interests are mainly in the field of speech and audio signal processing and include source separation, machine learning methods for speech/audio/music signals and music information retrieval.