A TENTATIVE TYPOLOGY OF AUDIO SOURCE SEPARATION TASKS

Emmanuel Vincent
Xavier Rodet
Axel Röbel

IRCAM, Analysis-Synthesis Group
1, place Igor Stravinsky
F-75004 PARIS
FRANCE
emmanuel.vincent@ircam.fr

Cédric Févotte
Éric Le Carpentier

IRCCyN, ADTS Group
1, rue de la Noë – BP 92 101
F-44321 NANTES CEDEX 03
FRANCE
cedric.fevotte@irccyn.ec-nantes.fr

Rémi Gribonval
Laurent Benaroya
Frédéric Bimbot

IRISA, METISS Project
Campus de Beaulieu
F-35042 RENNES CEDEX
FRANCE
remi.gribonval@irisa.fr

ABSTRACT

We propose a preliminary step towards the construction of a global evaluation framework for Blind Audio Source Separation (BASS) algorithms. BASS covers many potential applications that involve a more restricted number of tasks. An algorithm may perform well on some tasks and poorly on others. Various factors affect the difficulty of each task and the criteria that should be used to assess the performance of algorithms that try to address it. Thus a typology of BASS tasks would greatly help the building of an evaluation framework. We describe some typical BASS applications and propose some qualitative criteria to evaluate separation in each case. We then list some of the tasks to be accomplished and present a possible classification scheme.

1. INTRODUCTION

Blind Audio Source Separation (BASS) has been a subject of intense work during the latest years. Several models have emerged, such as Independent Component Analysis (ICA) [1] and Sparse Decompositions (SD) [2], and it is now more or less well known how to solve the separation problem under these models with efficient and robust algorithms. However BASS is not just about solving some tractable model (*e.g.* finding independent or sparse components), it is about recovering results that make sense according to the target application.

BASS covers many applications, such as high quality separation of musical sources, signal/speech enhancement, multimedia documents indexing, speech recognition in a "cocktail party" environment or source localization for auditory scene analysis. Depending on the application, BASS algorithms have to address different tasks. For example, some applications require finding the number of sources given the observations, and others require recovering the source signals given the observations and the structure of the mixing system.

A given separation algorithm may perform well on some tasks and poorly on others. Depending on the task, various factors affect the difficulty of the separation, and distinct criteria may be used to evaluate the performance of an algorithm, and compare it to other algorithms.

The first step to determine which task(s) a given separation algorithm may achieve is to list and classify some of the interesting tasks. As tasks and applications are related, this implies to list and classify typical applications of BASS too. We attempt to address these questions in this paper. As a further step, we propose in a companion paper [3] some numerical criteria to evaluate the performance of BASS algorithms on some of these tasks.

In Section 2, we present two large classes of BASS applications. In Sections 3 and 4, we give some examples among each class and we identify some candidate qualitative criteria to measure separation quality and separation difficulty. In Section 5, we list some of the tasks to be addressed by BASS algorithms and we present a possible typology.

Let us emphasize that this paper should be considered as a preliminary proposal that does not contain final results but rather presents some thoughts on the definition, the typology and the evaluation of BASS tasks. We hope the source separation community will consider these topics more closely, so that the construction of an agreed-upon evaluation framework for BASS algorithms will become possible.

This work is part of a Junior Researchers Project funded by GdR ISIS (CNRS). See `http://www.ircam.fr/anasyn/ISIS/` for some insights on the Project.

2. BASS APPLICATIONS

An important distinction that can be made among BASS applications is whether the output of the algorithm is a set of extracted sources that are intended to be listened to or not. We term these two categories Audio Quality Oriented (AQO) and Significance Oriented (SO) applications.

AQO applications extract sources that are listened to, straight after separation or after some post-processing audio treatment. Most of the literature focuses on this goal by using ICA- and SD-related methods (see [4] for a review of ICA methods applied to audio signals). Some criteria for separation quality and separation difficulty have been proposed in [5, 6], and we propose some others in our companion paper [3].

In SO applications, the extracted sources and/or mixing parameters are processed to obtain information at more abstract levels, in order to find a representation of the observations related to human perception. For instance, looking for the number and the kind of the instruments in a musical excerpt enters the scope of SO separation. Separation quality criteria are generally less demanding than in AQO applications because the aim of SO separation is only to keep specific features of the sources. Thus, a rough separation may be sufficient (possibly with high distortion), depending on the robustness of the subsequent feature extraction algorithms.

An important remark is that separation and information extraction do not have to be separated processes. For example, the auditory system uses *a priori* and contextual information to perform separation, which means recognition can help separation.

For the sake of clarity, let us define the few notations used in the rest of this paper. The general (possibly convolutive) BASS problem $x_i(t) = \sum_{j=1}^{n} (a_{ij} \star s_j)(t)$ is expressed using the matrix of filters formalism as $\mathbf{x} = \mathbf{A} \star \mathbf{s}$, where $\mathbf{s}$ is the vector of the n sources $(s_j)_{j=1}^{n}$, $\mathbf{x}$ is the vector of the m observations $(x_i)_{i=i}^{m}$, $\mathbf{A}$ is the matrix of mixing filters. Note that we limit ourselves to linear time-invariant mixing systems. A similar analysis could be carried out extending the model to non-linear time-variant systems to take into account the dynamic compression applied to radio broadcasts or the spatial movements of the sources for example.

3. AUDIO QUALITY ORIENTED SEPARATION

Within AQO applications, we can distinguish two major families of applications. The first category is related to applications where we are interested in each individual extracted source, while the second one corresponds to applications where the goal is to listen to a new mixture of the sources.

3.1. One versus all

The *one versus all* problem consists in extracting one sort of sound (the target source s_j) from a mixture. Generally the other sources are considered as noise.

Some examples include restoration of old monophonic musical recordings [7], speech de-noising and de-reverberation for auditory protheses or mobile phones [8] and extraction of some interesting sounds in a polyphonic musical excerpt for electronic music creation.

In this context, a "good" separation requires estimating the source with a high Signal to Noise Ratio (SNR). The SNR criterion can be modified to model the specificities of hearing, such as masking phenomena, as does the criterion introduced in [9]. Other criteria can be used to evaluate the separation quality when an exact estimation is not needed. This is often the case when indeterminacies arise due to convolutive mixing. For some applications, it may be sufficient to recover some filtered versions of the target source s_j and not s_j itself [10, 11]. The "naturalness" of these versions could be measured by criteria like timbre distortion [12] or comparisons with a database of room impulse responses. For other applications, one may wish to extract the contribution of s_j to each sensor [13], that is to say to estimate the multichannel signal $\mathbf{s^j_{img}} = A \star [0, \ldots, 0, s_j, 0, \ldots, 0]^T$. In such cases, quality criteria may have to take into account the difference between the perceived spatial direction of s_j when listening to $\hat{\mathbf{s}}^{\mathbf{j}}_{\mathbf{img}}$ and to $\mathbf{x}$.

The problem of extracting several sources in order to listen to them separately falls in the *one versus all* category too. For each source, a global SNR can be computed and can be decomposed into the contributions of crosstalk (remainder of the other sources), additive noise and algorithmic artifacts [3].

The *one versus all* problem may be tackled at various difficulty levels. The algorithms are influenced by the number of sources, the number of sensors, the noise level, the dependency between the sources, the kind of mixing (instantaneous *vs* convolutive), etc. The blind case is usually addressed by ICA [1] and SD [2]. Other algorithms can handle *a priori* information, like a model of the playing musical instrument and its musical score used in [14] or the video of the lips corresponding to a noisy speech signal used in [8].

3.2. Audio scene modification

Audio scene modification consists in obtaining a new mixture $\mathbf{x_{remix}} = \mathbf{B} \star [f_1(s_1), \ldots, f_n(s_n)]^T$, for example by

extracting all the sources $(\hat{s}_j)_{j=1}^n$ from the original observations $\mathbf{x}$, applying an adapted audio processing f_j to each source and remixing the tracks using a possibly different mixing matrix $\mathbf{B}$, in order to listen to the result $\hat{\mathbf{x}}_{\mathbf{remix}}$. Let us note that prior extraction of each source is not a requirement in such applications, it is only a convenient way to describe the desired result and a possible way to achieve it.

Examples include re-mastering of a stereo CD, blind multichannel diffusion of stereo recordings [15], spatial interpolation [16] and cancellation of the voice in a song for "automatic karaoke".

Evaluation of the separation results may rely on calculating the SNR of the estimated remixed scene *w.r.t.* the expected result (that is to say the scene constructed by remixing the true sources). Depending on the signal "zones" affected by post-processing and remixing, quality criteria may be a little less restrictive than for the *one versus all* problem. For example, when the purpose is to increase slightly the "presence" of an instrument in a CD, distortion or crosstalk in the extracted instrument won't account for much in the final result. Indeed, the other sources will most likely mask the zones containing crosstalk after remixing, and even larger zones since auditory masking effects usually come into play [9].

Difficulties encountered by the algorithms include those of the *one versus all* problem : they are influenced by the number of sources, the number of sensors, etc. But the amount of change introduced by the intermediate audio processing f_j and the new number of channels obtained after remixing with $\mathbf{B}$ also play a role. For instance, given a mono recording, it is more difficult to cancel one of its sources or to broadcast each of its sources on one or more channel(s) than to augment slightly the volume of one of its sources. Like in the *one versus all* problem, algorithms can also use *a priori* information or not. The use of such information is to determine satisfactory $\mathbf{B}$ and f_j to achieve the desired effect. In blind remixing, this can only be done by relying on directly computable features, such as the direction or the instantaneous power of the sources. When models of the sources are available, it becomes conceivable to name each source (*i.e.* to label it with the right model), so that it can undergo more specific treatments. One can think of raising the level of "the voice" in a recording if a model of "voice" is available.

4. SIGNIFICANCE ORIENTED SEPARATION

SO applications aim at retrieving source features and/or mixing parameters to describe complex audio signals at various cognitive levels, focusing on different aspects of sound [17]. The purpose of finding an exhaustive description of a complex audio scene is called Auditory Scene Analysis (ASA) [18]. Most SO applications can be seen as by-products of ASA.

The main applications of SO separation concern the indexing of audiovisual databases and the construction of intelligent hearing systems. Depending on the application, one may need low level descriptive elements, high level ones or both. Some examples of descriptive elements are the score of each instrument in a musical excerpt [19], the text pronounced by a speaker in a noisy environment [20, 21], or the spatial position of the sources *w.r.t.* the sensors in a "real world" recording [22], etc. Other descriptions consist in telling the name of each instrument and the musical genre [23], identifying the speaker [24], linking audio sources and corresponding visual objects on a video [25], etc.

The purpose of S0 separation is to preserve as much as possible the features used to compute the descriptive elements. To evaluate the quality of a global description consisting of many descriptive elements, the quality of each element is first evaluated separately by a distinct criterion. The quality of continuous-valued descriptive parameters, such as the positions of the sources, is measured by simple distances. The evaluation of discrete-valued descriptive parameters, such as the name of the instrument is done by calculating misclassification or recognition error rates [26]. The quality of the whole audio description may then be expressed by a weighted combination of all these criteria. When such a weighting is hard to choose objectively, it may be preferable to conduct a series of listening tests to obtain a global separation/description grade [27, 18].

Criteria for the separation difficulty depend on the application. The number of sensors and their selectivity and the amount of reverberation in the environment affect the retrieval of the mixing parameters. Source classification is more or less difficult according to the number of classes to recognize and the robustness of the features calculation. For some applications, a real-time constraint is also needed.

5. AUDIO SOURCE SEPARATION TASKS

As we have shown, AQO and SO separation are used in many different applications, each one having its own evaluation criteria. However, these applications correspond to a smaller number of tasks to be accomplished by BASS algorithms, depending on the relevant objects in the model, the kind of mixing and the amount of available information.

A task is specified by the nature of the objects that the algorithm takes as an input, the nature of its output, and

Task	Input	Output
Counting		$\hat{n}$
Blind mixing identification	structure of $\mathbf{A}$ (*not always*)	$\mathbf{\hat{A}PD}$
Blind source extraction	structure of $\mathbf{A}$	$\mathbf{PD\hat{s}}$ or $\{\mathbf{\hat{s}^j_{img}}\}_{j=1}^n$
Blind remixing	structure of $\mathbf{A}$, generic $\mathbf{B}$ and $(f_j)_{j=1}^n$	$\mathbf{\hat{x}_{remix}}$
Detection	sources models $(\mathcal{M}_k)_{k=1}^K$	number $\hat{\epsilon}_k$ of sources following $\mathcal{M}_k$
Identification Representation	model of $\mathbf{s}$	description of $\mathbf{s}$ and $\mathbf{A}$
Source extraction	model+description of $\mathbf{s}$ and $\mathbf{A}$	$\mathbf{\hat{s}}$ or $(\mathbf{\hat{s}^j_{img}})_{j=1}^n$
Remixing	model+description of $\mathbf{s}$ and $\mathbf{A}$, adapted $\mathbf{B}$ and $(f_j)_{j=1}^n$	$\mathbf{\hat{x}_{remix}}$

Table 1. Some BASS tasks (see Section 5 for comments and previous Sections for notations)

a qualitative description of how the quality of the output should be assessed.

Table 1 lists some tasks according to the input-output scheme of the algorithms using the notations of the previous Sections. In the 'Input' column, we only list what comes in addition to the observations $\mathbf{x}$. The (qualitative) evaluation criteria are implicit, but it is indeed a crucial step to define relevant and agreed upon procedures to evaluate the performance of an algorithm on a task.

Note that for some tasks the well-known indeterminacies of the BASS problem are explicitly expressed in the description of the output, using a permutation matrix $\mathbf{P}$ and a diagonal matrix of filters $\mathbf{D}$, or using the different delimiters $()$ and $\{\}$ for ordered and unordered sets.

For the remixing task, the input include the new mixing matrix and audio processing to perform on each source. In the non-blind case, it is conceivable to specify this intrinsically so that a given processing is performed on a source identified by some model ("the piano" for example). This is denoted by 'adapted' B and f. However, in the blind situation, it seems very hard to specify the nature of the source that should undergo a given audio processing. One may thus be restricted to specifying a source in terms of "the loudest source" or "the leftmost source" on a stereo recording. In Table 1, this is denoted by 'generic' B and f.

We tried to choose the names of the tasks in correspondence to what is used in the literature. For example, the Blind Mixing Identification task contains as a special case what is usually called Blind System Identification [28]. The Detection task is close to the Verification problem in speaker recognition [29] and to the Classification problem in audiovisual database indexing. The Remixing task includes the Cancellation problem, which consists in cancelling one source in the mixture.

The main distinction we propose between tasks is whether models of the sources (generally learned from a database of samples of the sources) are available or not. The difference between blind tasks and their semi-blind "counterparts" is indeed quite important.

However, contrary to other contributions concerning the subject [5, 6], we group in each task the instantaneous mixing case and the convolutive one. In fact, we believe the various possible structures of $\mathbf{A}$ should be considered as various difficulty criteria (or subtasks) for the same overall task, rather than separate tasks. By structure of $\mathbf{A}$, we mean information such as the number of sources or the length of the mixing filters (simple gain, gain-delay, short FIR, IIR with few parameters). For some tasks, this structure is given as input to the algorithms.

For some non blind tasks, we also group the problems where a model of the sources is given and those where a description of the sources is also available. The term model covers all sorts of general signal models, such as the hidden Markov models used in [8], the modified additive models used in [12] or even a physical model of the source instrument [30]. Source models can also contain learned information about source interaction, for example parameters describing the degree of independence between them [31]. There can also be models of the mixing system. In most cases, the definition of a task does not include a specific type of model that an algorithm can rely on in order to solve the task. Generally, the algorithm is trained –prior to running on "live" data– using some training samples of each source. In this context, a description may be any kind of knowledge that restricts the models depending on the particular piece of signal considered : a temporal segmentation, a musical score, the size of the recording room. As for the distinction between instantaneous and convolutive problems, we believe that giving or not this kind of descriptions to the algorithm is facing different difficulty levels inside an overall task, but not separate tasks. This puts together rather different problems, for example extracting a piano and a violin playing together with the only information that there are a piano and a violin, or performing the same extraction

knowing their scores. However, these are the extreme cases among many intermediate assumptions. Sometimes, only one of the sources is learned and only an imperfect score is available, like in [14].

6. CONCLUSION

In this paper, we described some of the most typical applications encountered in the field of BASS, and we proposed to group these applications into two main categories : AQO (Audio Quality Oriented) and SO (Significance Oriented) separation. AQO applications aim at extracting sources for a listening purpose, whereas in SO applications the extracted sources are used for classification and description.

For each application, we stressed some of the audio specificities to be taken into account when designing related BASS algorithms. We proposed some qualitative criteria to evaluate the performance of a given algorithm for the application, and the difficulty of the application itself depending on various factors such as the available amount of prior information, the noise level, the instantaneous or convolutive nature of the mixtures, etc.

This lead us to propose a tentative typology of the corresponding tasks to be solved by BASS algorithms, according to the input-output scheme of the algorithms. The main classification axis is the distinction between blind and non blind tasks. We retained the classical distinction between instantaneous and convolutive mixing as different levels of difficulty for a given task.

We should insist here that the proposed typology is intended to serve as a preliminary proposal, and we encourage researchers in the community to share their ideas about BASS tasks typology and evaluation or related topics using the discussion list on our dedicated web-site [32].

7. FUTURE WORK

This work constitutes an important step towards the definition of a global evaluation framework for BASS algorithms under different tasks. However in this article we have only described candidate qualitative criteria. Hence, the next steps should consist in transforming these criteria into numerical formulae and in building a "smart" benchmark of test signals according to the various BASS tasks we identified.

In a related paper [3], we expose quantitative criteria to measure the performance of source separation algorithms on the (Blind) Source Extraction and the (Blind) Remixing tasks. These criteria measure the performance in terms of interferences, noise and algorithmic artifacts, by properly taking into account the gain indeterminacies of source separation. They can be used in (over-)determined as well as under-determined problems. Other tasks require the design of other relevant numerical criteria.

A database structure and some labeled test signals are also readily available online [32].

Finally, it seems that some BASS applications such as audio scene modification have been less studied than the *one versus all* problem for instance, despite their generally less demanding requirements. We hope this work will trigger interest for new research goals.

8. ACKNOWLEDGMENTS

This work has been performed within a Junior Researchers Project "Resources for Audio Signal Separation" funded by GdR ISIS (CNRS). The goal of the project is to identify the specificities of audio signal separation, to suggest relevant numerical criteria to evaluate separation quality, and to gather test signals of calibrated difficulty level, in order to evaluate the performance of existing and future algorithms.

Some evaluation routines, a database of audio signals and a discussion list can be found on the web-site [32].

9. REFERENCES

[1] J.-F. Cardoso, "Blind source separation : statistical principles," in *IEEE Proc.*, 1998, vol. 90, pp. 2009–2026.

[2] M. Zibulevsky and B.A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionnary," *Neural Computation*, vol. 13, no. 4, 2001.

[3] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *Proc. Int. Workshop on ICA and BSS (ICA'03)*, 2003, submitted.

[4] K. Torkkola, "Blind separation for audio signals - are we there yet ?," in *Proc. Int. Worshop on ICA and BSS (ICA'99)*, 1999.

[5] D. Schobben, K. Torkkola, and P. Smaragdis, "Evaluation of blind signal separation methods," in *Proc. Int. Workshop on ICA and BSS (ICA'99)*, 1999, pp. 261–266.

[6] R.H. Lambert, "Difficulty measures and figures of merit for source separation," in *Proc. Int. Workshop on ICA and BSS (ICA'99)*, 1999, pp. 133–138.

[7] Olivier Cappé, *Techniques de réduction de bruit pour la restauration d'enregistrements musicaux*, Ph.D. thesis, Télécom Paris, 1993.

[8] J. Hershey and M.A. Casey, "Audiovisual sound separation via hidden Markov models," in *Proc. Advances in Neural Information Processing Systems (NIPS'02)*, 2002.

[9] C. Colomes, C. Schmidmer, T. Thiede, and W.C. Treurniet, "Perceptual quality assessment for digital audio (PEAQ) : the proposed ITU standard for objective measurement of perceived audio quality," in *Proc. AES Conf.*, 1999.

[10] H. Bousbiah-Salah, A. Belouchrani, and K. Abed-Meraim, "Jacobi-like algorithm for blind signal separation of convolutive mixtures," *Electronics Letters*, vol. 37, no. 1, pp. 1049–1050, 2001.

[11] D.C.B. Chan, P.J.W. Rayner, and S.J. Godsill, "Multi-channel blind signal separation by decorrelation," in *Proc. IEEE Worshop on Applications of Sig. Proc. to Audio and Acoustics*, 1995.

[12] K. Jensen, *Timbre models of musical sounds*, Ph.D. thesis, Datalogisk Institut, Copenhague University, 1999.

[13] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[14] Y. Meron and K. Hirose, "Separation of singing and piano sounds," in *Proc. Int. Conf. on Spoken Language Processing*, 1998.

[15] R. Dressler, "Dolby Surround Pro Logic II decoder : Principles of operation," Dolby Laboratories Information, 2000.

[16] R. Radke and S. Rickard, "Audio interpolation," in *Proc. AES 22nd Int. Conf. on Virtual, Synth. and Entertainment Audio*, 2002.

[17] P. Herrera-Boyer, "Setting up an audio database for Music Information Retrieval benchmarking," in *Proc. Int. Symp. on Music Information Retrieval (ISMIR'02)*, 2002, pp. 53–55.

[18] D.P.W. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, MIT, 1996.

[19] M. Goto and S. Hayamizu, "A real-time music description system : detecting melody and bass lines in audio signals," in *Proc. IJCAI'99 Workshop on Computational ASA*, 1999.

[20] H. Attias, J.C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Proc. Int. Workshop on Neural Information Processing Systems (NIPS'01)*, 2001.

[21] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sound sources," in *Proc. Int. Conf. on Speech and Language Proc. (ICSLP'00)*, 2000.

[22] L.C. Parra and C.V. Alvino, "Geometric source separation : merging convolutive source separation with geometric beamforming," *IEEE Trans. on Speech and Audio Proc.*, 2002, accepted.

[23] M. Casey, "Generalized sound classification and similarity in MPEG-7," *Organized Sound*, vol. 6, no. 2, 2002.

[24] The ELISA consortium, "The ELISA systems for the NIST'99 evaluation in speaker detection and tracking," *Digital Signal Processing*, vol. 10, no. 1-3, 2000, Special issue on the NIST'99 Speaker Recognition Workshop.

[25] K. Wilson and al., "Audio-video array source separation for perceptual user interfaces," Tech. Rep., MIT Artificial Intelligence Lab., 2001.

[26] F. Ehlers and H.G. Schuster, "Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment," *IEEE Trans. on Sig. Proc.*, vol. 45, no. 10, pp. 2608–2612, 1997.

[27] J. Reiss and M. Sandler, "Benchmarking Music Information Retrieval systems," in *Proc. Int. Symp. on Music Information Retrieval (ISMIR'02)*, 2002, pp. 37–42.

[28] K. Abed-Meraim, W. Qiu, and Y. Hua, "Blind system identification," in *Proc. IEEE*, 1997, vol. 85, pp. 1310–1322.

[29] F. Bimbot, *Traitement Automatique du Langage Parlé*, chapter Reconnaissance Automatique du Locuteur, Information–Commande–Communication (IC2). Hermès, 2003, to appear.

[30] J.O. Smith, "Physical modeling using digital waveguides," *Computer Music Journal*, vol. 16, no. 4, pp. 74–91, 1992, Special issue on physical modeling of musical instruments.

[31] L.K. Saul and M.I. Jordan, "Mixed memory Markov models: Decomposing complex stochastic processes as mixtures of simpler ones," *Machine learning*, vol. 37, no. 1, pp. 75–87, 1999.

[32] Action Jeunes Chercheurs du GDR ISIS (CNRS), "Ressources pour la séparation de signaux audiophoniques," http://www.ircam.fr/anasyn/ISIS/.