Advances in audio source seperation and multisource audio content retrieval

Emmanuel Vincent

INRIA, Centre de Rennes - Bretagne Atlantique
Campus de Beaulieu, F-35042 Rennes Cedex, France

ABSTRACT

Audio source separation aims to extract the signals of individual sound sources from a given recording. In this paper, we review three recent advances which improve the robustness of source separation in real-world challenging scenarios and enable its use for multisource content retrieval tasks, such as automatic speech recognition (ASR) or acoustic event detection (AED) in noisy environments. We present a Flexible Audio Source Separation Toolkit (FASST) and discuss its advantages compared to earlier approaches such as independent component analysis (ICA) and sparse component analysis (SCA). We explain how cues as diverse as harmonicity, spectral envelope, temporal fine structure or spatial location can be jointly exploited by this toolkit. We subsequently present the uncertainty decoding (UD) framework for the integration of audio source separation and audio content retrieval. We show how the uncertainty about the separated source signals can be accurately estimated and propagated to the features. Finally, we explain how this uncertainty can be efficiently exploited by a classifier, both at the training and the decoding stage. We illustrate the resulting performance improvements in terms of speech separation quality and speaker recognition accuracy.

Keywords: Speech, audio, source separation, robust automatic speech recognition

1. INTRODUCTION

In many environments, the audio modality consists of several sound sources. The target sound sources, which may be either specific speakers or sound events, are then masked by a variety of other sounds which make it difficult to understand speech or detect these events both for humans and computers. These issues can be broken down into two fundamental problems:

- *source separation*, that is estimating the audio signal of each source from the observed mixture,
- *classification* of the separated sources, *e.g.* automatic speech recognition (ASR) or acoustic event detection (AED).

In this paper, we review recent advances pertaining to these two problems in Sections 2 and 3, respectively. We conclude in Section by summarizing achievements and future perspectives 4.

2. SOURCE SEPARATION

2.1 State of the art

Audio source separation algorithms typically operate in the time-frequency domain by means of the short-time Fourier transform (STFT). The established *linear modeling* paradigm[1,2] assumes that the sources are static point sources and that the amount of reverberation is low. Under these assumptions, the vector $\mathbf{X}_{nf}$ of complex-valued STFT coefficients of the multichannel mixture signal in time frame n and frequency bin f is given by

$$\mathbf{X}_{nf} = \sum_{j=1}^{J} S_{jnf} \mathbf{A}_{jf} \tag{1}$$

Further author information: E-mail: emmanuel.vincent@inria.fr, Telephone: +33 (0)2 99 84 22 69

Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering X, edited by Harold Szu, Liyi Dai, Proc. of SPIE Vol. 8401, 840109 · © 2012 SPIE · CCC code: 0277-786X/12/$18
doi: 10.1117/12.926411

where S_{jnf} are the scalar STFT coefficients of the J underlying single-channel source signals indexed by j and $\mathbf{A}_{jf}$ are *steering vectors* representing the frequency response of the mixing filters. The source STFT coefficients are then typically assumed to be independent and identically distributed according to a *sparse* distribution and estimated in the maximum a posteriori (MAP) sense using convex or nonconvex optimization. Depending on the respective number of mixture channels and sources and on the chosen source distribution, this approach yields different algorithms called frequency-domain independent component analysis (FDICA) or sparse component analysis (SCA).

The series of Signal Separation Evaluation Campaigns (SiSEC)[3*] has shown that these algorithms achieve good results on mixtures of two sources with low reverberation but that their performance significantly degrades in the presence of many sources, background noise or medium to high reverberation. Indeed, they mostly rely on spatial cues, which are obscured in complex environments.

2.2 A general flexible probabilistic framework

In order to achieve more robust separation in such complex environments, a new *variance modeling* paradigm has emerged which enables the handling of diffuse or reverberated sources and the exploitation of spectral cues.[2] The mixture STFT coefficients are now modeled as

$$\mathbf{X}_{nf} = \sum_{j=1}^{J} \mathbf{C}_{jnf} \tag{2}$$

where $\mathbf{C}_{jnf}$ is the *spatial image* of source j, that is its contribution within the mixture. This multichannel quantity is subsequently modeled via the zero-mean Gaussian distribution

$$\mathbf{C}_{jnf} \sim \mathcal{N}(\mathbf{C}_{jnf}|\mathbf{0}, V^{\mathrm{ex}}_{jnf} V^{\mathrm{ft}}_{jnf} \mathbf{R}_{jf}) \tag{3}$$

where V^{ex}_{jnf} and V^{ft}_{jnf} represent respectively the spectral power of the excitation and that of the filter within a *source-filter model* of the source and $\mathbf{R}_{jf}$ is the *spatial covariance matrix* of the source.

In the case of point sources with low reverberation, the spatial covariance matrices are rank-1 matrices which can be expressed as $\mathbf{R}_{jf} = \mathbf{A}_{jf}\mathbf{A}^H_{jf}$ where $\mathbf{A}_{jf}$ are the steering vectors defined above. The main benefit of this new model comes when considering diffuse or reverberated sources: we have argued that $\mathbf{R}_{jf}$ then become full-rank matrices which do not only encode the spatial position of the sources but also their spatial width.[4]

We have also proposed to factor the spectral power of the excitation as[5]

$$V^{\mathrm{ex}}_{jnf} = \sum_{klm} W^{\mathrm{ex}}_{jfl} U^{\mathrm{ex}}_{jlk} G^{\mathrm{ex}}_{jkm} H^{\mathrm{ex}}_{jmn} \tag{4}$$

where W^{ex}_{jfl}, U^{ex}_{jlk}, G^{ex}_{jkm} and H^{ex}_{jmn} represent the *spectral fine structure*, the *spectral envelope*, the *temporal envelope* and the *temporal fine structure* of the source, respectively. A similar factorization may be assumed for the spectral power of the filter. Each of these quantities may be either fixed depending on the available prior knowledge about the sources at hand, or estimated from the data in an unsupervised fashion. For instance, in Figure 1, a speech source is modeled by assuming that its spectral fine structure is either harmonic or wideband and that its temporal fine structure exhibits a smooth temporal decay and by estimating its spectral envelope and its temporal envelope from the data. The estimated temporal envelope G^{ex}_{jkm} exhibits peaks for certain rows k encoding the pitch of the voiced part and the shape of the unvoiced part over time m. This flexible factorization makes it possible to exploit many more cues for separation than FDICA or SCA, thereby improving robustness to difficult mixtures where spatial cues do not suffice.

The model parameters can be estimated in the maximum likelihood (ML) sense via the expectation-maximization (EM) algorithm and the source STFT coefficients $\mathbf{C}_{jnf}$ are then recovered by multichannel Wiener filtering. The source signals are finally obtained by STFT inversion. The resulting algorithm has been implemented in the Flexible Audio Source Separation Toolkit (FASST)[†] for Matlab. Table 1 shows the average signal-to-distortion

[*] http://sisec.wiki.irisa.fr/

[†] http://bass-db.gforge.inria.fr/fasst/

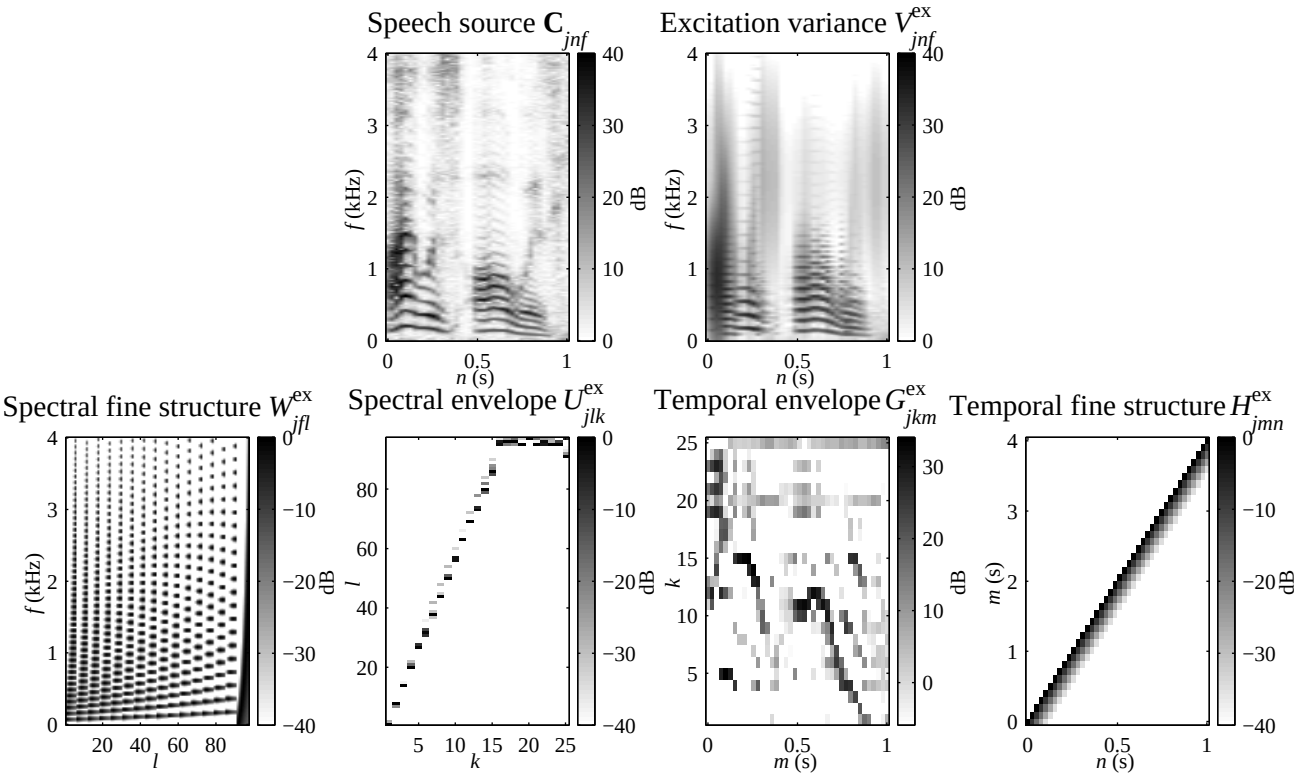

Figure 1. Example factorization of the spectral power of a speech source.

ratio (SDR) in decibels (dB) achieved by FASST on two-channel reverberant mixtures of three and four speech sources from the SiSEC 2010 evaluation campaign, depending on the rank of the spatial covariance matrices and on whether the spectral and the temporal fine structures are constrained or not. The best results are achieved when both the spectral and the temporal fine structures are constrained as in Figure 1. Also, full-rank spatial covariance matrices greatly improve performance compared to rank-1 spatial covariance matrices when the microphone spacing is large.

Table 1. Average SDR achieved by FASST on two-channel reverberant mixtures of three and four speech sources from the SiSEC 2010 evaluation campaign, depending on the rank of the spatial covariance matrices and on whether the spectral and temporal fine structures are constrained (X) as in Figure 1 or not. Two different microphone spacings are considered.

Spatial, spectral, and temporal constraints			Average SDR (dB)	
rank	spec	temp	5 cm	1 m
1			2.2	2.5
2			2.0	3.0
1	X		2.2	2.8
2	X		2.3	3.2
1		X	2.4	2.6
2		X	2.1	2.9
1	X	X	**2.5**	3.9
2	X	X	2.3	**5.0**

3. CLASSIFICATION OF SEPARATED SOURCES

3.1 State of the art

While the source signals estimated by FDICA, SCA or FASST can be directly input to an ASR or an AED system,[6] this approach is suboptimal as shown in particular by the PASCAL CHiME Speech Separation and Recognition Challenge[7‡]. Indeed, the audio features such as mel frequency cepstral coefficients (MFCC) used by classification systems are typically not robust to the distortions of the sources produced by source separation systems. In order to overcome this issue, additional information must be passed along with the sources to the classification system describing which parts of the source signals have been estimated with high confidence and which have not.

The *uncertainty decoding* (UD) framework[8] addresses this issue in the case of Gaussian mixture model (GMM) or hidden Markov model (HMM)-based classification systems by assuming that the feature vectors are modeled by a Gaussian distribution whose mean and covariance matrix represent the expected value of the features and the uncertainty about this value, respectively. Assuming that a clean model has been learned for each class, this uncertainty information is integrated with the GMM/HMM likelihood so as to form a *noise-robust likelihood*In the case of Gaussian mixture model (GMM) or hidden Markov model (HMM)-based classification used for ML classification. In practice, the uncertainty over the features is obtained by *uncertainty estimation* over the source STFT coefficients followed by *uncertainty propagation* to the features via moment matching or unscented transform techniques.[9]

Despite advances in this area, uncertainty estimation techniques remain mostly heuristic to date. Also, the original UD framework exploits uncertainty information only at the classification stage but not at the training stage, so that clean data are needed for training the classifier.

3.2 Robust feature extraction

Let us denote by θ the set of parameters of the considered source separation system, *e.g.* the variables $\mathbf{R}_{jf}$, W^{ex}_{jfl}, U^{ex}_{jlk}, G^{ex}_{jkm}, H^{ex}_{jmn}, W^{ft}_{jfl}, U^{ft}_{jlk}, G^{ft}_{jkm} and H^{ft}_{jmn} in the case of FASST. One established uncertainty estimation technique consists of estimating $\widehat{\theta} = \arg\max P(\mathbf{X}|\theta)$ in the ML sense as explained in Section 2 and approximating the uncertainty $P(\mathbf{C}|\mathbf{X})$ over the source STFT coefficients by that of the multichannel Wiener filter $P(\mathbf{C}|\mathbf{X}, \widehat{\theta})$ derived from $\widehat{\theta}$.

We have argued that uncertainty is better estimated in theory by marginalizing over the parameters θ:

$$P(\mathbf{C}|\mathbf{X}) \propto \int P(\mathbf{X}|\mathbf{C}, \theta) p(\mathbf{C}|\theta) d\theta. \tag{5}$$

Due to the large number of parameters, this integral is typically intractable so we proposed a variational Bayesian (VB) inference algorithm yielding a tractable approximation.[10]

Table 2 shows the average root mean square error (RMSE) between the true and the estimated source MFCCs for the two-channel reverberant speech mixtures and the various source models of Table 1, depending on the chosen uncertainty estimation technique. The proposed VB-based uncertainty estimation provides a small improvement compared to ML-based uncertainty estimation for all source models. The estimated covariance of the source MFCCs (not shown in the table) is also more accurately estimated.

3.3 Robust classifier training

The inability of UD to train classifiers from noisy data is even more crucial. Firstly, clean training data are not always available in the case of, e.g., field recording or mobile recording where the whole recording might be corrupted by noise. Secondly, even when sufficient clean data are available for training, the uncertainty over the test data is never perfectly estimated in practice such that some noise may remain that is not accounted for. Yet, training from noisy data is known to be an efficient technique to account for noise in the test data and improve the robustness of the classifier.

‡http://spandh.dcs.shef.ac.uk/projects/chime/challenge.html

Table 2. Average RMSE in dB between the true and the estimated source MFCCs for the two-channel reverberant mixtures and the various source models of Table 1, depending on the chosen uncertainty estimation technique.

Spatial, spectral, and temporal constraints			Average RMSE (dB)	
rank	spec	temp	ML-based	VB-based
1			6.66	6.63
2			6.85	6.84
1	X		6.72	6.69
2	X		6.82	6.80
1		X	6.59	**6.54**
2		X	6.76	6.76
1	X	X	6.61	6.57
2	X	X	7.23	6.92

We proposed an EM algorithm termed *uncertainty training* that estimates the parameters of a GMM-based classifier by maximizing the noise-robust likelihood of UD over the training data.[11] Each iteration of this algorithm consists of estimating the first and second order moments of the feature vectors by Wiener filtering given the uncertainty about the training data and subsequently updating the GMM parameters.

Table 3 shows the results for a speaker classification task over mixtures of speech and real-world domestic background noise taken from the PASCAL CHiME Speech Separation and Recognition Challenge dataset. For both training and test data, separation is performed via FASST, uncertainty estimation via the ML-based approach cited above and MFCC uncertainty propagation via the vector Taylor series (VTS) technique.[11] Speaker recognition accuracy is evaluated when training either from clean data, matched training data exhibiting the same signal-to-noise ratio (SNR) as the test data, unmatched data exhibiting a different SNR, or multi-condition data spanning all SNRs. Source separation combined with UD improves accuracy by 6 to 18% absolute compared to classification from the noisy signal. Uncertainty training further increases performance by 3 to 4% absolute when training from noisy data. The resulting performance is then much higher when training from multi-condition or even unmatched data than when training from clean data.

Table 3. Average speaker recognition accuracy (in %) for all training and decoding algorithms as a function of the training condition.

Source separation	Training strategy	Decoding strategy	Training condition			
			Clean	Matched	Unmatched	Multi
No	Conventional	Conventional	65.17	71.81	69.34	84.09
Yes	Conventional	Conventional	55.22	82.11	80.91	90.12
Yes	Conventional	Uncertainty	**83.48**	87.92	87.19	90.12
Yes	Uncertainty	Uncertainty	**83.48**	**91.79**	**90.61**	**94.04**

4. CONCLUSION

In the last five years, source separation has become a mainstream topic in audio signal processing. A few recent algorithms such as FASST have reached a sufficient level of maturity which enables their use in real-world reverberant, noisy application scenarios. The boom of mobile computing has also created a demand for robust ASR systems for handheld devices. Although commercial systems already exist, room remains for improving their robustness to challenging acoustic conditions. Recent advances in uncertainty decoding will most probably play a role in that context.

ACKNOWLEDGMENTS

This work was jointly performed with Kamil Adiloğlu, Frédéric Bimbot, Ngoc Q. K. Duong, Rémi Gribonval, Alexey Ozerov and Laurent S. Simon and supported by Oseo under the Quaero program and the EUREKA Eurostars i3DMusic project.

REFERENCES

[1] Makino, S., Lee, T.-W., and Sawada, H., [*Blind speech separation*], Springer (2007).

[2] Vincent, E., Jafari, M., Abdallah, S. A., Plumbley, M. D., and Davies, M. E., "Probabilistic modeling paradigms for audio source separation," in [*Machine Audition: Principles, Algorithms and Systems*], ch. 7, 162–185, IGI Global (2010).

[3] Vincent, E., Araki, S., Theis, F. J., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, B. V., Lutter, D., and Duong, N. Q. K., "The Signal Separation Evaluation Campaign (2007–2010): Achievements and remaining challenges," *Signal Processing* (to appear). doi: 10.1016/j.sigpro.2011.10.007.

[4] Duong, N. Q. K., Vincent, E., and Gribonval, R., "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech and Language Processing* **18**, 1830–1840 (Sep. 2010).

[5] Ozerov, A., Vincent, E., and Bimbot, F., "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech and Language Processing* **20**, 1118–1133 (May 2012).

[6] Ozerov, A. and Vincent, E., "Using the FASST source separation toolbox for noise robust speech recognition," in [*International Workshop on Machine Listening in Multisource Environments (CHiME 2011)*], 86–87 (September 2011).

[7] Barker, J., Vincent, E., Ma, N., Christensen, H., and Green, P., "The PASCAL CHiME Speech Separation and Recognition Challenge," *Computer Speech and Language* (to appear).

[8] Deng, L., Droppo, J., and Acero, A., "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Transactions on Speech and Audio Processing* **13**, 412– 421 (2005).

[9] Astudillo, R. F. and Orglmeister, R., "A MMSE estimator in mel-cepstral domain for robust large vocabulary automatic speech recognition using uncertainty propagation," in [*Proc. Interspeech*], 713–716 (2010).

[10] Adiloğlu, K. and Vincent, E., "A general variational Bayesian framework for robust feature extraction in multisource recordings," in [*Proc. 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*], (to appear).

[11] Ozerov, A., Lagrange, M., and Vincent, E., "GMM-based classification from noisy features," in [*Proc. 1st International Workshop on Machine Listening in Multisource Environments (CHiME)*], 30–35 (2011).