On-line multichannel estimation of source spectral dominance

Francesco Nesta[1], Trausti Thormundsson[1], Zbyněk Koldovský[2]

1) Conexant System, 1901 Main Street, Irvine, CA (USA)
2) Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic
francesco.nesta,trausti.thormundsson@conexant.com,zbynek.koldovsky@tul.cz

Abstract. Despite its popularity, multichannel source demixing is intrinsically limited in real-world applications due to the model mismatch between the convolutive mixing model and the actual recordings. Varying number of sources, reverberation, diffuseness and spatial changes are common uncertainties that need to be handled. Post-processing is commonly adopted to compensate for these mismatches, generally in the form of non-linear spectral filtering. In this work we analyze the property of the normalized differences between the output magnitudes of a linear spatial filter. We show that thanks to the time-frequency sparsity of acoustic signals, such distributions can be approximatively modeled by a bimodal Gaussian mixture model. An on-line bimodal constrained GMM fitting is proposed, in order to estimate the posterior probability of source spectral dominance. It is shown that the estimated posteriors can be used to produce a filtered output with very low distortion, outperforming traditional non-linear methods.

Keywords: source separation, GMM, binary masking, speech enhancement

1 Introduction

Multichannel spatial filtering has shown to be effective with the enhancement of a given sound source of interest from the remaining noise. Supervised methods exploit prior geometrical information in the form of source position, leading to classical beamformer [2] or to more sophisticated cancellation filter bank (CFB)-based methods [3]. On the other hand, unsupervised methods are traditionally based on separation frameworks using Independent Component Analisys (ICA) [9] or spatial/spectral clustering [1]. Differently from traditional single channel enhancement methods, multichannel filtering exploits spatial cues to discriminate between multiple sources and do not necessarily need any strong assumption on the nature of the sound signal. As a main advantage such methods are able to deal with the separation of highly non-stationary signals such as concurrent speech sounds.

[1] The work of Zbyněk Koldovský was partially supported by California Community Foundation through Project No. DA-15-114599.

Despite their popularity, spatial filtering methods have intrinsic limitations due to the approximated mixing system modeling. Mixtures are generally approximated as a linear combination of signals generated by a finite number of spatially localized sources, often referred to as coherent sources. However, this condition is only partially fulfilled in real-world since the noise spatial covariance can be highly time-varying. Furthermore, in presence of high reverberation, linear deximing with short filters is suboptimal and leads to a large cross-source output signal leakage.

For the above limitations, spatial demixing is rarely used alone for source separation and is usually complemented by post-filtering methods exploiting other spectral cues. For instance, in classical beamforming a GSC structure is employed to remove the residual noise in the target channel [2]. In source separation systems with a limited number of microphones, spectral masking is generally adopted in the form of binary masks [6] or Wiener-like gains [3]. However, these methods do not explicitly model the uncertainty of the spatial filter and as a result, they require heuristic tuning hyperparameter optimized to avoid distortion in the target signal.

In this work we discuss on the meaning of the normalized cross-output-channel magnitude differences, i.e. the normalized differences of magnitudes measured at the outputs of the spatial filter, and show how its pdf can be used to predict the posterior probability of source dominance, compared to the Ideal Binary Mask (IdBM)[11][4]. Then, we propose an on-line fitting of a constrained GMM model whose posteriors are used to generate spectral gains for the filtering.

2 Models for multichannel observations

In this work we limit the analysis to the case of recordings made by 2 microphones but the discussion can be easily extended to a generic multichannel case. We indicate with $s(t)$ the time-domain signal generated by a target speech and with $x_1(t)$ and $x_2(t)$ the signal sampled at the first and second microphones which can be modeled as $x_i(t) = s_i(t) + n_i(t)$, where $s_i(t)$ and $n_i(t)$ $\forall i = 1, 2$ indicates the reverberant image of the target source and the noise contributions to each microphone (which can be viewed as generated by a multiplicity of coherent noise sources). We assume that a generic spatial filtering system is trained to produce an estimate of $s_i(t)$ and of $n_i(t)$. If $s(t)$ is a coherent source, regardless of the spatial characteristic of $n(t)$, the output of the system can be approximatively modeled as

$$\hat{s}_i(t) = s_i(t) + \alpha[n_i(t)], \quad \hat{n}_i(t) = \gamma[n_i(t)] + \beta[s_i(t)] \tag{1}$$

where $\alpha[\cdot]$ and $\beta[\cdot]$ are time-varying convolutive transformations modeling the residual of noise and speech in the corresponding channels, and $\gamma[\cdot]$ models the distortion of the estimated noise due to the approximated linear demixing. This model comes from the application of the Minimal Distortion Principle (MDP)[5] to a generic inverse multichannel filter where the noise sources might exceed the number of microphones (see [7] section 5.2 for details).

By means of a time-frequency analysis, e.g. a weighted short-time Fourier transform (STFT), each signal can be transformed from time-domain to a discrete time-frequency representation. Therefore, let $S_i(k,l)$ and $N_i(k,l)$ (with $i = 1, 2$) be the downsampled subband representation of the time-domain signals where k and l indicates the frequency bin and subband frame, respectively. Assuming that the convolutive transformations are approximatively stationary within the STFT analysis window and that the target speech and noise are uncorrelated, we can model the output magnitude of the spatial filter as

$$\begin{aligned} |\widehat{S}_i(k,l)| &\simeq |S_i(k,l)| + \alpha(k,l)|N_i(k,l)| \\ |\widehat{N}_i(k,l)| &\simeq \gamma(k,l)|N_i(k,l)| + \beta(k,l)|S_i(k,l)| \end{aligned} \tag{2}$$

where $\alpha(k,l)$, $\beta(k,l)$ and $\gamma(k,l)$ are positive constants. In oracle conditions, if the magnitudes of the target source and noise were available the Ideal Binary Mask extracting the target source could be estimated as

$$IdBM(k,l) = 1, \text{if } \ |S_i(k,l)| > LC \cdot |N_i(k,l)|, \qquad IdBM(k,l) = 0, \text{otherwise} \tag{3}$$

where LC is the local signal-to-noise ratio (SNR) in linear scale, typically set to 1 (i.e. 0 dB) [11][4]. In our case we only observe $\widehat{S}_i(k,l)$ and $\widehat{N}_i(k,l)$ from which we cannot directly infer the IdBM. However, by modeling the statistical distribution of features derived from $|\widehat{S}_i(k,l)|$ and $|\widehat{N}_i(k,l)|$ it is possible to estimate the probability that the IdBM is 1 in a particular time-frequency point.

3 Normalized cross-output-channel magnitude differences as discriminative features for T-F source dominance

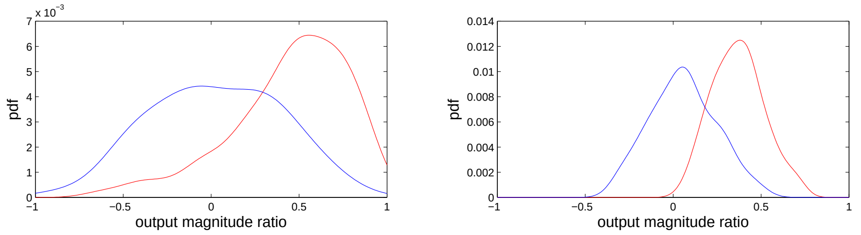
First, we define the normalized cross-output-channel magnitude differences as

$$f_k(l) = \frac{|\widehat{S}(k,l)| - |\widehat{N}(k,l)|}{|\widehat{S}(k,l)| + |\widehat{N}(k,l)|} \tag{4}$$

where we remove the dependence on the channel i to simplify the notation. By substituting (2) in (4) we get

$$f_k(l) \simeq \frac{[1-\beta(k,l)]\sqrt{SNR(k,l)} + [\alpha(k,l) - \gamma(k,l)]}{[1+\beta(k,l)]\sqrt{SNR(k,l)} + [\gamma(k,l) + \alpha(k,l)]} \tag{5}$$

where $SNR(k,l)$ is the true instantaneous signal-to-noise ratio (in linear scale) in the (k,l) T-F point. Due to the sparseness of acoustic signals in the T-F domain, the SNR will assume values close to 0 or close to infinity according to which source is dominant/active for a particular time-frequency point. As a consequence, the gains are expected to cluster around the centers $\frac{\alpha(k,l)-\gamma(k,l)}{\gamma(k,l)+\alpha(k,l)}$ and $\frac{1-\beta(k,l)}{1+\beta(k,l)}$ for T-F points dominated by the noise and target source, respectively. Therefore, we can expect the density of $f_k(l)$, $\forall k$, being approximatively



(a)

(b)

Fig. 1. Kernel density of $f(k,l)$ for a specific subband-based k a) and for the entire frame b).

bimodal with each component displaced and spread according to the statistic of the variables $\alpha(k,l), \gamma(k,l), \beta(k,l)$.

To help the understanding of this discussion, we consider the case where $\widehat{S}(k,l)$ and $\widehat{N}(k,l)$ are estimated through a two-channel spatial filter based on a geometrically constrained Independent Component Analysis [8]. By using the true oracle images $S(k,l)$ and $N(k,l)$, we define the IdBM and cluster T-F points in two classes, target source and noise source dominated points. Then, the kernel density estimate of $f_k(l)$ is computed for each class separately. In figure 1(a) the resulting densities for the two classes are shown with different colors. It can be observed that the full distribution resembles a mixture of exponential components which can be approximatively modeled with two Gaussian components. Note, this is only a convenient approximation since the observations in $f_k(l)$ are bounded in the range $[-1;1]$ and in general each component cannot be symmetric. However, in practice, a simple Gaussian model is enough accurate to describe the uncertainty of $f_k(l)$ in representing the dominance classes.

Figure 1(b) shows the density obtained with the $f_k(l)$ of all the subbands. On average a bimodal GMM fits well the empirical distribution, which implies that the model can be also used for a frame-based multichannel target source activity detection. Note, function in eq. (4) produces a convenient 1-dimensional discriminative representation correlated to the IdBM class probabilities. However, an alternative function could be used as long as a suitable model is available for describing its pdf.

4 Constrained on-line GMM parameter fitting

According to the above analysis, in each T-F point, we model the density of $f_k(l)$ with a bimodal GMM as

$$p[f_k(l)] = w_{1,k}(l) \cdot N[\mu_{1,k}(l), \sigma^2_{1,k}(l)] + w_{2,k}(l) \cdot N[\mu_{2,k}(l), \sigma^2_{2,k}(l)] \qquad (6)$$

where the $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ represent the distribution of the spectral gains for points dominated by the target source and noise respectively. Following the interpretation of $f_k(l)$, $\mu_{1,k}(l)$ is expected to be larger than $\mu_{2,k}(l)$ and therefore proper constraints need to be added to the model. In this work we fit the model variables using a sequential approximation of the expectation-maximization approach as in [12]. We define with $c \in \{1, 2\}$ the class labels, where 1="target speech dominant", 2="noise dominant". We are interested in the probability $p[c = 1|f_k(l), \lambda_k(l)]$, where $\lambda_k(l) = [\mu_{1,k}(l), \sigma^2_{1,k}(l), w_{1,k}(l), \mu_{2,k}(l), \sigma^2_{2,k}(l), w_{2,k}(l)]$ is the parameter vector for the target speech and noise component models, estimated at the frame l. The probability of T-F target speech dominance $p_k(l)$ can be computed using the Bayes formula as

$$p_k(l) = p[c = 1|f_k(l), \lambda_k(l)] = \frac{w_{1,k}(l)p[f_k(l)|c = 1, \lambda_k(l)]}{\sum_{c=1}^{2} w_{c,k}(l)p[f_k(l)|c, \lambda_k(l))} \tag{7}$$

In the sequential on-line learning, within the frame l, the mixture parameters are updated $\forall c = 1, 2$ and $\forall k$ as

$$w_{c,k}(l) = (1 - \eta_c) \cdot w_{c,k}(l - 1) + \eta_c \cdot p[c|f_k(l), \lambda_k(l - 1)] \tag{8}$$

$$\mu_{c,k}(l) = \frac{(1 - \eta_c) \cdot \mu_{c,k}(l - 1)}{w_{c,k}(l)} + \frac{\eta_c \cdot p[c|f_k(l), \lambda_k(l - 1)]f_k(l)}{w_{c,k}(l)} \tag{9}$$

$$\sigma_{c,k}(l) = \frac{(1 - \eta_c) \cdot \sigma_{c,k}(l - 1)}{w_{c,k}(l)} + \frac{\eta_c \cdot p[c|f_k(l), \lambda_k(l - 1)](f_k(l) - \mu_{c,k}(l))^2}{w_{c,k}(l)} \tag{10}$$

where η_c is a learning rate step-size. Iterating equations (7)-(10) the GMM parameters are updated on-line with the incoming data. To avoid divergence in trivial solutions some constraints are necessary. First, the weights $w_{1,k}(l)$ or $w_{2,k}(l)$ can approach zero if either the target or the noise signal is absent for a long time. To avoid this divergence the values of the weights are constrained within the iteration as

$$w_{1,k}(l) = \min[\max(w_{1,k}(l), \epsilon), 1 - \epsilon], \quad w_{2,k}(l) = 1 - w_{1,k}(l) \tag{11}$$

where ϵ is set to a small value (e.g. 0.05). To guarantee that the estimated components are in the correct order, for each frequency bin k we impose:

$$\mu_1(k, l) > \mu_2(k, l), \tag{12}$$

and to avoid $\sigma^2_{1,k}(l)$ and $\sigma^2_{2,k}(l)$ approach 0, the following constraint is applied:

$$\sigma^2_{c,k}(l) = \min(\sigma^2_{c,k}(l), \epsilon_{\sigma^2}), \forall c \tag{13}$$

where ϵ_{σ^2} is a small value (e.g. 0.0001).

5 Proposed system architecture

Figure 2(b) shows the block architecture of the proposed filtering scheme. The multichannel recordings are sent to the input of a convolutive spatial filter which

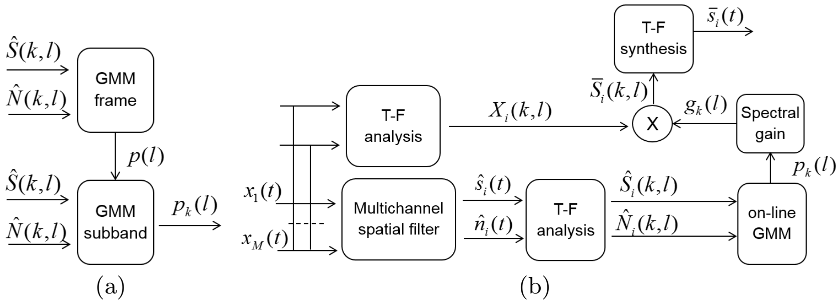


Fig. 2. Block diagram of the proposed filtering structure.

decomposes the input in target speech and noise signal components. In this work we evaluate the performance with both a blind spatial filter learned with a geometrically constrained ICA [8] and with a semi-blind filter based on a pre-trained cancellation filterbank [3]. The time-domain outputs are transformed in T-F representation by means of a STFT analysis with Hanning windows of 2048 points with increments of 128 samples.

To improve the robustness of the on-line fitting, a two level hierarchy structure is proposed as shown in figure 2(a). In the top level, a GMM model is used to fit all the subband features $f_k(l),\ \forall k$ in a single model and estimate the probability of speech dominance $p(l)$ in the entire frame. This block can be considered as an adaptive multichannel Voice Activity Detector (VAD). The frame-based posterior dominance probability $p(l)$ is then used to weight the learning rate of the subband-based GMM fitting, as follow:

$$\eta_1 = \eta \cdot p(l), \quad \eta_2 = \eta \cdot (1 - p(l)) \tag{14}$$

where η is the maximum step-size used for the subband parameter tracking. This approach prevents the GMM to update the target model parameters during speech pauses, which improves the convergence speed of the on-line learning. The subband posterior probabilities $p_k(l)$ are then used to compute spectral gains used to filter the multichannel input. Two alternative filtering approaches are proposed:

- M1: use the posteriors to directly compute the gains as $g_k(l) = p_k(l)$, if $p_k(l) > 0.5$ (0 otherwise). This approach is justified by the interpretation of the posteriors learned by the GMM model, which are related to the probability of IdBM(k,l) being equal to 1.
- M2: use the posteriors indirectly to estimate the expectation of the noise power $P(k,l)$ as

$$\begin{gathered} P(k,l) = (1 - a_k(l)) \cdot P(k,l) + a_k(l) \cdot |\hat{N}(k,l)|^2] \\ a_k(l) = 1 - p_k(l), \text{ if } p_k(l) < 0.5, \text{ (0 otherwise)} \end{gathered} \tag{15}$$

  and then use the noise power to estimate the spectral gains with conventional single-channel methods (in this work we used a standard spectral subtraction [10])

6 Experimental evaluation

Sources are recorded at $f_s = 16$kHz, in a room of size 5x5x2.5 meters with T_{60}=300ms with two microphones spaced of $0.2m$ and at a distance of 2 meters from the center of the array. Two datasets of 100 mixtures are generated:

- the first dataset is obtained by combining a target and noise interfering speakers (randomly chosen from a collection of male and female speakers). The target speaker is assumed to be in an angular region of $+/-10^o$ around the center of the array, while the noise is randomly displaced in any direction out of the target region. The average SNR at the input is of about -2dB.
- the second dataset is obtained from the first dataset but reducing the dynamic of the interfering speaker by 6 dB and adding a stereo real-world cafeteria noise to the mixture. The average SNR at the input is of about -0.5dB.

Tables 1 and 2 show the SNR and SDR improvement obtained for the two datasets with the blind and the semi-bind spatial filter. Performance are compared to conventional spectral masking methods such as binary masking $BM(k,l) = |\hat{S}(k,l)| > c{\cdot}|\hat{N}(k,l)|$ and parametric wiener-like filter $GW(k,l) = \frac{|\hat{S}(k,l)|}{|\hat{S}(k,l)|+c\cdot|\hat{N}(k,l)|}$ where the hyper parameter c was tuned to optimize the SDR scores. It can be noted that method $M1$ achieves the best overall scores which highlight that the GMM model fits well the IdBM class probabilities as long as the noise can be considered enough sparse in the time-frequency domain. On the other hand, in real-world diffuse noise, the indirect use of the posteriors in $M2$ delivers the best overall results since the noise has a lower degree of sparsity.

7 Conclusions

In this paper we discuss on the property of the normalized cross-output-channel magnitude difference of a spatial filter. It is shown that under certain conditions, its statistic can be approximatively modeled with a bimodal GMM whose posteriors can predict the probability of target source dominance. An on-line constrained GMM learning structure together with a filtering scheme is then proposed. Through an experimental evaluation with both coherent and real-world reverberant challenging recordings, it is shown that the proposed method can generate masks able to enhance the target source with a limited amount of distortion.

Future works may concern the use of better density models to generate more accurate posteriors, in combination with more advanced spectral filtering structures.

References

1. Araki, S., Nakatani, T., Sawada, H., Makino, S.: Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem. In: ICA '09: Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation. pp. 742–750. Springer-Verlag, Berlin, Heidelberg (2009)

Dataset1	BM	GW	M1	M2
SNRi	5.46 (3.21)	6.46 (3.11)	8.42 (2.90))	**11.27** (4.51)
SDRi	5.04 (1.55)	5.42 (1.45)	**5.62** (1.60)	5.47 (1.63)
Dataset2	BM	GW	M1	M2
SNRi	2.20 (1.01)	3.09 (1.29)	5.40 (1.99)	**9.49** (2.62)
SDRi	3.59 (0.96)	4.03 (0.88)	**4.09** (1.26)	3.94 (1.38)

Table 1. Mean and (standard deviation) SNRi and SDRi performance when using a blind ICA-based spatial filter.

Dataset1	BM	GW	M1	M2
SNRi	6.97 (3.90)	6.26 (2.28)	10.23 (3.01)	**12.08** (4.04)
SDRi	5.14 (1.68)	5.46 (1.35)	**6.21** (2.17)	5.64 (2.18)
Dataset2	BM	GW	M1	M2
SNRi	1.35 (0.50)	2.68 (1.18)	8.83 (1.32)	**11.78** (1.72)
SDRi	3.60 (0.82)	3.95 (0.70)	**5.64** (1.44)	5.37 (1.46)

Table 2. Mean and (standard deviation) SNRi and SDRi performance when using a semi-blind CFB-based spatial filter.

2. Brandstein, M., Ward, D.: Microphone Arrays. Springer Verlag (2001)
3. Koldovský, Z., Málek, J., Tichavský, P., Nesta, F.: Semi-blind noise extraction using partially known position of the target source. IEEE Transactions on Audio, Speech & Language Processing 21(10), 2029–2041 (2013)
4. Loizou, P., Kim, G.: Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. Audio, Speech, and Language Processing, IEEE Transactions on 19(1), 47–56 (Jan 2011)
5. Matsuoka, K., Nakashima, S.: Minimal distortion principle for blind source separation. In: Proceedings of International Symposium on ICA and Blind Signal Separation. San Diego, CA, USA (Dec 2001)
6. Mori, Y., Saruwatari, H., Takatani, T., Ukai, S., Shikano, K., Hiekata, T., Morita, T.: Real-time implementation of two-stage blind source separation combining simo-ica and binary masking. In: IWAENC. pp. 229–232 (2005)
7. Nesta, F., Matassoni, M.: Blind source extraction for robust speech recognition in multisource noisy environments. Comput. Speech Lang. 27(3), 703–725 (May 2013)
8. Nesta, F., Matassoni, M., Astudillo, R.F.: A flexible spatial blind source extraction framework for robust speech recognition in noisy environments. Proc. CHiME pp. 33–40 (2013)
9. Pedersen, M.S., Larsen, J., Kjems, U., Parra, L.C.: A survey of convolutive blind source separation methods. In: Springer Handbook of Speech (Nov 2007)
10. Tashev, I., Lovitt, A., Acero, A.: Unified framework for single channel speech enhancement. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (Aug 2009)
11. Wang, D.: Time-frequency masking for speech separation and its potential for hearing aid design. Trends in Amplification 12(4), 332–53 (2008)
12. Ying, D., Yan, Y., Dang, J., Soong, F.: Voice activity detection based on an unsupervised learning framework. Audio, Speech, and Language Processing, IEEE Transactions on 19(8), 2624–2633 (Nov 2011)