A NEW MODEL-BASED UNDERDETERMINED SPEECH SEPARATION

Zaher El Chami[1]*, Antoine Dinh-Tuan Pham*[2]*, Christine Servière*[3]*, Alexandre Guerin*[1]

[1]Orange Labs - TECH/SSTP, 2 Avenue Pierre Marzin, 22307 Lannion Cedex, France
[2]Laboratory of Modeling and Computation, B.P 53X, 38041 Grenoble cedex 9, France
[3]GIPSA-lab, Department of Images and Signals, BP 46, 38402 St Martin d'Hère Cedex, France
E-mail: {zaher.elchami;alexandre.guerin}@orange-ftgroup.com
dinh-tuan.pham@imag.fr; christine.serviere@inpg.fr

ABSTRACT

In this paper we present a new model-based blind speech separation for underdetermined case. Under sparsity assumption, separation is achieved by applying soft time frequency masks to observations. The masks are derived by estimating the parameters of an ad-hoc distribution of the Interchannel Level/Phase Difference (ILD/IPD). These parameters are estimated using an expectation-maximization (EM) procedure. The performance of the algorithm is evaluated on real-world convolved mixtures using the database of the first audio source evaluation campaign [1]. Results show that the proposed algorithm outperforms the algorithm presented in the campaign in terms of artifacts and distortion noise.

Index Terms— Sparse Source Separation, Time Frequency Masking, ILD/IPD modeling, EM algorithm.

1. INTRODUCTION

Recently, Underdetermined Blind Speech Separation (UBSS) has received much attention and especially the stereo case where researchers tend to limit their observations to two microphones, mimicking the natural ear separation. Most of the proposed UBSS methods [2, 3, 4] mainly rely on the assumption of sparse speech sources being disjoint in the Time-Frequency (TF) domain [2]. This results in the presence of a dominant source at each time-frequency (TF) point. Thus, sources can be separated by applying binary TF masks where features, like ILD/IPD, are used to estimate the index of the dominant source.

In case of anechoic mixing, the IPD, normalized by frequency, is constant and may be used as a signature to indentify each sources over the whole TF domain (see for instance [2, 3]). In real-world convolved mixtures we cope with, the anechoic assumption is no longer valid, since the normalized IPD varies with frequency. This implies that feature estimation and source separation should be done in each frequency band independently. In [4] the authors propose to separate sources using a hierarchical clustering in each frequency band. Another approach, also by frequency band, consists in modelling the interchannel cues using an *a priori* distribution. In [5], the authors proposed to model the ILD/IPD variables as Gaussian distribution with the assumption of a dominant path. Nevertheless, in real-world situations, the underlying linear phase assumption deriving from the dominant path constraint reveals not applicable due to early reflections and reverberation.

In this paper, we propose to use a two dimension probabilistic model as in [5], without any constraint on the phase. A parametric joint distribution is proposed for the couple of features ($\log$(ILD), IPD) considered as random variables. The parameter estimation and the mask computation are derived using an EM algorithm, yielding a soft TF masks that represents the probabilities of each source at each TF point. Note that, as we focus on the separation performance, the traditional frequency permutation problem [6] is supposed to be solved.

2. PROBLEM FORMULATION

Consider the convolved stereo mixture model:

$$x_j(t) = \sum_{i=1}^{N} \sum_{k} a_{ji}(k)\, s_i(t-k) = \sum_{i=1}^{N} (a_{ji} * s_i)(t) \quad (1)$$

where $s_i(t), i = 1,..,N$ are the sources, $*$ denotes the convolution symbol and $a_{ji}(k)$ are the impulse response of the acoustic path from source i to microphone j with $j = 1, 2$. The time-domain observed signals $x_j(t)$ are converted into frequency-domain time-series $x_j(t,\omega)$ signals using the Short-Time Fourier Transform (STFT):

$$X_j(t,\omega) = \sum_{k=0}^{L-1} w(k) x_j(t+k) e^{j\omega k} = \sum_{i=1}^{N} X_j^{(i)}(t,\omega)$$

where $w(k)$ is a window (e.g. Hamming) and $X_j^{(i)}(t,\omega)$ is the STFT of the contribution of the i^{th} source to the j^{th} sensor, that is of $x_j^{(i)}(t,\omega) = (a_{ij} * s_i)(t)$. Working in the TF domain has two advantages: sparseness of speech signals becomes prominent and the convolutive mixtures in (1) can be approximated as instantaneous mixtures at each frequency:

$$X_j(t,\omega) = \sum_{i=1}^{N} X_j^{(i)}(t,\omega) \approx \sum_{i=1}^{N} A_{ji}(\omega) S_i(t,\omega) \quad (2)$$

where $A_{ji}(\omega)$ is the frequency response of the filter $\{a_{ji}(k)\}$ and $S_i(t,\omega)$ is the STFT of $s_i(t)$. Under the disjoint assumption of the source TF supports, only one source is dominant at each TF bin (t,ω), so that the sum in (2) contains one single non negligible term. Thus (2) can be approximated as:

$$X_j(t,\omega) \approx X_j^{(q)}(t,\omega) \approx A_{jq}(\omega) S_q(t,\omega)$$

where $S_q(t,\omega)$ is the dominant source at the TF point (t,ω). The ratio would then be:

$$R(t,\omega) = \frac{X_1(t,\omega)}{X_2(t,\omega)} \approx \frac{X_1^{(q)}(t,\omega)}{X_2^{(q)}(t,\omega)} \approx \frac{A_{1q}(\omega)}{A_{2q}(\omega)}. \quad (3)$$

Being frequency and source dependent, the ratio $R(t,\omega)$ is used in each frequency band to identify the sources. Note that the modulus and argument of this ratio are the traditional ILD and IPD used in previous method [2, 3, 4, 5].

3. MODELLING OBSERVED RATIO

The second approximation made in (3) is only valid when the length L of the analysis window is much larger than that of the filter response $a_{ji}(k)$. However, in real reverberant room and due to speech stationarity constraint, L is usually smaller than the reverberation time; therefore such approximation is no longer valid. In fact, a detailed analysis not shown here for lack of space, reveals that for a given source i, the logarithm of the one-source observed ratio $\log[X_1^{(i)}(t,\omega)/X_2^{(i)}(t,\omega)]$ is not constant in each frequency band. Instead, at a given frequency ω, the time set of this log ratio can be shown theoretically to be a random variable with real and imaginary parts (x,y) that admits a joint density of the form:

$$p_{\rho_i(\omega)}\left[x - \log|r_i(\omega)|, y - \arg r_i(\omega)\right] \quad (4)$$

where $r_i(\omega)$ and $\rho_i(\omega)$ are specific to the i^{th} source. They respectively stand for its position in space and for the reverberation degree of the acoustic path between source i and the set of microphones. Further, the distribution probability function p_ρ is given by:

$$p_\rho(x,y) = \frac{1}{4\pi} \frac{1-\rho^2}{\left[\cosh(x) - \rho\cos(y)\right]^2} \quad (5)$$

So for a given frequency band ω and for the set of considered time points $t \in T$, we are led to assume the following model for the distribution of the real and imaginary parts of $\log[R(t,\omega)]$:

$$p(x,y|\boldsymbol{\rho},\boldsymbol{r},\boldsymbol{\mu}) = \sum_{i=1}^{N} \mu_i p_{\rho_i}\left\{x - \log|r_i|, y - \arg r_i\right\} \quad (6)$$

where $\boldsymbol{\rho} = [\rho_1,..,\rho_N]$, $\boldsymbol{r} = [r_1,..,r_N]$ and $\boldsymbol{\mu} = [\mu_1,..,\mu_N]$, μ_i denoting the *a priori* probability of the i^{th} source in the considered frequency band. Note that the parameters $\boldsymbol{\rho}$, $\boldsymbol{r}$ and $\boldsymbol{\mu}$ depend on frequency ω but for simplicity and readability, we will omit this variable in the rest of paper. Since from now, we implicitly work at a given frequency band. The parameters in (6) can be estimated, for example, by the maximum likelihood method based on the data $(\log|R(t)|, \arg R(t))$, $t \in T$. Once estimated, the *a posteriori* probability that i^{th} source is dominant at the TF point (t,ω) can be obtained by:

$$\pi_i(t) = \frac{\mu_i p_{\rho_i}\left[\log|R(t)/r_i|, \arg R(t)/r_i\right]}{\sum_{q=1}^{N} \mu_q p_{\rho_q}\left[\log|R(t)/r_q|, \arg R(t)/r_q\right]} \quad (7)$$

The source separation can thus be performed, independently in each frequency band, by directly applying these *a posteriori* probabilities to the observations.

Note that, in many practical situations, the *a posteriori* probabilities $\pi_i(t)$ given in (7) are either nearly one or zero If we suppose that the points $r_1,..,r_N$ are well distant so that there is only one term in the denominator of which dominates all other terms. Thus, if location parameters r_i are correctly estimated, the exact form of the density (4) is not very important since it affects little the *a posteriori* probabilities. Moreover, as this family of densities will lead to a likelihood function not easy to maximize, we propose to replace the density (5) by the following one:

$$p_{\alpha,\beta}(x,y) = \frac{e^{-|x|/\alpha}}{2\alpha} \frac{e^{(\cos y - 1)(1-\beta)/\beta}}{C(\beta)}, \alpha > 0,\ \beta \in [0,1] \quad (8)$$

where $C(\beta) = \int_{-\pi}^{\pi} e^{(\cos y-1)(1-\beta)/\beta} dy$ is the normalising constant. Thus in (6) we replace p_ρ by $p_{\alpha,\beta}$ and in (7) p_{ρ_i} by p_{α_i,β_i}. The model density is changed but still belongs to a location family admitting $\log|R(t)|$,$\arg R(t)$ as mean, mode and median (since the function p_ρ and $p_{\alpha,\beta}$ are both symmetric in both its arguments). This is important as we have seen that location parameters r_i should be correctly estimated. The density $p_{\alpha,\beta}$ has exponential tail (in it first argument) like p_ρ. The distribution (8) implicitely assumes that the variables are independent. The form of distribution given by (5) does not guarantee such hypothesis: nevertheless, a statistical analysis of the couple (x,y) shows that they may be assumed as decorrelated. Note that we have used a 2-parameters family (8) to approximate a 1-parameter family (5). This is convenient for computational purpose but also provides flexibility for a better approximation.

4. THE EXPECTATION-MAXIMIZATION

To estimate the model parameters μ_i, r_i, α_i, and β_i for each source at each frequency band ω, we propose to use as criteria the maximum of the log-likelihood of the data $\log|R(t)|, \arg R(t), t \in T$. A standard way to maximize the likelihood in presence of hidden variable is to use an EM algorithm [7]. In our case, the hidden variable is the index that indicates which source is dominant at time t (again, the frequency ω is fixed). Under the independence assumption between the time set of observations, the EM algorithm operates as follow:

4.1. The E-step

This step computes the expectation of the full log likelihood given the data. The full likelihood will be computed at generic parameters μ_i', r_i', α_i', and β_i' and the conditional expectation is computed relative to the model specified by the current parameter μ_i, r_i, α_i, and β_i. The result can be shown to be:

$$\sum_{t\in T}\sum_{i=1}^{N}\pi_i\left(t\right)\log\left\{\mu_i' p_{\alpha_i',\beta_i'}\left[\log\left|\frac{R\left(t\right)}{r_i'}\right|,\arg\frac{R\left(t\right)}{r_i'}\right]\right\} \quad (9)$$

where $\pi_i(t)$ is the *a posteriori* probability given in (7).

4.2. The M-step

This step maximises the above conditional expectation of the full log likelihood with respect to the generic parameters μ_i', r_i', α_i', and β_i'. The maximum point is then taken as the new parameter. It is easily seen that the maximization of (9) with respect to μ_i' (under the constraint $\sum_{i=1}^{N}\mu_i'(t)=1$) and with respect to the set (r_i', α_i', β_i') can be performed independently. The first maximisation yields the new μ_i: $\mu_i=\sum_{t\in|T|}\pi_i(t)$ where $|T|$ denotes the number of points in T. The second one is reduced to the maximisation of:

$$\sum_{t\in T}\pi_i\left(t\right)\left\{-\log\alpha_i-\frac{\left|\log\left|R\left(t\right)/r_i\right|\right|}{\alpha_i}\right\}+$$
$$\sum_{t\in T}\pi_i\left(t\right)\left\{\frac{\left(1-\beta_i\right)}{\beta_i}\cos\arg\frac{R\left(t\right)}{r_i}-\log C\left(\beta_i\right)\right\}$$

with respect to r_i, α_i, and β_i. it's maximization with respect to r_i yields to:

$$r_i=\exp\left\{\text{med}\left[\log\left|R\left(t\right)\right|,\pi_i\left(t\right)\right]\right\}\text{sign}\sum_{t\in T}\pi_i\left(t\right)\text{sign}\,R\left(t\right)$$

where med$\{\xi(t),\mu_i(t)\}$ is the median of $\xi(t)$ with the relative probabilities $\mu_i(t)$ and sign$(z)=z/|z|$. Inserting this value into the above expression, it can be seen that it is maximised when:

$$\alpha_i=\frac{\sum_{t\in T}\pi_i\left(t\right)\left|\log\left|R\left(t\right)/r_i\right|\right|}{\sum_{t\in T}\pi_i\left(t\right)}$$

and β_i is the solution of:

$$\frac{\beta^2}{C\left(\beta\right)}\frac{dC\left(\beta\right)}{d\beta}=\frac{\sum_{t\in T}\pi_i\left(t\right)\left\{1-\cos\arg R\left(t\right)/r_i\right\}}{\sum_{t\in T}\pi_i\left(t\right)}$$

The above left hand side can be shown to be an increasing function of β: tabulating this function allows us to find the solution by interpolation.

4.3. EM initialisation

A simple way to initialise our algorithm is to choose randomly N times points t_1,.., t_N and initialise r_i by $r_i=R(t_i)$, $i=1..N$. Thus, $\pi_i(t)$ can be initialised by formula (7) using the (non approximated) density (5) with ρ_i set to one. This yields:

$$\pi_i(t)=\frac{d[R(t),r_i]}{\sum_{i=1}^{N}d[R(t),r_i]}$$

where $d\left[R\left(t\right),r_i\right]=\cosh\log\left[R\left(t\right)/r_i\right]-\cos\arg\left[R\left(t\right)/r_i\right]$

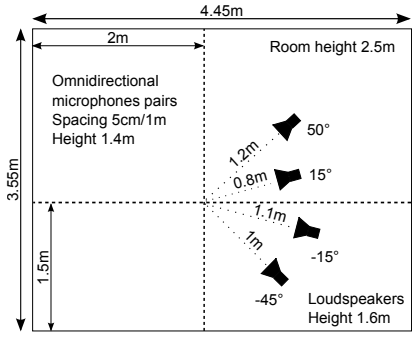

Fig. 1. Recording arrangement used for development data.

5. EXPERIMENTS AND RESULTS

The algorithm is evaluated on the same database used in the audio source separation campaign [1, 8]. The room's dimension and the respective position of sources and microphones are given in Fig.1. We focus here on live recording of four male and four female speech signals, sampled at 16 kHz and with a 10s duration. In order to evaluate the robustness of the method, two microphones configurations are considered: 5 cm and 1 m spacing. The algorithm uses a 2048 samples length Hanning window with a 75% overlap and reconstruction is achieved using the overlap add method. The separation performance was evaluated for the estimated source i by the Signal to Interference Ratio (SIR_i), Image to Signal Ratio (ISR_i), Signal to Distortion Ratio (SDR_i) and Signal to Artefact Ratio (SAR_i) improvements given by:

$$SDR_i=10\log_{10}\frac{\sum_{j=1}^{M}\sum_t\left[s_j^{(i)}\left(t\right)\right]^2}{\sum_{j=1}^{M}\sum_t\left[\text{es}_j^{(i)}\left(t\right)+\text{ei}_j^{(i)}\left(t\right)+\text{ea}_j^{(i)}\left(t\right)\right]^2}$$

$$ISR_i=10\log_{10}\frac{\sum_{j=1}^{M}\sum_t\left[s_j^{(i)}\left(t\right)\right]^2}{\sum_{j=1}^{M}\sum_t\left[\text{es}_j^{(i)}\left(t\right)\right]^2}$$

$$SIR_i=10\log_{10}\frac{\sum_{j=1}^{M}\sum_t\left[s_j^{(i)}\left(t\right)+\text{es}_j^{(i)}\left(t\right)\right]^2}{\sum_{j=1}^{M}\sum_t\left[\text{ei}_j^{(i)}\left(t\right)\right]^2}$$

$$SAR_i=10\log_{10}\frac{\sum_{j=1}^{M}\sum_t\left[s_j^{(i)}\left(t\right)+\text{es}_j^{(i)}\left(t\right)+\text{ei}_j^{(i)}\left(t\right)\right]^2}{\sum_{j=1}^{M}\sum_t\left[\text{ea}_j^{(i)}\left(t\right)\right]^2}$$

where $\text{es}_j^{(i)}(t)$, $\text{ei}_j^{(i)}(t)$ and $\text{ea}_j^{(i)}(t)$ represent filtering distortion, interference and artifacts. These three distinct errors are obtained by decomposing the estimated contribution of source i to the j^{th} channel, $\hat{s}_j^{(i)}(t)$, into:

Live Recordings		Female 5cm				Male 5cm				Female 1m				male 1m				OAP
		S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	MeanS_i
MBUSS Algorithm	SDR(dB)	2.9	2.5	4.0	4.1	2.1	1.4	3.9	3.9	4.1	2.7	5.8	3.9	3.1	2.5	4.7	4.1	3.5
	ISR(dB)	5.2	4.3	5.4	9.0	4.8	2.4	7.9	6.7	6.7	5.2	8.5	6.0	5.3	5.2	6.9	7.2	6
	SIR(dB)	3.8	3.6	5.7	4.6	1.0	0.7	4.0	5.5	5.9	2.5	8.4	5.3	4.0	1.4	7.1	5.6	4.3
	SAR(dB)	6.7	6.4	8.4	8.3	6.7	5.1	8.8	8.1	7.8	6.4	10.0	7.5	6.0	6.2	8.0	7.3	7.4
Sawada Algorithm	SDR(dB)	2.6	-0.8	1.7	4.2	3.0	0.4	3.2	3.9	4.5	3.8	7.4	3.3	3.0	1.5	5.2	2.3	3.1
	ISR(dB)	6.5	4.5	2.8	10.4	7.4	1.4	10.5	9.9	9.1	8.0	13.1	6.2	7.9	4.7	9.0	6.5	7.4
	SIR(dB)	4.4	-2.2	5.4	7.7	5.6	1.7	4.0	6.7	8.0	7.1	12.2	7.4	5.1	2.6	11.0	4.7	5.7
	SAR(dB)	5.6	6.3	3.7	6.1	4.6	0.7	7.1	6.2	6.3	5.4	9.5	4.7	4.7	2.7	6.1	4.6	5.3

Table 1. Results for live recording with two microphone spacing (5cm, 1m) and two different types of speaker the overall performance of the source separation is presented in the last column.

$$\hat{s}_j^{(i)}(t) = s_j^{(i)}(t) + \text{es}_j^{(i)}(t) + \text{ei}_j^{(i)}(t) + \text{ea}_j^{(i)}(t)$$

Roughly, the $\text{es}_j^{(i)}(t)$ stands for the distance between the estimated $\hat{s}_j^{(i)}(t)$ source and the filtered version of the source, $\text{ei}_j^{(i)}(t)$ is the quantity of other sources present in the estimated source and $\text{ea}_j^{(i)}(t)$ stands for the degradation of the estimated source itself (see for instance [8]).

Performance results are given in Table 1 and are compared with one of the most efficient algorithm in the campaign. First of all, let us note that, in terms of SDR which is the global performance, the proposed algorithm is more or less equivalent to the Sawada algorithm: the overall performance (OAP) column shows a slight advantage for the proposed method (3.5dB vs 3.1dB). The intermediate errors show that the compromise operated by the algorithms is different: Sawada's method favours weak interferences (SIR=5.7) to the cost of more degraded separated speech (SAR=5.3), when the proposed algorithm present stronger interference (SIR=4.3) but with a clearer separated speech signal (SAR=7.4).

These results are confirmed by informal listening tests: less interference is audible in the Sawada's results, at the expense of more audible artefact and attenuation on speech. Besides the separation method itself, the difference in the objective and subjective results seems also be connected to the smoothing of the separation masks: on the contrary to the Sawada algorithm, no smoothing is applied to our separation masks. Smoothing permits to reduce the interference characterized by small TF supports (few TF bins), at the expense of slight degradation and attenuation of the separated source.

6. CONCLUSION

In this paper, we proposed a novel Model-Based Underdetermined Speech Separation (MBUSS) algorithm based on standard binaural cues, i.e. IPD and $\log$(ILD). Based on the sparseness assumption and a specific model for the (IPD,$\log$(ILD)), the algorithm demonstrated its ability to separate undetermined reverberant mixtures, in terms of objective criteria as well as in terms of subjective listening. Nevertheless, some work remains to be done: the frequency band permutation has to be treated and more research needs to be done on the effect of the model regardings to source position and room reverberation.

7. REFERENCES

[1] P. Bofill S. Makino E. Vincent, H. Sawada and J.P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," *ICA*, 2007.

[2] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *Signal Processing, IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[3] Shoko Araki, Hiroshi Sawada, Ryo Mukai, and Shoji Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, 2007.

[4] Stefan Winter, Walter Kellermann, Hiroshi Sawada, and Shoji Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1-norm minimization," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 81–81, 2007.

[5] Michael I. Mandel and Daniel P. W. Ellis, "Em localization and separation using interaural level and phase cues," *WASPAA 2007*, pp. 275–278, 21-24 Oct. 2007.

[6] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain bss," *ISCAS 2007*, pp. 3247–3250, May 2007.

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

[8] ," http://sassec.gforge.inria.fr/.