IMPROVED COCKTAIL-PARTY PROCESSING

Alexis Favrot, Markus Erne
Scopein Research
Aarau, Switzerland
postmaster@scopein.ch

Christof Faller
Audiovisual Communications Laboratory, LCAV
Swiss Institute of Technology
Lausanne, Switzerland
christof.faller@epfl.ch

ABSTRACT

The human auditory system is able to focus on one speech signal and ignore other speech signals in an auditory scene where several conversations are taking place. This ability of the human auditory system is referred to as the "cocktail-party effect".

This property of human hearing is partly made possible by binaural listening. Interaural time differences (ITDs) and interaural level differences (ILDs) between the ear input signals are the two most important binaural cues for localization of sound sources, i.e. the estimation of source azimuth angles.

This paper proposes an implementation of a cocktail-party processor. The proposed cocktail-party processor carries out an auditory scene analysis by estimating the binaural cues corresponding to the directions of the sources. And next, as a function of these cues, suppresses components of signals arriving from non-desired directions, by speech enhancement techniques. The performance of the proposed algorithm is assessed in terms of directionality and speech quality.

The proposed algorithm improves existing cocktail-party processors since it combines low computational complexity and efficient source separation. Moreover the advantage of this cocktail-party processor over conventional beam forming is that it enables a highly directional beam over a wide frequency range by using only two microphones.

1. INTRODUCTION

1.1. Overview

The "cocktail-party effect" is the ability of the human auditory system to select one desired sound from an ambient background of noise, reflections, or other sounds. For instance, at a party, where many talkers are speaking simultaneously, humans may focus their attentions on one voice and ignore other voices and noise which are possibly equally strong in loudness.

The concept of a cocktail-party processor, motivated by simulating electronically the "cocktail-party effect", has been introduced earlier in [1]. The algorithm simulated neural excitation patterns based on specific physiological assumptions about the auditory system. Next, by a model of central stages of the signal processing in the auditory system, a spatial analysis of the auditory scene was performed in order to predict the azimuth angles of the sound sources. These spatial parameters were then used to control the transfer function of a time-variant filter, removing the components of signal arriving from non-desired directions.

For low computational complexity, the proposed cocktail-party processor makes simplified physiological assumptions compared to [1]. Using a FFT based time-frequency representation of the ear input signals, the proposed algorithm first estimates the binaural localization cues (ITDs and ILDs) related to the azimuth angles of the sources to be recovered. Next, different speech enhancement techniques are controlled as a function of these binaural cues: in addition to conventional short-time spectral modification, as in [1], the proposed algorithm applies blind source separation in order to improve source separation.

1.2. Mixing model

The goal of the proposed cocktail-party processor is to recover a desired speech signal given two linear mixtures of speech signals, representing the right and left ear input signals, $x_R[n]$ and $x_L[n]$:

$$x_R[n] = \sum_{i=1}^{N} s_i[n] \quad \text{and} \quad x_L[n] = \sum_{i=1}^{N} a_i s_i[n - d_i] \,, \qquad (1)$$

where $s_1, \ldots, s_N$ are the N speech sources, spatially distributed as represented in Figure 1. a_i and d_i are the attenuation coefficient and time delay associated with the path from the i^{th} source to the left ear. The azimuth angle of the i^{th} source is Φ_i. Note that it is assumed that all sources are in different directions[1], and only the direct paths are considered, i.e. we assume anechoic conditions.

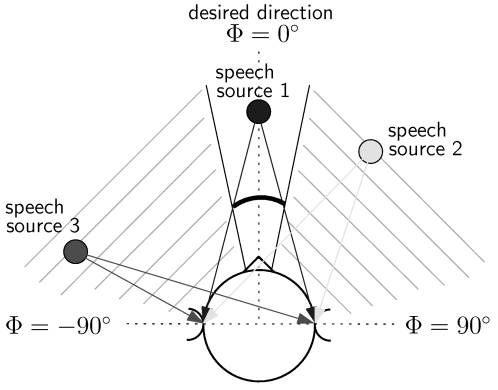

Figure 1: *The ear input signals are linear mixtures of the speech signals coming from spatially distributed sound sources.*

[1] If several sources are in the same direction they are considered as a single source.

Let $s_1[n]$ be the desired speech signal arriving from the direction defined by $\Phi_1 = 0°$. The other speech sources are considered as interfering speech sources. For a low computational complexity, we consider the short-time spectra of the ear input signals, $\mathcal{X}_R[m,k]$ and $\mathcal{X}_L[m,k]$, obtained by a windowed short-time Fourier transform (STFT). m denotes the frequency index, and k the frame number.

2. CONSIDERING BINAURAL CUES

2.1. Definition of binaural cues

Localization of sound is partly made possible by capturing the slight differences between sound signals at the right and left ear entrances. In order to understand how the auditory system estimates the direction of arrival of a sound, we first consider a single sound source [2]. The Ear input signals can be seen as filtered versions of the source signal, as shown in Figure 2(a): the filters used are referred to as head related transfer function (HRTF). But a more simple manner to model the ear input signals is to assume only a difference of path length from the source to both ears, as shown in Figure 2(b). As a result of this path length difference,

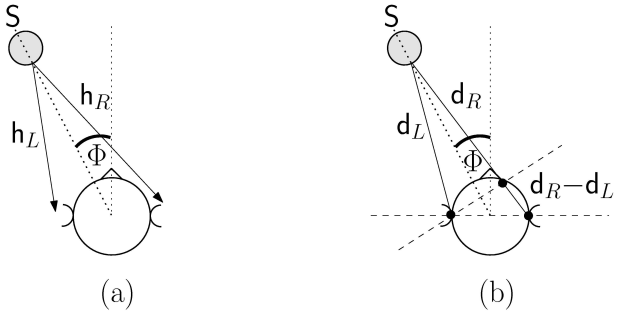

Figure 2: *(a): Ear input signals modelled as filtered versions of the source signal, by the HRTFs h_R and h_L . (b): Ear input signals modelled with a difference in length of paths $d_R - d_L$ to both ears. HRTFs and the difference in length of paths are linked to the azimuth angle Φ of the source.*

there is a difference in time of arrival of sound, denoted interaural time difference (ITD). Additionally, the shadowing of the head results in an intensity difference to the right and left ear input signals, denoted interaural level difference (ILD). ITD and ILD are the binaural localization cues of the considered sound source. They are directly linked to the azimuth angle Φ of this source.

2.2. Auditory scene analysis

The auditory scene analysis (source localization) is important for ultimately estimating the source signals. The directions of the sources are evaluated based on the ITDs. Next, the corresponding ILDs are computed by considering HRTF data lookup.

Source localization is mainly based on the coherence function between right and left ear input signals:

$$\Gamma_{LR}[m,k] = \frac{\Psi_{LR}[m,k]}{\sqrt{\Psi_{LL}[m,k]\Psi_{RR}[m,k]}} \ , \qquad (2)$$

where m is the frequency index, k is the frame number, and

$$\Psi_{LR}[m,k] = \mathrm{E}\left\{\mathcal{X}_L[m,k]\mathcal{X}_R^*[m,k]\right\} \ , \qquad (3)$$

where $\mathrm{E}\{\ldots\}$ stands for the mathematical expectation [3]. In the time domain the coherence function $\Gamma_{LR}[m,k]$ corresponds to the normalized cross-correlation function $\gamma_{LR}[n,k]$ between $x_R[n]$ and $x_L[n]$. $\gamma_{LR}[n,k]$ is evaluated over time lags in the range of $[-1,1]$ ms, i.e $n/f_s \in [-1,1]$ ms, where f_s is the sampling rate. If only a single source s_i is emitting sound, the ITD is estimated as the lag of the peak of the normalized cross-correlation function:

$$\mathrm{ITD}_i = \arg\max_n \gamma_{LR}[n,k] \ . \qquad (4)$$

In a more complex auditory scene, where a number of sources are emitting simultaneously, we assume that the auto-correlation functions $\gamma_{s_i}[n,k]$ of the source signals $s_i[n]$ do not overlap. And thus the resulting cross-correlation function $\gamma_{LR}[n,k]$ is the sum of auto-correlation functions $\gamma_{s_i}[n,k]$, shifted in time by the corresponding ITD_i. Figure 3 illustrates the peaks detection of the normalized cross-correlation function. A peak corresponds to a source emitting from a direction leading to a time lag, corresponding to the ITD, in the normalized cross-correlation $\gamma_{LR}[n,k]$.

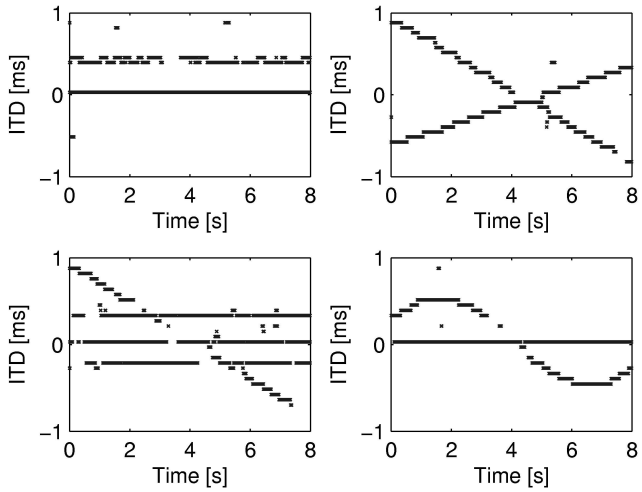

Figure 3: *Different auditory scenes analyzed by ITD estimation. Two static sources are emitting simultaneously with two different ITDs (right top corner). Two sources are moving linearly over time (left top corner). Three static sources with an additional source moving linearly (left bottom corner). A static source with a source moving by following a cosine law (right bottom corner).*

So far, we have only considered ITD. However, in order to analyze precisely the auditory scene, ILD needs to be taken into account. For each sound source s_i, the missing cue (ILD_i) is evaluated from a head related transfer function (HRTF) data lookup.

ITD and ILD can be described as functions of azimuth angle and frequency: $\mathrm{g}_T(m,\Phi)$ and $\mathrm{g}_L(m,\Phi)$, respectively. While ITD can be approximatively considered as independent of frequency, ILD is highly frequency dependent and is relevant for spatial perception for frequencies above about 1.5 kHz. However, we only estimate a single full band ILD in order to complete the scene analysis. In this case ITD and ILD are independent of frequency

by considering a weighted sum of the two dimensional functions, $\mathrm{g}_T(m,\Phi)$ and $\mathrm{g}_L(m,\Phi)$, among frequencies,

$$\begin{aligned}\text{ITD} &= \mathrm{g}_T(\Phi) = \sum_m c_T^m \mathrm{g}_T(m,\Phi) \\ \text{ILD} &= \mathrm{g}_L(\Phi) = \sum_m c_L^m \mathrm{g}_L(m,\Phi)\ ,\end{aligned} \tag{5}$$

where c_T^m and c_L^m are the frequency dependent scale factors for ITD and ILD, determined by the HRTF CIPIC Database [4].

Now, the azimuth angle Φ_i of each source can be calculated from the ITD_i by using the inverse function $\mathrm{g}_T^{-1}(\text{ITD}_i)$. And next from this azimuth angle we can estimate the corresponding ILD_i, as shown in Figure 4.

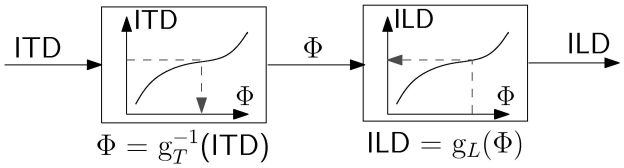

Figure 4: *Evaluation of azimuth angle from estimated ITD by an inverse function, followed by the estimation of ILD by the function which directly links azimuth angles to ILD.*

The auditory scene analysis yields a single pair of full band binaural cues (ITD_i and ILD_i) for each speech source s_i. These binaural cues are directly linked to the direction of arrival of source s_i.

2.3. Using binaural cues for speech enhancement techniques

In the proposed algorithm, the main application of previously binaural cues estimation is to get source dependent parameters in order to control the used speech enhancement techniques: blind source separation (BBS) and noise-adaptive spectral magnitude expansion (NASME), presented in Sections 3 and 4, respectively.

BBS needs to solve the mixing model in equation (1). The estimated binaural cues are directly linked to the attenuation coefficients a_i and the time delays d_i, defined in equation (1). Indeed, for a source s_i, a positive azimuth angle Φ_i corresponds to a point source situated at the right side with respect to the head of the listener. Also the sound of a source localized on the right side of the head will arrive first at the right ear ($d_i > 0$ ms) and the signal level will be stronger at the right ear ($a_i < 0$ dB). Because of $\mathrm{g}_T(\Phi)$ and $\mathrm{g}_L(\Phi)$ being monotonically increasing, a positive Φ_i yields positive ITD ($\text{ITD}_i > 0$ ms) and ILD ($\text{ILD}_i > 0$ dB). More generally we can write for the source i,

$$\begin{aligned} d_i &= \text{ITD}_i \quad \text{[samples]} \\ 20\log_{10}(a_i) &= -\text{ILD}_i \quad \text{[dB]}\ . \end{aligned} \tag{6}$$

Moreover, NASME requires signal statistics in order to be carried out. The variances of the sources signals are estimated from the power spectra of the input signals. since speech signals are assumed stationary over short time periods between 10 ms and 20 ms. A short-time estimate of the frequency domain cross-correlation between $x_R[n]$ and $x_L[n]$, defined in (3), is obtained by:

$$\Psi_{LR}[m,k] = \alpha \mathcal{X}_L[m,k]\mathcal{X}_R^*[m,k] + (1-\alpha)\Psi_{LR}[m,k-1]\ , \tag{7}$$

where the factor α determines the degree of smoothing over time. With an inverse fourier transform, the smoothed time domain cross-correlation $\psi_{LR}[n,k]$ is obtained. As before, under the assumption that the resulting cross-correlation function $\psi_{LR}[n,k]$ is the sum of auto-correlation functions $\psi_{s_i}[n,k]$ of the source signals $s_i[n]$, shifted in time by the corresponding ITD_i, the variance $\sigma_{s_i}^2[k]$ is:

$$\sigma_{s_i}^2[k] = \psi_{LR}[\text{ITD}_i, k]\ . \tag{8}$$

The variances $\sigma_{x_R}^2[k]$ and $\sigma_{x_L}^2[k]$, of ear input signals $x_R[n]$ and $x_L[n]$, are computed in the same manner by considering the auto-correlation functions $\Psi_{RR}[m,k]$ and $\Psi_{LL}[m,k]$.

As a conclusion, the mixing model defined in equation (1) has been solved. Additionally, the short-time variances of the signals $s_i[n]$, $x_R[n]$ and $x_L[n]$ have been estimated. The resulting parameters, a_i, d_i, $\sigma_{s_1}^2[k]$, $\sigma_{x_R}^2[k]$ and $\sigma_{x_L}^2[k]$, are used to control speech enhancement techniques.

3. BLIND SOURCE SEPARATION

3.1. W-disjoint orthogonality

The first speech enhancement technique to be used is blind source separation (BSS) [5]. The goal of BSS is to recover the original source signals, given linear mixtures of these source signals. The considered linear mixtures are defined in equation (1). By performing a discrete windowed STFT, with a suitable window function $W[n]$, the mixing model can be expressed in the frequency domain as:

$$\begin{aligned} \mathcal{X}_R[m,k] &= \sum_{i=1}^{N} \mathcal{S}_i[m,k], \\ \mathcal{X}_L[m,k] &= \sum_{i=1}^{N} a_i \mathcal{S}_i[m,k] e^{-J\frac{2\pi m d_i}{M}}\ , \end{aligned} \tag{9}$$

where M is the length of the discrete fourier transform (DFT).

In BSS, it is assumed that the spectra, $\mathcal{S}_1[m,k], \ldots, \mathcal{S}_N[m,k]$, of the N source signals satisfy the W-disjoint orthogonality condition. W-disjoint orthogonality corresponds to non-overlapping windowed STFT representations of the sources. This condition

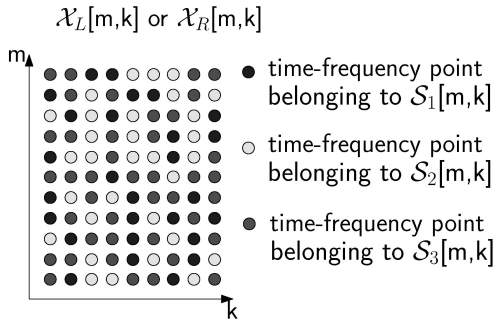

Figure 5: *Time-frequency representation of the ear input signals for a scenario as is shown in Figure 1. The spectrogram with two-dimensional time-frequency grid shows the basic of W-disjoint orthogonality assumption that each point of this grid is related to only one of the three sources.*

means that at most one source is active at each time-frequency point $[m, k]$. That is, each point of the time-frequency grid represents only one source, as illustrated in Figure 5.

3.2. Time-frequency masks

In order to decide on the pairing between speech sources and time-frequency points, for each source i, the maximum likelihood function is evaluated:

$$L_i[m,k] = \frac{1}{2\pi}^{-\frac{1}{2(1+a_i^2)}|a_i e^{-j\frac{2\pi m d_i}{M}}\mathcal{X}_L[m,k]-\mathcal{X}_R[m,k]|^2}, \quad (10)$$

where the parameters a_i and d_i have been evaluated estimated by the audio scene analysis in (6). $L_1[m,k]$ is the likelihood that the source s_1 is dominant at time-frequency point $[m,k]$. The points $[m,k]$ of the time-frequency grid, which represent the source s_1, satisfy:

$$\forall\, i \neq 1,\ L_i[m,k] < L_1[m,k]\ . \quad (11)$$

Then the binary time-frequency mask, used for extracting the contributions of the source s_1 from the ear input spectra, is computed as follows:

$$\mathrm{M}_1[m,k] = \begin{cases} 1 & \forall\, i \neq 1,\ L_i[m,k] < L_1[m,k] \\ 0 & \text{otherwise}\ . \end{cases} \quad (12)$$

From this mask, the source s_1 is recovered from the mixtures by:

$$\begin{array}{ll} \mathcal{S}_{1_R}[m,k] = \mathrm{M}_1[m,k].\mathcal{X}_R[m,k] & \text{for the right ear,} \\ \mathcal{S}_{1_L}[m,k] = \mathrm{M}_1[m,k].\mathcal{X}_L[m,k] & \text{for the left ear;} \end{array} \quad (13)$$

and by considering both ears, the spatial distribution of the sources is preserved.

4. NOISE-ADAPTIVE SPECTRAL MAGNITUDE EXPANSION

4.1. Gain filter

The second speech enhancement technique implemented is noise-adaptive spectral magnitude expansion (NASME) [6]. This technique combines both compandors and conventional noise reduction techniques such as parametric spectral subtraction. The main idea is to adapt the spectral magnitude expansion as a function of noise level and spectral components. NASME focuses on the suppression of uncorrelated additive background noise,

$$x[n] = s_1[n] + v[n]\ , \quad (14)$$

where $v[n]$ is the noise measured at the ear entrance. Note that $x[n]$ can represent either the right or the left ear input signal, since NASME is performed separately on each channel and thus the parameters a_i and d_i are not considered.

By analogy with parametric spectral subtraction, the estimated desired speech signal magnitude spectrum can be computed with the gain filter H by:

$$\hat{\mathcal{S}}_1[m,k] = \mathrm{H}[m,k].|\mathcal{X}[m,k]|e^{j\arg\mathcal{X}[m,k]} = \mathrm{H}[m,k].\mathcal{X}[m,k]\ . \quad (15)$$

The phase remains unchanged by this filtering, which has no consequence since the human perception is relatively insensitive to phase corruption.

In NASME, the gain filter, H, is given by:

$$\mathrm{H}\left(\frac{|\hat{\mathcal{V}}[m,k]|}{|\mathcal{X}[m,k]|}\right) = \left[\mathrm{A}[m,k]\frac{|\hat{\mathcal{V}}[m,k]|}{|\mathcal{X}[m,k]|}\right]^{1-\theta[m,k]}\ . \quad (16)$$

Moreover |H| is upper-bounded by 1. $\mathrm{A}[m,k]$ is the crossover point, used to adapt the gain filter to the estimated noise magnitude spectrum $|\hat{\mathcal{V}}[m,k]|$ and $\theta[m,k]$ controls the expansion as a function of the inverse signal to noise ratio (SNR).

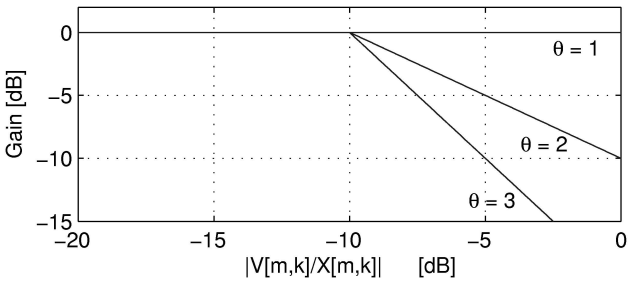

Figure 6: *Gain filters $H[m,k]$ for several parameters θ and a constant crossover point $A = 10$ dB are plotted as functions of the inverse signal to noise ratio.*

The gain curves, as a function of the inverse SNR, for several expansion powers θ and a constant crossover point $\mathrm{A} = 10$ dB, are shown in Figure 6.

4.2. Extension to speech signals

In the proposed cocktail-party processor, NASME is used to enhance a desired speech signal out of the spatial distribution of concurrent speech signals. In this case, the noise signal $v[n]$ is composed of speech signals which are not completely uncorrelated with the desired speech signal and are not stationary:

$$v[n] = \sum_{i=2}^{N} s_i[n]\ . \quad (17)$$

But such signals are considered as statistically reasonably independent if they are observed over a sufficient long period of time. And over a sufficient short period of time they can be considered as stationary. By choosing a suitable analysis frame size, we assume that speech signals satisfy statistical independence and stationarity. By doing some approximations, we adapt NASME to mixtures of speech signals, as explained next.

The first approximation is related to the estimated noise spectrum. Indeed the noise spectrum is a combination of the spectra $\mathcal{S}_2[m,k],\ldots,\mathcal{S}_N[m,k]$. And since each of the $N-1$ noise sources can not a priori be separated, the noise magnitude spectrum can not be directly estimated. But by assuming that $s_1[n]$ and $v[n]$ are uncorrelated, the instantaneous power spectrum of the noise $v[n]$ can be recovered by subtracting an estimate of $|\mathcal{S}_1[m,k]|$ from the estimate $|\hat{\mathcal{X}}[m,k]|$:

$$|\hat{\mathcal{V}}[m,k]|^2 = |\hat{\mathcal{X}}[m,k]|^2 - |\hat{\mathcal{S}}_1[m,k]|^2\ . \quad (18)$$

The corresponding noise spectral magnitude is,

$$|\hat{\mathcal{V}}[m,k]| = \sqrt{|\hat{\mathcal{V}}[m,k]|^2} = \left[|\hat{\mathcal{X}}[m,k]|^2 - |\hat{\mathcal{S}}_1[m,k]|^2\right]^{\frac{1}{2}}\ . \quad (19)$$

A more general form can be derived by introducing the parameters α, β and γ,

$$|\hat{\mathcal{V}}[m,k]| = \left[|\hat{\mathcal{X}}[m,k]|^{\alpha} - \gamma|\hat{\mathcal{S}}_1[m,k]|^{\alpha}\right]^{\beta} . \qquad (20)$$

where α and β are exponents and γ controls the estimation of $|\mathcal{S}_1[m,k]|$ in case it is under or over estimated. The estimated noise magnitude spectrum is calculated only from the spectra of the ear input signals and the desired signal. This method has the advantage that the computation time is reduced, and if the number N of sources becomes large, the computation time stays the same.

As a second approximation, the variances of signals are estimated rather than their entire power spectrum. The magnitude spectrum $|\hat{\mathcal{S}}_1[m,k]|$ is estimated according to:

$$|\hat{\mathcal{S}}_1[m,k]| = \sqrt{\sigma^2_{s_1}[m,k]} \approx \sqrt{|\mathcal{S}_1[m,k]|^2} . \qquad (21)$$

The variance of the desired speech signal, $\sigma^2_{s_1}[m,k]$, as well as the variances, $\sigma^2_{x_R}[m,k]$ and $\sigma^2_{x_L}[m,k]$, of the ear input signals have been estimated from the auditory scene analysis.

Then, the noise magnitude spectrum is directly computed from equations (20) and (21):

$$|\hat{\mathcal{V}}[m,k]| = \left[\sqrt{\sigma^2_x[m,k]}^{\alpha} - \gamma\sqrt{\sigma^2_{s_1}[m,k]}^{\alpha}\right]^{\beta} . \qquad (22)$$

Finally the gain filter H defined in equation (16) becomes:

$$\mathrm{H}_1[m,k] = \left(\mathrm{A}[m,k]\left[\frac{|\sqrt{\sigma^2_x[m,k]}^{\alpha} - \gamma\sqrt{\sigma^2_{s_1}[m,k]}^{\alpha}|}{\sqrt{\sigma^2_x[m,k]}^{\alpha}}\right]^{\beta}\right)^{1-\theta[m,k]} , \qquad (23)$$

where $\mathrm{A}[m,k]$ defines the crossover point, and $\theta[m,k]$ controls the expansion power. Finally, the source s_1 is recovered from the mixtures by:

$$\begin{array}{ll} \mathcal{S}_{1_R}[m,k] = \mathrm{H}_1[m,k].\mathcal{X}_R[m,k] & \text{for the right ear,} \\ \mathcal{S}_{1_L}[m,k] = \mathrm{H}_1[m,k].\mathcal{X}_L[m,k] & \text{for the left ear.} \end{array} \qquad (24)$$

5. THE PROPOSED COCKTAIL-PARTY PROCESSOR

The proposed cocktail-party processor combines both BSS and NASME as illustrated in the block diagram in Figure 7. The first step is concerned with time-frequency transform adapted to speech signals. Then the scene analysis is carried out by estimating the binaural cues (6) related to the directions of the sources to be recovered. The different source dependent parameters, are evaluated using these binaural cues. Next, as the function of these parameters, the speech enhancement techniques, blind source separation and noise-adaptive spectral magnitude expansion, are performed simultaneously. However, the uniform spectral resolution of the STFT is not well adapted to human perception. Therefore, BSS and NASME are carried out within critical bands, which are formed by grouping the STFT coefficients such as each group corresponds to a critical band.

The binary time-frequency mask defined in equation (12) and the gain filter defined in equation (23) are combined together in a combined gain filter $\mathrm{G}_1[m,k]$. The combined gain filter which is used to recover speech source s_1, is given by:

$$\mathrm{G}_1[m,k] = \mathrm{M}_1[m,k].\mathrm{H}_1[m,k] . \qquad (25)$$

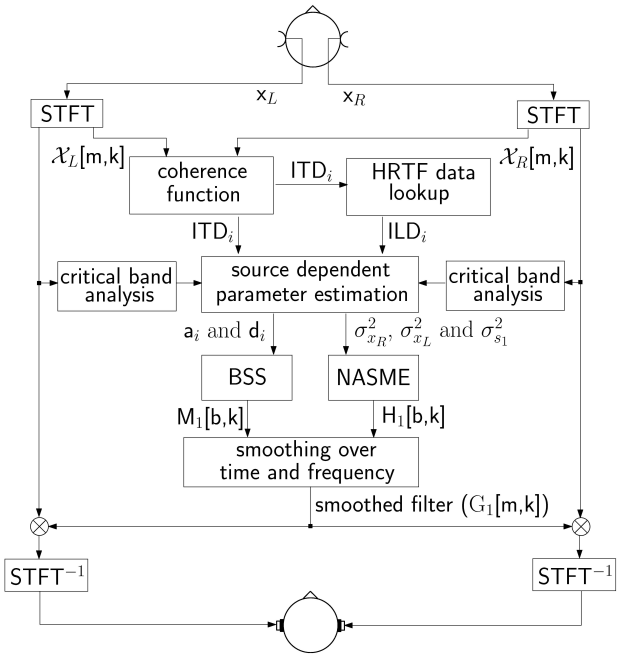

Figure 7: *Detailed block diagram of the proposed algorithm for the cocktail-party processor.*

In order to reduce artifacts and distortions, the last step is devoted to time and frequency smoothing of the combined gain filter applied to the ear inputs signals, which are converted back into the time domain.

6. PERFORMANCE

6.1. Directionality pattern

The ability of the proposed cocktail-party processor to suppress interfering sources arriving from non-desired directions can be expressed by means of directionality patterns. The desired direction is defined by azimuth angle $\Phi_1 = 0^{\circ}$. The simulations involve input signals, coming from different directions between -90° and 90°, which have been obtained by convolving white noises with HRTFs. The attenuations of the output signals have been plotted within different critical bands.

The resulting directionality patterns are shown in Figure 8: they are narrow even at low frequencies and their widths are nearly independent of frequency.

The cocktail-party processor enables a highly directive beam over a wide range of frequency with only two microphones placed at the ear entrances. With two microphones, conventional beam former are much more limited in terms of directionality.

6.2. Intelligibility

For performance evaluation, a concurrent speech signal s_2 is added to the desired speech signal s_1 at the mean SNR of 0 dB. In the processed signal, at the output of the cocktail-party processor, the concurrent signal is attenuated by 15 dB and only slight changes compared to the desired signal can be observed by visual inspection of Figure 9.

concurrent source(s)	mean STI	min STI	max STI
1	0.95	0.92	0.98
2	0.89	0.78	0.95
3	0.82	0.73	0.88
4	0.73	0.61	0.80

Table 1: *The STI is evaluated under diverse acoustical conditions.*

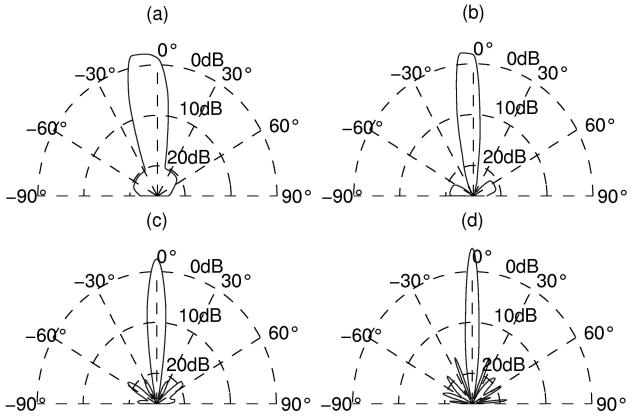

Figure 8: *Directionality patterns of the cocktail-party processor within different critical bands. (a): Critical band* $125 - 250$ *Hz. (b): Critical band* $620 - 745$ *Hz. (c): Critical band* $1425 - 1800$ *Hz. (d): Critical band* $3535 - 4400$ *Hz.*

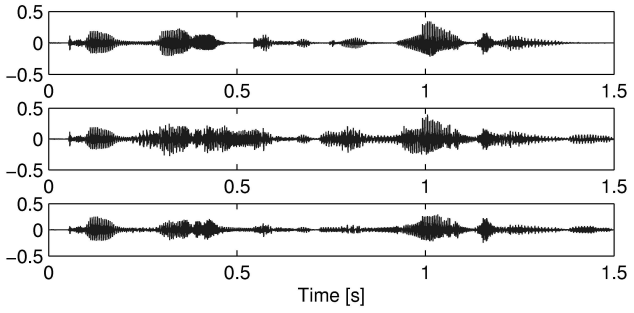

Figure 9: *The desired speech signal (top) of a female speaker* ($\Phi_1 = 0°$)*. A concurrent male speech signal* ($\Phi_2 = 30°$) *is added to the desired speech (middle). The output of the cocktail-party processor (bottom).*

The intelligibility of the processed signal is evaluated by calculating the speech transmission index (STI) of the cocktail-party processor, in order to find a good trade-off between the degree of suppression of signal components and resulting distortions. The STI is a single number between 0 (unintelligible) and 1.0 (perfectly intelligible) [7].

The STI, for the proposed cocktail-party processor, is calculated for a set of different HRTFs and under diverse acoustical conditions: with several interfering sources coming from different directions of arrival. The results are presented in Table 1. For only one concurrent signal the intelligibility remains nearly perfect, but for two concurrent signals the intelligibility starts to be deteriorated. By increasing the number of concurrent signals the intelligibility becomes worse, but remains still excellent (that means larger than 0.75 for the STI scales) with up to three interfering sources.

7. CONCLUSIONS

In this paper, we presented a cocktail-party processor controlled by binaural localization cues and signal statistics. The proposed algorithm improves source separation in existing cocktail-party processors by implementing blind source separation. The good performance of this algorithm has been demonstrated in terms of directionality and intelligibility by using the STI.

The proposed algorithm is expected to be of advantage for many applications such as automatic speech recognition, intelligent hearing aids, or speaker identification. Often, a low computational complexity is needed for real-time application. The proposed algorithm, implemented with a FFT, offers such a computational complexity.

8. ACKNOWLEDGEMENTS

Numerous individuals, at LCAV and Scopein Research, have contributed suggestions, thoughts, references, potential problems, and perspectives that have shaped this work.

9. REFERENCES

[1] M. Bodden, "Modeling human sound source localization and the cocktail-party-effect," *Acta Acustica 1*, vol. 1, pp. 43–55, Ferbuary/Apr. 1993.

[2] J. Blauert, *Spatial hearing: The psychophysics of human sound localization*, revised ed. Cambridge, Massachusetts, USA: The MIT Press, 1997.

[3] C. Faller, "Parametric coding of spatial audio," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, July 2004, thesis No. 3062, [Online] http://library.epfl.ch/theses/?nr=3062.

[4] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop Appl. of Dig. Sig. Proc. to Audio and Acoust.,* New Palz, NY, Oct. 2001, pp. 99–102.

[5] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[6] W. Etter and G. S. Moschytz, "Noise reduction by noise-adaptive spectral magnitude expansion," *J. Audio Eng. Soc.*, vol. 42, pp. 341–349, May 1994.

[7] H. J. M. Steeneken and T. Houtgast, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, Mar.. 1985.