A Sector-Based, Frequency-Domain Approach to Detection and Localization of Multiple Speakers

Guillaume Lathoud [a,b]
Mathew Magimai.-Doss [a,b]

IDIAP–RR 04-54

September 2004

to appear in
Proceedings of the 2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2005)

[a] IDIAP Research Institute, CH-1920 Martigny, Switzerland
[b] Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny − Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 04-54

A Sector-Based, Frequency-Domain Approach to Detection and Localization of Multiple Speakers

Guillaume Lathoud Mathew Magimai.-Doss

September 2004

to appear in
Proceedings of the 2005 IEEE International Conference on Acoustics, Speech and Signal Processing
(ICASSP 2005)

Abstract. Detection and localization of speakers with microphone arrays is a difficult task due to the wideband nature of speech signals, the large amount of overlaps between speakers in spontaneous conversations, and the presence of noise sources. Many existing audio multi-source localization methods rely on prior knowledge of the sectors containing active sources and/or the number of active sources. This paper proposes sector-based, frequency-domain approaches that address both detection and localization problems by measuring relative phases between microphones. The first approach is similar to delay-sum beamforming. The second approach is novel: it relies on systematic optimization of a centroid in phase space, for each sector. It provides major, systematic improvement over the first approach as well as over previous work. Very good results are obtained on more than one hour of recordings in real meeting room conditions, including cases with up to 3 concurrent speakers.

1 Introduction

Speaker segmentation and tracking are useful in multi-party context such as in meeting rooms, to analyze large amounts of data automatically and to enhance browsing experience. Microphone arrays are useful for such tasks, by providing means for instantaneous detection and localization of multiple concurrent speakers.

A previous work [1] motivated and introduced a sector-based approach for localization of multiple wideband sources in general, and speech sources in particular. To summarize, spontaneous indoor multi-party speech features many overlaps between speakers, as well as reverberations and noise sources such as laptops and projectors. Therefore, detection and localization of multiple concurrent, often wideband sources is needed. The direction taken here is to use Uniform Circular Arrays (UCAs), because their characteristics are almost invariant with respect to direction [2], therefore not depending on particular room dimensions, and imposing no constraint on the location of the source(s).

Existing approaches for microphone array source localization can be divided in two groups: parametric [3] and non-parametric [4]. The review made in [1] suggests that for both groups of methods, there is a need for detection *and* localization of acoustic waves coming from a sector of the space – i.e. “sector-based”, rather than from a specific point or direction – i.e. “point-based”. One sucessful work in this direction is [5]: a coarse-to-fine approach that relies on beamsteering heuristics and prior knowledge of room dimensions, among other things. On the contrary, [1] defined a generic Sector-based Activity Measure (SAM), that relies only on knowledge of the geometry of the microphone array. A time-domain implementation called SAM-PHAT was introduced and tested on real data, achieving correct localization of up to 3 concurrent speakers. However, it did not address the detection issue, due to excessive “leakage”: for a single active source, several sectors would feature a high SAM-PHAT value. In other terms, not only the correct sector would have a high activity value, but also its neighbours (false alarms).

Section 2 briefly summarizes the previously proposed time-domain approach, then proposes two frequency-domain approaches. While the first one is similar to point-based delay-sum beamforming, the second one is novel: it relies on systematic centroid optimization in phase space, *for all points of each sector*. Section 3 describes the meeting room recordings. A preliminary experiment given in Sect. 4 motivates metrics and experiments, which are described in Sect. 5 and 6, respectively. Near optimal results are obtained on hard multisource cases, and applicability to human speech is demonstrated. In particular, the novel approach provides major, systematic improvement over both delay-sum approach and previous work. Section 7 provides a discussion and concludes with future directions.

2 SAM methods

Parametric methods for source localization search physical space for local maximum(s) of a “point-based” measure of activity, which is estimated for each point of space, from the recorded multichannel signals. On the contrary, SAM methods partition the space into sectors, and define an activity measure for each sector.

One time frame of multichannel samples is denoted by vectors $x_1, \ldots, x_m, \ldots, x_M$, with M the number of channels and $x_m \in \mathbb{R}^{N_{samples}}$. The corresponding positive frequency Fourier coefficients representations are denoted by $X_1, \ldots, X_m, \ldots, X_M$, with $X_m \in \mathbb{C}^{N_{\mathrm{bins}}}$ (e.g. obtained by FFT). Each SAM method defines an *activity function* $A_k\left(X_1, \ldots, X_M\right) \in \mathbb{R}$ for each sector $k = 1 \ldots N_{\mathrm{S}}$. The higher this value is, the more likely the sector is to contain at least one active source, which can then be used to take a hard decision, for example by applying a threshold on A_k.

2.1 Partition of the Search Space into Sectors

The search space is partitioned into N_{S} connected volumes (sectors) [1]. For example, the space around a horizontal planar microphone array can be partitioned in "vertical slices": for k=1 . . . N_{S}:

$$\mathbf{S}_k = \left\{ (\mathrm{r}, \mathrm{az}, \mathrm{el}) \in \mathbb{R}^3 \;\middle|\; \begin{array}{l} \mathrm{r} \geq \mathrm{r}_0,\ 2\pi\frac{k-1}{N_{\mathrm{S}}} \leq \mathrm{az} < 2\pi\frac{k}{N_{\mathrm{S}}}, \\ 0 \leq \mathrm{el} \leq \frac{\pi}{2} \end{array} \right\} \tag{1}$$

where r, az, el designate radius, azimuth and elevation w.r.t. the microphone array center; microphones are all in the sphere $r < r_0$.

2.2 SAM-PHAT approach (SP)

SAM-PHAT defines sector activity A_k^{SP} with a 1-dimensional integration. For each microphone pair, the time-domain GCC-PHAT function [6] is summed over a range of time-delays corresponding to the sector. For complete details the reader is invited to refer to [1]. One important point for the following discussion is the implementation of the time-domain GCC-PHAT, for each possible microphone pair $p = 1 \ldots P$:

$$R_{\mathrm{PHAT}}{}^{(p)}(\mu) \triangleq \mathcal{R}e\left[\mathrm{IFFT}\left(\frac{G_p(f)}{|G_p(f)|}\right)\right], \tag{2}$$

where $\mu \in \mathbb{N}$ is a time-delay in samples, $f \in \mathbb{N}$ is a discrete frequency ($1 \leq f \leq N_{\mathrm{bins}}$), $\mathcal{R}e(\cdot)$ denotes real part, and $G^{(p)}(f)$ is the frequency domain cross-correlation for microphone pair p:

$$G^{(p)}(f) \triangleq X_{i_p}(f) \cdot X_{j_p}^*(f), \tag{3}$$

where $(\cdot)^*$ denotes complex conjugate, i_p and j_p are indices of the 2 microphones: $1 \leq i_p < j_p \leq M$. Note that $P = M(M-1)/2$.

2.3 SAM-SPARSE approaches

Previous experiments [1] showed that SP suffers from "leakage", as explained in the Introduction. A possible cause is the non-linearity introduced by Eq. 2. It assumes a single source occupying the entire spectrum, which is rarely the case in practice (noise and/or other speakers). The methods proposed here perform analysis in the frequency domain only. They rely on a sparsity assumption, which is reasonable in speech [7]: within each frequency bin f, only one sector $k_{max}(f)$ is judged as active:

$$k_{max}(f) \triangleq \arg\max_k \left(a_{k,f}\right), \tag{4}$$

where $a_{k,f}\left(X_1(f), \ldots, X_M(f)\right)$ is a *frequency bin* activity function. Next, for each sector the global activity is estimated by counting the number of active frequency bins:

$$A_k^{\mathrm{SAM}}\left(X_1, \ldots, X_M\right) \triangleq \mathrm{card}\left\{f \,\middle|\, k_{max}(f) = k\right\}. \tag{5}$$

Both SAM-SPARSE methods presented below define $a_{k,f}$ as:

$$\begin{aligned} a_{k,f} \quad &\triangleq \quad \mathcal{R}e\left[\textstyle\sum_{p=1}^{P} \frac{G^{(p)}(f)}{|G^{(p)}(f)|} \cdot e^{-j\Phi_{k,f}^{(p)}}\right] \\ &= \quad \textstyle\sum_{p=1}^{P} \cos\left[\angle G^{(p)}(f) - \Phi_{k,f}^{(p)}\right], \end{aligned} \tag{6}$$

where $\angle(\cdot)$ is the argument of a complex number, $\Phi_{k,f}^{(p)}$ is a phase value (angle), fixed for each triplet (k, f, p). $a_{k,f}$ is maximized when several pairs have $\angle G^{(p)}(f)$ close to $\Phi_{k,f}^{(p)}$. The hope is to circumvent spatial aliasing limitations of a single microphone pair, by combining information from

Table 1: Target values of the $P_{N_c \geq 1}$ and $\overline{N_c}$ metrics, as a function of the α statistic and N_a, the number of simultaneously active sources in the annotation.

N_a	$P_{N_c \geq 1}(N_a)$	$\overline{N_c}(N_a)$
1	α	α
2	$1-(1-\alpha)^2$	$2\alpha(1-\alpha)+2\alpha^2$
3	$1-(1-\alpha)^3$	$3\alpha(1-\alpha)^2+6\alpha^2(1-\alpha)+3\alpha^3$

Table 2: Results (in %) of the two working points WP1 and WP2 on the development set. "min" and "mean" figures are computed on utterance recall. "–" denotes that the target is unknown.

	Seq. #1				Seq. #4			
System	FAR	FRR	min	mean	FAR	FRR	min	mean
SP-WP1	1.0	52.7	0.0	47.3	0.1	85.3	0.0	15.6
SSD-WP1	0.4	8.5	0.0	91.5	0.8	50.7	24.3	50.6
SSC-WP1	**0.4**	**2.8**	**15.3**	**97.2**	1.4	41.7	**36.4**	58.6
target	0.5	0	100	100	–	–	–	67.4
SP-WP2	23.6	**0.50**	**90.3**	99.5	**10.8**	**30.3**	35.1	70.9
SSD-WP2	20.5	0.6	83.1	99.4	**33.7**	**15.7**	57.6	84.7
SSC-WP2	**6.8**	0.4	81.9	**99.6**	**14.7**	**21.3**	59.1	79.1
target	0	0.5	100	100	–	–	–	67.4

multiple microphone pairs. This way, not only low frequencies, but also higher frequencies can be used for low resolution ("sector-based") analysis. Each frequency bin is treated independently and phase unwrapping is not needed.

Note that:

$$k_{max}(f) \triangleq \arg\max_k (a_{k,f}) = \arg\min_k (d_{k,f}) \tag{7}$$

with:

$$d_{k,f} \triangleq \sum_{p=1}^{P} \sin^2\left(\frac{\angle G^{(p)}(f) - \Phi_{k,f}^{(p)}}{2}\right). \tag{8}$$

Note that, as SAM-PHAT, the SAM-SPARSE methods rely on phase information only, but not on energy. This choice is inspired by (1) the GCC-PHAT weighting, which is well adapted to reverberant environments, (2) the fact that Interaural Level Difference is known to be much less reliable than time-delays, as far as localization is concerned. As for computational complexity, it must be noted that since each frequency bin is processed independently, the SAM-SPARSE methods can be parallelized in a straightforward manner.

Sections 2.4 and 2.5 propose two definitions of $\Phi_{k,f}^{(p)}$.

2.4 SAM-SPARSE-D (SSD)

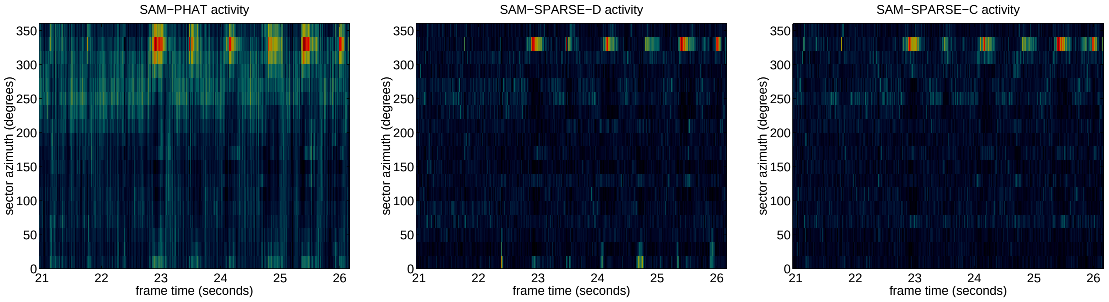
"D" stands for delay-sum. Similarly to classical steered beamformer source localization approaches [3], a linear phase is used:

$$\Phi_{k,f}^{(p)\ \mathrm{SSD}} \triangleq \pi \frac{f}{N_{\mathrm{bins}}} \mu_k^{(p)}, \tag{9}$$

where $\mu_k^{(p)}$ is the time-delay (in samples) associated with the geometric center point of sector k, and microphone pair p.

2.5 SAM-SPARSE-C (SSC)

"C" stands for centroid. We note that SSD defines sector-based activity values based on point-based analysis. In SSC, the linear phase in Eq. 9 becomes an unconstrained phase value $\Phi_{k,f}^{(p)}$, such that $a_{k,f}$



(a) (b) (c)

Figure 1: Activity values A_k for each method, on the beginning of Seq. #4 (development set). Dark colors denote low activity, and light colors denote high activity. The true speaker location is 326.3^o. The subject starts speaking at time 22.7 s and says a few isolated words.

is optimized *for all points in sector* k, not just for its geometric center:

$$\Phi_{k,f}^{(p)\,\mathrm{SSC}} \triangleq \arg \min_{\psi\in[-\pi,+\pi]} \sum_{n=1}^{N} \sin^2\left(\frac{\theta_{k,n}^{(p)}-\psi}{2}\right), \qquad (10)$$

where $\theta_{k,n}^{(p)}$ (n=1...N) are phase values corresponding to microphone pair p and a grid of N points in sector k. This grid of points is defined according to (1) microphone array's geometry, (2) sectors' definition. In experiments below, a circular array is used with radial sectors, as in Eq. 1. Thus, the grid is defined uniformly in spherical coordinates $(\mathrm{r}, \mathrm{az}, \mathrm{el})$: it has constant angular intervals for varying radius.

Note that the sum in Eq. 10 can be written $B\cos(C-\psi)$, where B and C do not depend on ψ. Therefore, numerical search for its unique minimum over $[-\pi,+\pi]$ is simple and fast.

3 Data

Five real 16kHz audio sequences were taken from a meeting room audio-visual corpus available online [8], recorded with a horizontal circular 8-mic array (10 cm radius) set on a table. Complete data and description can be found at `http://mmm.idiap.ch/Lathoud/05-ICASSP` Seq. #1 to #3 were recorded with with either 2 or 3 simultaneously active loudspeakers, at various locations. Seq. #4 has a single human speaker. Seq. #5 has multiple concurrent human speakers. Total duration exceeds 1 hour.

4 Preliminary experiment

This section summarizes conclusions of an experiment made with a point-based source localization approach. The reader can refer to [1] for complete details and to [8] for implementation. The point-based localization was applied on Seq. #4 (single human speaker). It appeared that correct localization is achieved on a proportion of frames annotated as "speech", equal to α=67.4%.

Another conclusion of this experiment is that frame energy does *not* correlate with correct localization. That is why Sect. 6 reports results on *all* frames annotated as "speech". One consequence on real human multispeaker cases (Seq. #5) is that results should be compared to expectations derived from α.

5 Metrics

The major difference with metrics used in [1] is that here a source is considered as correctly detected if it is *exactly* within a sector marked as active in the result. In [1], the source was allowed to be ±5^o outside of an active sector to be considered as detected.

Three types of metrics are used: (1) False Alarm Rate (FAR) and False Rejection Rate (FRR) to evaluate global performance, (2) "frequency distributions" to determine whether a method succeeds to detect and localize concurrent sources, and (3) "utterance recall" to verify practical usability on real human speech.

FAR/FRR: a varying threshold is applied on sector activity A_k, true/false positives/negatives are counted for each threshold, and used to derive FAR, FRR and Equal Error Rate (EER).

Frequency distributions: *On each frame annotated as "speech"*, the number N_c of correctly found sources is counted. Two metrics are derived: $P_{N_c \geq 1}$ the proportion of "speech" frames where at least one source is correctly found, and $\overline{N_c}$ the average number of simultaneous sources correctly found. The target (see Table 1) is based on the α value, which is 0.674 in the case of humans (as justified in Sect. 4), and 1.0 in the case of loudspeakers. It is important to note that this target is exact in the case of loudspeakers, but only approximate in the case of humans.

Utterance recall: final applications of source detection and localization include speech/silence segmentation and/or speaker tracking. Thus, "utterance recall" is defined for *each* utterance, as the proportion of frames where the corresponding source is correctly localized. The average target for "utterance recall" is α.

6 Experiments

The focus is on loudspeaker recordings (Seq. #1 to #3), since their speech/silence annotation is perfect, while Seq. #4 and #5 are used to verify applicability on human speech (approximate target). Time frames were defined as 32 ms long, with 16 ms overlap, and the following parameters were fixed: $N_{samples}$=512, N_{bins}=512 (i.e. zero-padded FFT was used), and N_{S}=18 sectors of 20^o each.

As explained in [9], evaluating FAR/FRR curves on some test data is *not* enough, since in a real system the threshold has to be fixed *before* seeing any of the test data, rather than after. A development set was defined: the first 148 s of Seq. #1 (2 loudspeakers) and 30.3 s of Seq. #4 (single human) to determine two Working Points: (WP1) a conservative threshold T_{WP1} (FAR=0.5%), (WP2) a lower threshold $T_{\mathrm{WP2}} < T_{\mathrm{WP1}}$ (FRR=0.5%). The test set is the rest of the data (more than 1 hour duration).

6.1 Development set

Fig. 1 shows activity values for the 3 methods on human data. SSD and SSC approaches produce much less noise (leakage) than SP, as confirmed by the EER values on the 2-source case (Seq. #1): 9.9% for SP, 3.8% for SSD and 1.6% for SSC.

Next, the thresholds T_{WP1} and T_{WP2} are determined on Seq. #1 to be as close as possible to the desired FAR (resp. FRR). Results are reported in Table 2. WP1 is reasonably close to the desired FAR on both loudspeaker and human data, but FAR and FRR obtained for WP2 vary a lot for different methods, on human speech. Thus, only WP1 is used in experiments on the test set.

6.2 Test set

Table 3a gives detailed results on loudspeaker recordings for all three methods. Overall the SP method fails producing results close to the target in all cases (see FRR metric). The SSC method provides the best results for all metrics, especially on the 3-source cases (see $\overline{N_c}(3)$ metric).

Seq. #2 appeared to be the most difficult for all three methods. This suggests that disparity in received power of the various sources is a major factor of degradation for all methods, although

Table 3: Test set results at Working Point WP1. Values are in %, except for $\overline{N_c}$, which is a number of speakers. Values closest to the target are indicated in bold ("–" means unknown target). "u.r." means utterance recall.

		Seq. #1			Seq. #2			Seq. #3		
Metric	Target	SP	SSD	SSC	SP	SSD	SSC	SP	SSD	SSC
FAR	0.5	0.5	0.5	**0.5**	0.4	0.4	**0.5**	0.3	0.4	**0.6**
FRR	0	82.2	28.5	**12.4**	88.2	29.8	**17.1**	87.0	34.4	**11.2**
$P_{N_c \geq 1}(2)$	100	79.0	**100**	**100**	49.6	**100**	**100**	49.2	99.9	**100**
$\overline{N_c}(2)$	2.0	0.8	1.8	**2.0**	0.5	1.8	**1.9**	0.5	1.7	**1.9**
$P_{N_c \geq 1}(3)$	100	34.6	99.7	**99.8**	19.1	99.7	**99.8**	19.1	98.3	**99.8**
$\overline{N_c}(3)$	3.0	0.4	2.0	**2.5**	0.2	1.9	**2.4**	0.22	1.7	**2.6**

(a) Loudspeakers

Seq.	Metric	Target	SP	SSD	SSC
#4	FAR	–	0.4	1.4	1.6
	FRR	–	70.7	43.8	34.7
	min u.r.	–	3.6	19.4	13.0
	mean u.r.	67.4	31.6	57.2	**65.6**
#5	FAR	–	1.4	3.0	3.0
	FRR	–	72.9	50.5	42.6
	$P_{N_c \geq 1}(2)$	89.4	51.9	80.0	**90.8**
	$\overline{N_c}(2)$	1.3	0.6	1.0	**1.3**
	$P_{N_c \geq 1}(3)$	96.5	51.4	86.7	**95.1**
	$\overline{N_c}(3)$	2.0	0.7	1.4	**1.6**

(b) Human speakers

the SSC results are still close to optimal. For both SSD and SSC, results on Seq. #1 and #3 are comparable. We can therefore expect these methods to cope better with low angular separation of the sources. This has to be confirmed with further experiments.

Results on recordings with real human speakers are given in Table 3b. Although the target values are indicative only (see Sect. 5), it is possible to draw some conclusions. SP obtains very high FRR – more than 70%, which makes it unusable in practice. SSC has results which are the closest to the target for all metrics. FAR is slightly higher than on loudspeaker data, which can be explained by non-speech sounds (body motion, throat noises) on segments annotated as "silence". In particular, SSC achieves much better results on multispeaker cases.

On Seq. #4, the minimum utterance recall of SSC is more than 8 times the FAR, so post-processing is likely to separate the speech from the noise so that each short utterance is detected.

7 Discussion, Conclusion and Future Plans

Frequency-domain analysis provides a major improvement over time-domain analysis: the new SAM-SPARSE methods do not suffer from the leakage problem of SAM-PHAT. The consequence is that SAM-SPARSE methods are able to both detect and locate multiple sources, while SAM-PHAT only provides location information. One cause for this improvement is likely to be the sparsity assumption, which prevents two sectors from being active in the same frequency bin. On the other hand, the sparsity assumption has an intrinsic limitation: a simple experience of thought shows that SAM-SPARSE methods will not, by definition, differentiate between all sectors inactive and all sectors equally active. However, the latter may not be likely to happen in practice.

Overall, we can safely state that SAM-SPARSE methods provide results that are good enough to detect and locate up to 3 concurrent speakers in an indoor environment, which makes them usable in a practical application. Computational complexity was low enough for SAM-SPARSE-D to be implemented in real-time on a single DSP chip by one of us.

The present study also confirms the fundamental difference between point-based analysis and sector-based analysis: SAM-SPARSE-C always performed better than SAM-SPARSE-D. Future directions include investigating more complex activity functions, as well as applicative experiments (e.g. coarse-to-fine search). Also including more speech-specific features may provide additional improvement. We note that more complex SAM-SPARSE methods could still be implemented in real-time: parallelization is straightforward, since they treat each frequency bin separately.

8 Acknowledgments

The authors acknowledge the support of the European Union through the M4 and HOARSE projects. This work was also carried out in the framework of the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM)2. The authors wish to thank Dr I.A. McCowan, Dr D. Barber and O. Masson for fruitful discussions.

References

[1] G. Lathoud and I.A. McCowan. A Sector-Based Approach for Localization of Multiple Speakers with Microphone Arrays. In *Proc. SAPA 2004*, Oct. 2004.

[2] J.J. Fuchs. On the Application of the Global Matched Filter to DOA Estimation with Uniform Circular Arrays. *IEEE Trans. SP*, 49(4), April 2001.

[3] H. Krim and M. Viberg. Two Decades of Array Signal Processing Research: The Parametric Approach. *IEEE SP Mag.*, 13:67 – 94, July 1996.

[4] P. Stoica and R. Mose. *Introduction to Spectral Analysis.* Prentice-Hall, 1997.

[5] R. Duraiswami, D. Zotkin, and L.S. Davis. Active Speech Source Localization by a Dual Coarse-to-Fine Search. In *IEEE Proc. ICASSP*, 2001.

[6] C. Knapp and G. Carter. The Generalized Correlation Method for Estimation of Time Delay. *IEEE Trans. ASSP*, ASSP-24(4):320–327, August 1976.

[7] S.T. Roweis. Factorial Models and Refiltering for Speech Separation and Denoising. In *Proc. Eurospeech*, 2003.

[8] G. Lathoud, J.M. Odobez, and D. Gatica-Perez. AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking. IDIAP-RR 28, IDIAP, 2004.

[9] S. Bengio, M. Keller, and J. Mariéthoz. The Expected Performance Curve. IDIAP-RR 85, IDIAP, 2003.