Evaluation of CASA and BSS models for subband cocktail-party speech separation

Frédéric Berthommier[*] *and Seungjin Choi*[§]

[*]Institut de la Communication Parlée/INPG
46, Av. Félix Viallet, 38031 Grenoble, France
e-mail : bertho@icp.inpg.fr

[§]Department of Computer Science and Engineering
Pohang University of Science and Technology, Korea
e-mail : seungjin@postech.ac.kr

ABSTRACT

For speech segregation, a blind separation model (BSS) is tested together with a CASA model which is based on the localisation cue and the evaluation of the time delay of arrival (TDOA). The test database is composed of 332 binary mixture sentences recorded in stereo with a static set-up. These are truncated at 1 second for the simulations. For applying the two models, we cut the frequency domain in a variable number of subbands, which are processed independently. Then, we evaluate the gain, using reference signals recorded in isolation. Without using this reference, a coherence index is also established for the BSS model, which measures the degree of convergence. After a careful analysis, we find gains of about 1-3dB for the two methods, which are smaller than those published for the same task. The variation of the number of subbands allows an optimisation, and we obtain a significant peak at 4 subbands for the CASA model, and a smaller maximum at 2 subbands for the BSS model.

1. INTRODUCTION

The aim of Blind Source Separation (BSS) is to be data driven and to adapt thanks to a criterion of independence of the different emission processes. This is a reasonable assumption for auditory scene analysis when the different sound sources are not physically coupled, i.e., when each sound is produced by an independent excitation+resonance process. The aim of the computational auditory scene analysis (CASA) is to integrate more abstract levels of description, and hence, to perform an unblind decomposition of the scene. But, the idea is to proceed by steps and to describe intermediate levels of organisation of the scene, without jumping directly to upper levels. In this sense, the speech signal has low-level properties which are fine (harmonicity) and coarse (onset, offset, formant trajectories) spectro-temporal structures, as well as acoustical properties (echos, spatial localisation).

The classical cocktail party paradigm is based on the assumption of independence of emission of different speech signals. Hence, this is not the case if we consider a normal conversation between people because this involves a lot of cognitive aspects and high level representations. But, in the general case, this remains a reasonable assumption, at the signal level, for describing the interference between several speeches. Then, the speech separation task has been adopted early for the application of BSS algorithms [6]. For CASA modelling, the localisation cue is the most simple to exploit for source segregation and cocktail party processing, within the above series of low level speech properties. This is psycho-acoustically relevant for increasing speech intelligibility [4], and CASA assumes, as BSS, that it is involved in a separation process occurring before speech identification. The main difference is the use, in CASA models, of an explicit identification of the spatial location of each source. This can be achieved in short-term frames (tens of ms), whereas the BSS algorithm requires long time frames of about one second to adapt.

We propose a comparison between two algorithms we have previously developed and tested as front-end for robust speech recognition [5][8] on the same database (ST-Numbers95). A novelty is we make a comparable application in subbands. This is in line with the recent development of subband speech recognition for improving robustness [3]. This comparison is significant about similarities and differences between the two approaches. For this evaluation, we take into account and we extent the methodology described in the literature [6][7]. Because we work with only two sources, a static set-up and loudspeakers in a soundproof room, we also bear in mind the limitation of the current solutions in regarding the whole complexity of the speech separation problem [10].

2. THE ST-NB95 DATABASE

The original database Numbers95 is composed of sentences of several words (numbers within a small vocabulary of 32 words). These are pronounced by different speakers and transmitted by telephone. This is dedicated to the development of robust speech recognition algorithms because the speech is somewhat noisy and distorted.

A motivation for making a new recording was to set the background for a close comparison between cocktail-party techniques in which two sources are targets for recognition. The stereo database ST-NB95 was built at ICP from the monophonic NB95 in order **(1)** to spatialise the signal of NB95 in azimuth and **(2)** to mix the signals of NB95 with a relative level controlled well. A minimal distortion of the original signal is introduced during the new recording and recognition tests are feasible without a great specific development. The baseline given by a normal recognition system applied on mixtures is low (about 72% Word Error Rate [5]), and this allows a sensitive measure of any improvement.

The record was carried out in a soundproof an-echoic room by playing and recording the files of NB95 simultaneously with the same computer. The geometry of the set-up is shown Fig.1. The 40cm distance between the microphones has been chosen in order to have a large time difference of arrival (TDOA). Arbitrarily, the source s=1 is the left loudspeaker and has a positive TDOA. This geometry is static for all the records of the database.

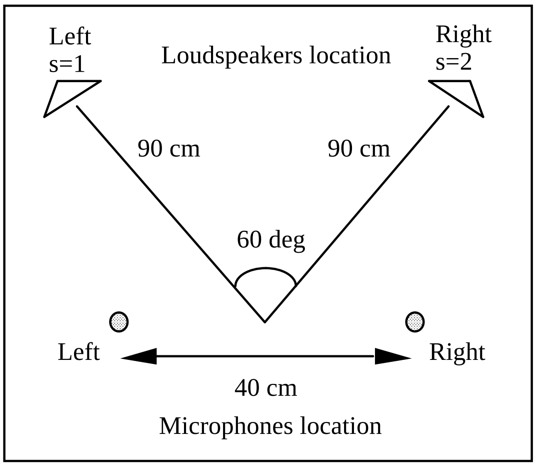

Figure 1: Geometrical set-up for the recording of ST-Numbers95.

The ST-NB95 database is composed of sentences selected from NB95: **(1)** 2*613 sentences are played left or right in isolation in order to have in hands a reference signal **(2)** 613 binary mixtures of the same sentences are arranged for having the highest speech overlap as possible. The signal is resampled at 8kHz. The global relative level between left and right sources is tuned at 0dB separately for each pair of sentences. The isolated records are precisely synchronised with the mixture. This allows estimating the segregation gain by close comparison between the segregated signal and the reference signal recorded in isolation. In the present simulations, 332 mixture sentences have been taken out of the 613 ones. The mean duration of the files is 2 seconds, but the duration has been truncated at 1 second in the present study. Only 52 files have less than one-second duration.

In this paper, we propose a quantification of the gain of the two front-end methods using the reference signals. The front-end principle consists in feeding the recognition system with segregated signals. So, the performance will depend on the degree of cross-talk suppression as well as on the distortion of the processed signal. The different stereo signals used in the simulations and the evaluation will be noted X for the mixture, R for the reference, and Y after processing.

3. THE CASA MODEL

The aim of the CASA model is to perform the segregation according to a primitive feature analysis. Hypothetically, the human auditory system streams the different sound sources thanks to acoustical properties of the signal as harmonicity, localisation, temporal continuity and amplitude co-modulation in different frequency bands. For modelling, these features are extracted by simple signal analysis methods (as the cross-correlation) and the signal is segmented according this estimation. In this vein, we have developed and tested a model similar to the Bodden&Blauert's [2] cocktail party processor and we have incorporated some interesting simplifications on the signal processing point of view. The principle is to weight the spectrogram differently for each source. One characteristic of the CASA model is to exploit the bias of the TDOA estimation relatively to the known TDOA (here, the geometrical set-up is fixed for all the simulations) for the weight estimation. A more detailed description of this model can be found in [8],[9].

3.1 Description of the model

Another characteristic of the CASA model is to operate in large time frequency regions adjusted by two parameters: the number of subbands (nbsb) and the time-frame duration (256 or 512 bins). The filterbank, which produces this decomposition is designed in order to have a unit gain, and to vary the number (and then the size) of subbands in which the process is applied independently. The filters are Bark-scaled and quasi-rectangular (Fig. 2).

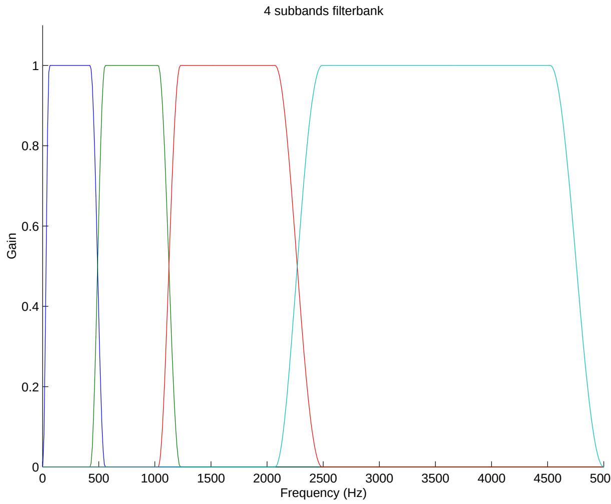

Figure 2: Filterbank design for nbsb=4. The four quasi-rectangular filters F_i i=1..4 are built by grouping and summation of 16 initial hanning and Barkscaled filters.

At first, the spectrogram is computed frame by frame. Here, we vary the time-frame duration. For each time frame of the spectrogram, the spectrum is decomposed into nbsb subbands i:

$$\left|X_i(\omega)\right| = F_i(\omega)\left|X_{\mathrm{left}}(\omega)\right|$$

We have chosen the c="left" input microphone channel for applying the weights. These weights are estimated using the local TDOA observed for each subband and each time-frame. Then, this requires a synchronous decomposition of the "right" input channel. For the TDOA estimation, the subband waves arising from the nbsb filters after inverse FFT are demodulated by half-wave rectification and band-pass filtering in the pitch domain. A cross-correlation is computed between left and right signals and the estimated TDOA is the position of the maximum within an observation window [-10, 10]bin. Then, we use two weighting functions, one for each source, to evaluate the weights.

The weighting functions of the sources are characterised by a symmetric slope (Fig. 3), and their sum is one:

$$W_{s,i}(\mathrm{TDOA}_i) = (1 - W_{s',i}(\mathrm{TDOA}_i))$$

These are adapted to the current geometrical set-up, and to the current number of sources. The maximum (one) is assigned to the TDOA of the target source and the minimum (zero) to the TDOA of the other source (Fig. 3).

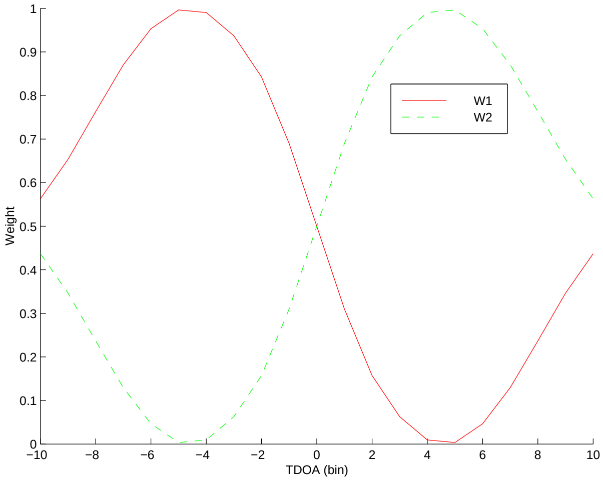

Figure 3: Weighting functions of the two sources s=1 and s'=2 adapted to the current geometrical set-up.

The estimation of the local spectrum of each source s is a product between the weight and the local (left) module spectrum of the mixture data:

$$\left|Y_{s,i}(\omega)\right| = W_{s,i}(\mathrm{TDOA}_i)\left|X_i(\omega)\right|$$

Then, the reconstructed spectrum Y is, for each source s, the sum of the subband contributions. For each time-frame of the spectrogram, we have:

$$\left|Y_s(\omega)\right| = \sum_{i=1}^{nbsb}\left|Y_{s,i}(\omega)\right|$$

Finally, the FFT resolution and the fine details of the spectrogram are preserved because just the amplitude of the signal is modified in large time-frequency regions in order to suppress the competing source by segmentation. The re-synthesis of the temporal signal Y is done by inverse FFT.

3.2 Estimation of the gain

The recording of isolated sentences allows a reference to estimate the accuracy of the reconstruction of each source s. Following [11], we define a distance: the Reconstruction Accuracy (RA) measure. We fix the time frame duration analysis at 1024 bins. We make for each frame a comparison between the full-band spectra of the reference R and of the product of segregation Y. All these spectra are pre-normalised for avoiding global amplitude differences:

$$\mathrm{RA}(R_{s,\mathrm{left}}, Y_s) = 10\log\frac{\int_\Omega \left|R_{s,\mathrm{left}}(\omega)\right|^2}{\int_\Omega (\left|R_{s,\mathrm{left}}(\omega)\right| - \left|Y_s(\omega)\right|)^2}$$

$$\text{where } \Omega/2\pi = [100,4000]\mathrm{Hz}$$

A statistic of RA is established for all 1024 bins time frames (silence included) of the 332 sentence pairs truncated at 1s.

The effect of the two factors: **(1)** the number of subbands nbsb, varied from 1 to 5, and **(2)** the length of the processing window, 256 and 512 bins, is shown in Figure 4. We observe a significant maximum at nbsb=4 and 256 bins. Each source is reconstructed according to an estimation of the TDOA having an accuracy that also depends on these two parameters. Then, the maximum observed for nbsb=4 is due to the trade-off between the accuracy of the TDOA estimation (which decreases when the bandwidth decreases) and the accuracy of the spectrogram segmentation which increases when the bandwidth decreases.

For taking into account the initial RA of the mixture X, we subtract it from the RA of Y to obtain the effective gain for each source:

$$\mathrm{Gain}_s = \mathrm{RA}(R_{s,\mathrm{left}}, Y_s) - \mathrm{RA}(R_{s,\mathrm{left}}, X_{\mathrm{left}})$$

A statistic of the gain for all time frames of the same sentences shows a high correlation between the gain values obtained for the two sources (Fig. 5).

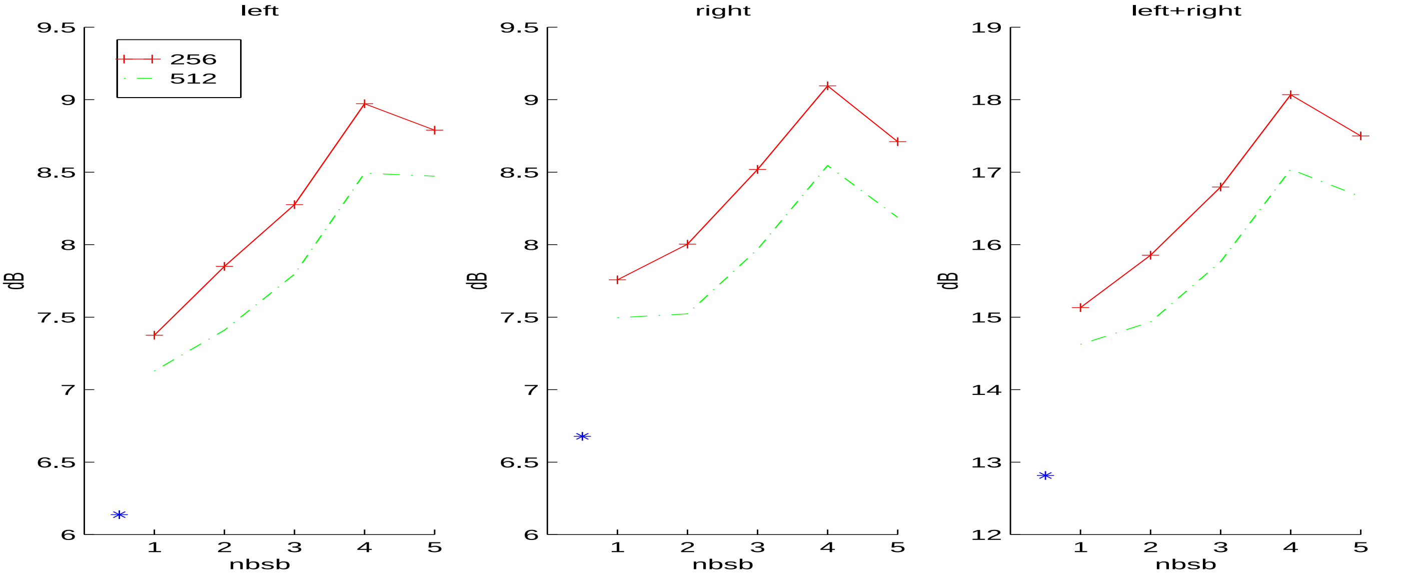

Figure 4: Effect of the number of subbands (nbsb) for the CASA model on the RA (in dB). From left to right: averaged left source RA, averaged right source RA, averaged left+right RA over all frames. The number of subbands varies from 1 to 5 and the two curves correspond to duration= 256 and 512 bins. The RA of the mixture, which is subtracted for gain evaluation is labelled (*).

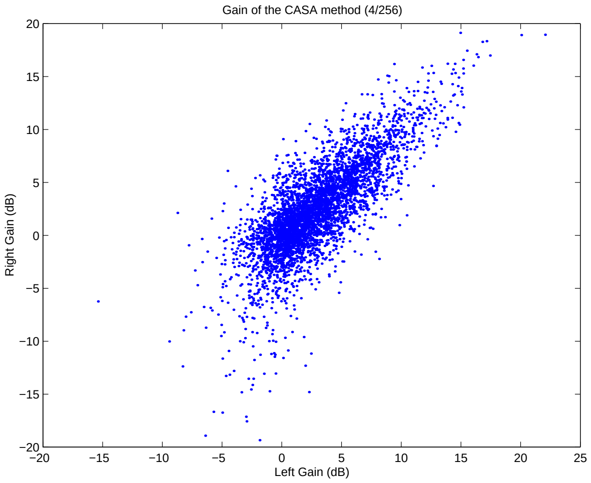

Figure 5: Gain in dB of the CASA method (CASA 4) for the left (s=1) and the right (s=2) sources (nbsb=4, duration=256bins). Each point is for a frame of 1024 bins.

4. THE BSS MODEL

The BSS model we use is standard and it is based on the main assumption that the sources are independent. A dual assumption, which motivated the extension with delay lines of the initial Hérault&Jutten's model, is that each source is temporally correlated [1]. This is the case for the speech signal, mainly because the glottal source is regular and periodic. This led to a recurrent implementation of the BSS, in which the goal is to adapt vectors of weights W having several hundreds of taps. Moreover, the use of delay lines allows the implicit processing of the TDOA when the sources are located in space, and finally, this model is powerful to tackle the current cocktail party task.

4.1 Algorithm

The separation process is temporal, and we have developed a subband version of the algorithm presented in [5]. We apply the same algorithm independently on the nbsb waves obtained after filtering the signal with the filterbank previously defined (Fig. 2). For each sample t of the input signal (which is truncated at 1 s):

$$Y_c^{(i)}(t) = X_c^{(i)}(t) + \sum_{p=0}^{L} W_{cc',p}^{(i)} Y_{c'}^{(i)}(t-p)$$

where c and c' are the two (left and right) channels and (i) the subband. For multiple nbp passes, t is simply reset at the beginning of the signal without re-initialisation. The variation of the weight vectors is given by:

$$\Delta W_{cc',p}^{(i)}(t) = -\eta \lfloor 1 - W_{cc',0}^{(i)}(t) \rfloor \mathit{sign}((Y_c^{(i)}(t))\, Y_{c'}^{(i)}(t-p)$$

p = 0…L and the length L+1 of W is 200 taps
The learning rate is fixed at $\eta=10^{-7}$
The demixing filters W are initialised at 0
These are not re-initialised between each pass

Because the filterbank is unity gain, the output of these nbsb processes are simply added to obtain Yc and Yc'.

4.2 Evaluation of the BSS segregation

The computation of RA and gain for the BSS method does not differ from this defined for the CASA method, excepted the use of the right input channel reference for the right source s=2. In our simulations, there is no permutation problem and the left source arises in the left channel and respectively for the right source:

$$\mathrm{Gain}_s = \mathrm{RA}(R_{s,c}, Y_c) - \mathrm{RA}(R_{s,c}, X_c)$$

$$\mathrm{Gain}_{s'} = \mathrm{RA}(R_{s',c'}, Y_{c'}) - \mathrm{RA}(R_{s',c'}, X_{c'})$$

with c=left, s=1 and c'=right, s'=2

The effect on the RA of the two factors **(1)** the number of subbands nbsb, varied from 1 to 4, and **(2)** the number of passes, nbp=2,3,10, is shown in Figure 7 (for the same previous 332 sentence pairs). The frame distribution of the gain is shown Fig. 6. We add two conditions to this set of simulations: **(1)** the "BSS given" is the one pass application without convergence of the two averaged demixing filters W obtained after simulation of the nbsb=1, nbp=10 (332 vectors) **(2)** the "BSS ori" condition uses the selection of the same data set extracted from the output data of our previous study [5].

Additionally, a coherence index between signal pairs is calculated. Interestingly, this does not require the presence of a reference signal R as for the RA index. A first step consists in evaluating the coherence values (varying between 0 and 1 for each frequency bin) for consecutive [n,n+1] windows of 256 bins overlapping by half:

$$\mathrm{Coh}(Y_c, Y_{c'}, n) = \frac{\left|\sum_{n,n+1} Y_c(\omega) Y^*_{c'}(\omega)\right|^2}{\sum_{n,n+1} \left|Y_c(\omega)\right|^2 \sum_{n,n+1} \left|Y_{c'}(\omega)\right|^2}$$

Then, the coherence spectrogram is averaged over the sentence duration, and this mean value also varies between 0 and 1. For the interpretation of the variations of Coh(Y_c,$Y_{c'}$) observed after applying the BSS method, we need two reference values:

(1) CohX the coherence between the mixture channels Coh(X_c,$X_{c'}$) which is the initial value.

(2) CohR the coherence between the reference signals Coh($R_{s,c}$,$R_{s',c'}$) which is a floor value, at about 0.5 for this modality of computation.

The result of coherence calculation is shown figure 7 (right). In all cases, we observe a steep variation after applying the BSS method (compare to CohX), and the index stays above CohR after nbp=10 passes. There is a small decrease between nbp=2 and nbp=3 and no difference between nbp=3 and 10. The index Coh decreases monotonically when nbsb increases. The correlation with the RA is small and only for nbp. So, we conclude this index is useful for testing the convergence of the algorithm, but not for evaluating the quality of the separation. Remarkably, it varies only for the BSS method. The CASA method does not modify the fine spectral structure of the signal, and consequently, the value stays at CohX. The decreasing of the coherence index depends on the specific criterion of separation of the BSS method (i.e., the independence of outputs). Furthermore, in the "given" condition, the convergence is blocked and the algorithm cannot adapt to the input signal. The value of Coh is intermediate. This confirms that the coherent index is significant about the criterion of the BSS that is partially fulfilled in this case.

We retrieve the known variation of the gain with the number of passes nbp (Fig. 7). A good convergence is obtained after nbp=3. We observe a small effect of the application in subbands with a maximum at nbsb=2, but this is less pronounced than for the CASA method. We remarked that the demixing filters obtained for each subbands are bandpass filtered versions of the fullband demixing filter (not shown). The frequency bands of these filters are well related to the filterbank allowing the decomposition (Fig. 2 for nbsb=4). This suggests that the convergence is facilitated in subbands because the complexity is decreased. But this has not significant effect on the gain. Another observation is that the de-mixing filters W have a minimum at 5 bin, corresponding to the absolute value of the TDOA of the sources. Finally, the gains for left and right sources are also correlated above 0dB (Fig. 6) as for the CASA method (Fig. 5), and there is an interesting (not interpreted) flooring effect of the gain of one source around 0dB.

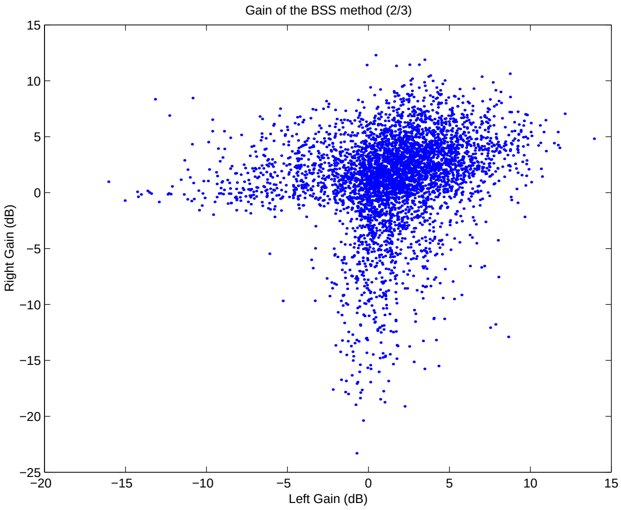

Figure 6: Gain of the BSS 2/3 model (nbsb=2, nbp=3). Each point is for a frame of 1024 bins.

5. DISCUSSION AND CONCLUSION

We have summarised the main results of this study in Table 1. When we correct the RA to evaluate the effective gain, the results are less optimistic than those published in the literature (about 10 dB in the same condition are reported in [6]). A part of this difference is due to specific factors of our study (e.g., our study includes silence). But, we obtained a good improvement of speech recognition with the same "BSS ori" output data [5] and this outperformed a CASA method similar to CASA 4 [8]. A gain of about 2 dB is not negligible for improving recognition scores.

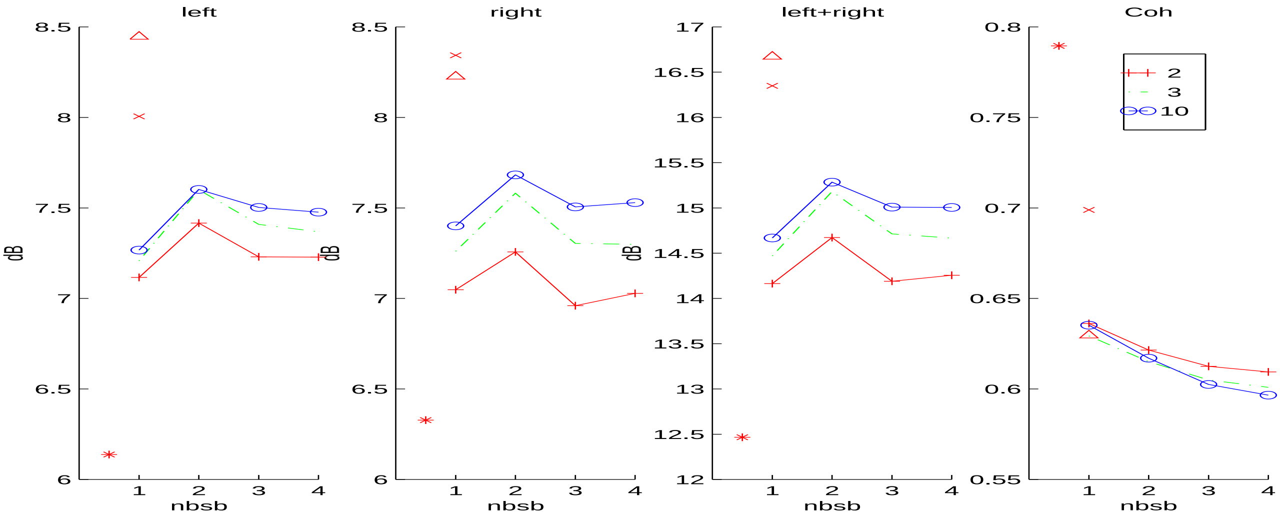

Figure 7: Effect of the number of subbands (nbsb) for the BSS model on the RA (in dB). From left to right: averaged left source RA, averaged right source RA, averaged left+right RA over all frames, coherence. The number of subbands varies from 1 to 4 and the three curves correspond to nbp= 2,3,10. The RA of the mixture, which is subtracted for gain evaluation is labelled (*). The CohX coherence between the two mixture channels is labelled (*) in the right figure. In each figures, two points are added at nbsb=1 for the "BSS giv" condition (×) and for "BSS ori" data (Δ).

dB	**RA 1+2**	**RA 1**	**RA 2**	**Gain 1**	**Gain 2**
CASA 4	18.06	8.97	9.09	2.83	2.41
BSS 2/3	15.18	7.60	7.58	1.46	1.25
BSS ori	16.67	8.44	8.22	2.29	2.13
BSS giv	16.34	8.0	8.34	1.86	2.01

Table 1: Summary of the main results expressed in dB. This is an average over all frames.

Now, the CASA method is found to have a gain higher than the BSS method. Moreover, in the present simulation, we do not retrieve the same gain for the BSS method, as observed with "BSS ori" data (Table 1). This could be due to the truncation at 1 second of the sentences, and also to some differences in the implementation and in the condition of application of the algorithm (e.g., L is at 200 instead of 250). Remarkably, the "given" condition ("BSS giv" in table 1) allows a higher gain than the adaptive condition BSS 2/3. The two filters W are specific for the static mixing condition and not for a particular signal pair. We conclude that when this condition is fixed, better is to use the average demixing filters for making a simple cancellation.

These results confirm that the judgement about the quality of a method is not easy, and that a methodology of comparison has to incorporate different indexes applied at different levels. This could include a perceptual study for appreciation of the degree of distortion because, consistently with our previous recognition results, the artefacts produced by the segregation are not the same for the two methods: BSS allows less distortion, but this is subjective.

We have shown that the two domains CASA and BSS/ICA are close together because they share common questioning about the biological modelling, the development of useful applications as well as common paradigms as the cocktail-party problem. One promising compromise is to consider these are complementary approaches for modelling the functioning of sensory pathways with a switch between fix (signal processing) and adaptive (learning) modes.

Acknowledgements: The French-Korean collaboration is supported by ARIEL-KOSEF and this work is a part of the EEC RESPITE project.

REFERENCES

[1] Amari, S. (1999) ICA of temporally correlated signals – learning algorithm, in proc. of ICA'99, Aussois, pp. 239-244.

[2] Bodden, M. (1993) Modeling human sound-source localization and the cocktail-party-effect, Acta Acustica, vol. 1 (1), p. 43-55.

[3] Bourlard, H. & Dupont, S. (1997) Subband-based speech recognition, in proc. of ICASSP'97, pp. 1251-1254.

[4] Bronkhorst, A. (2000) The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker condition, Acustica, 86:117-128.

[5] Choi, S., Hong, H., Glotin, H. & Berthommier, F. (2000) Multichannel signal separation for cocktail party speech recognition: a dynamic recurrent network, in proc. of ICSLP-2000, Beijing.

[6] Deville, Y. (2001) Applications of blind source separation and independent component analysis methods, in proc. "*De la séparation de sources à l'analyse en composantes indépendantes*", C. Jutten et al. (Eds), Villard de Lans, pp. 177-212.

[7] Schobben, D., Torkkola, K. & Smaragdis, P. (1999) Evaluation of blind separation methods, in proc. of ICA'99, Aussois, pp. 261-266.

[8] Tessier, E., Berthommier, F., Glotin, H. & Choi, S. (1999) A CASA front-end using the localisation cue for segregation and then cocktail-party speech recognition, in proc. ICSP'99, Seoul, Korea, pp. 97-102.

[9] Tessier, E. & Berthommier, F. (2001) Speech enhancement and segregation based on the localisation cue for cocktail-party processing , to appear in proc. of CRAC workshop, Aalborg.

[10] Torkkola, K. (1999) Blind separation for audio signals – are we there yet ?, in proc. of ICA'99, Aussois, pp. 239-244.

[11] Yang, X., Wang, K., Shamma, S., A. (1992) Auditory representations of speech signals, IEEE Trans. on Inf. Theory, 38:2:824-839.