Several improvements of the Hérault-Jutten model for speech segregation

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

We have adopted the original Hérault-Jutten model for the segregation of two concurrent voices in realistic conditions. Firstly, with the new Daimler-Chrysler in-car database, we confirm our previous results obtained with a similar model, and how to make good use of the algorithm. Secondly, we create different mixture conditions from the DC database, and we test 3 kinds of improvement. (1) With the introduction of non-causal FIR filters, the H-J model is able to process asymmetric configurations of two speakers. (2) The benefit of the frequency decomposition is well retrieved with four subbands. Two compatible methods for completing the segregation by post-processing (3) are described: The amplification of the input/output relative level gain by joint-enhancement and the beamforming. We show a great overall reduction of the cross-talk, even in noise, which is promising for speech recognition applications.

1. INTRODUCTION

The original Hérault-Jutten (H-J) model [8,9] is the precursor of a wide family of BSS and ICA models, but this is not recognised to well support realistic audio applications. The first reason is this was initially adapted for the segregation of instantaneous additive mixtures. As main characteristics early established from biological considerations [8], the H-J algorithm **(1)** is a recurrent network, **(2)** it operates in the temporal domain, and **(3)** it is adaptive, thanks to the following updating rule of the scalar weight pairs:

$$\Delta W_{cc'}(t) = -\eta\, f(Y_c(t))\, g(Y_{c'}(t))$$

in which f(.) and g(.) are functions chosen for having an independence criterion more powerful than output de-correlation. Remarkably, this was the root of the ICA concept. For BSS of spatially localised audio sources [14], as well as for convolved mixtures [12,13], the weighting has been extended in the temporal domain by adding several taps. After extension, the estimation of the unknown mixture weights distributed in the temporal domain, and forming a FIR filter vector, follows the same rule:

$$\Delta W_{cc',p}(t) = -\eta\, f(Y_c(t))\, g(Y_{c'}(t-p)), \text{ with } p \in [0,L]$$

Then, within the set of existing BSS and ICA applications, we can integrate the algorithms having these three main characteristics in the vein of the H-J family, in contrast with other methods (reviewed in [17]). Many demonstrations of the ability of these adaptations of H-J to segregate audio sources, speech or music have been carried out, with a success depending on the difficulty of the task [6]. For superimposed (additive) and delayed mixtures, there is a considerable gain, but this declines when other factors are involved, as an echoic environment, the noise, the movement of the sources, the use of loudspeakers or not, etc... Moreover, there is a gain estimation problem arising when there is no reference recorded in isolation, i.e., when loudspeakers are not used, and the lack of reliable quantification of the performance does not favour the development of demonstrations in realistic conditions. As far as we know, the gain obtained with the H-J model was quite small in realistic environments as in a running car, in which it would be greatly beneficial to pre-process the speech before automatic recognition, in order to remove noise or, better, to segregate concurrent speeches because this is major source or errors.

Following an empirical point of view, the idea is to adapt the H-J model for realistic audio applications, assuming that the three original characteristics taken together are operant for a useful speech processing. On a similar basis, we presented in [5] a successful experiment of cocktail-party speech recognition, in which we have shown a reduction of the recognition error rate after BSS for a static and symmetric setting, and with loudspeakers. In the same way, we will propose several cumulative improvements of the H-J algorithm, each allowing a modest gain, instead of a unique solution. Then, two post-processing solutions illustrate a manner to introduce some a priori knowledge. Hence, the time-extended H-J model is slow and the price to pay for these improvements is an increase of the computational load, but the running of the basic H-J unit we will define is tractable on a PC, and the increment of the complexity is always linear. The database we use for our simulations is well adapted to this goal, because this is recorded inside a car with real speakers who pronounced short repetitive sentences. Each improvement is illustrated by a simulation involving a different mixture condition, and we focus on the concurrent speech segregation task. We also propose a new gain measure working without reference, which is complementary to the spectral gain measure we used in previous studies [2, 16].

2. DATABASE AND METHODS

2.1 The Daimler-Chrysler database

The database [11] was provided to us by courtesy of Daimler-Chrysler in the context of the EC RESPITE project, which focused on the improvement of robust speech recognition. The DC database was precisely designed for studying in-car speech enhancement and segregation, with or without presence of car noise. Comparatively to other resources, the main advantage of this database is to offer a realistic setting, and at the same time to be not too difficult, partly because the in-car environment is weakly echoic. One, two or three speakers (driver, co-driver, passenger) repeated a sequence of German sentences, in silence, or

when the car is running at 120Km/h. They had no precise instruction to not move the head. In the full content, there are two available configurations of microphones: 4 microphones close to each speaker (for details, see [11], Part I), or an array of 6 microphones (Part II). Because the later section is more compatible with the standard BSS task, we have selected in the Part II the recordings of two long sentences of about 30 seconds recorded at 24 KHz we have re-sampled at 8 KHz for most simulations.

We combine these recordings in order to build four conditions of mixing: **(1)** Delayed Mixtures (**DM**), using additive and delayed mixtures of isolated speaker mono recordings **(2)** Convolved Mixtures (**CM**), in which the same mono recordings are convolved and mixed using the HRTF (Head Related Transfer Function, refer to [7]) **(3)** Stereo Mixtures (**SM**) by adding without delay stereo recordings of isolated speakers, in which the echoic structure is preserved **(4)** Real Mixtures (**RM**) of simultaneous speakers, without change of the original data of part II. These four conditions allow a control of the three main factors; delays, convolution, echoes; involved in the mixing process, but there is a compromise. The condition **(1)** is flexible and allows the choice of the delays as well as of the global input relative levels, but it is artificial and the easier to process, whereas **(4)** includes all main factors but the parameters are fixed by the recording and the reference signals are not available.

2.2 Estimation of the spectral gain

For the conditions (1-3), the mixing process is controlled, and it involves the reference signals themselves. As in [2, 16], we define a spectral distance using these references: the Reconstruction Accuracy (RA) measure. We fix the time-frame duration analysis at 1024 samples. A full-band spectral distance is calculated between the reference source R and a signal Y, this for each source/channel pair. All these spectra are normalised at 1 for removing the effect of global amplitude differences:

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

where $\Omega/2\pi = [0,4000]$Hz

The evaluation of the RA is an intermediate step, and the gain obtained for each source/channel pair is the difference between the output RA (Y is an output) and the input RA (X is an input):

$$Gain_{s,c} = RA(R_{s,c},Y_c) - RA(R_{s,c},X_c)$$

This spectral gain estimate removes the overall amplitude gain, thanks to the normalisation, and it is sensitive to the spectral distortions, so this is less optimistic than the similar measures commonly adopted (see [15] for a review). However, this is not sensitive to the small delays. We assume that this measure matches the audible quality of the segregation.

2.3 Estimation of the relative level gain

To complement this, we introduce a relative level gain measure. This directly evaluates the amplitude of the cross-talk reduction produced by the H-J unit. Here, the reference signal is not required. Practically, for two channels, we fix the frame duration at 1024 samples and we evaluate the relative RMS level in dB between both input (RLin) and both output channels (RLout). The input/output relationship (Figure 1) is significant about the performance of the segregation process. Then, the relative level gain can be quantified (Table 1) by the averaged input/output difference (in dB):

$$RLgain = |RLout| - |RLin|$$

as well as by the slope of the linear fit of this distribution (L slope). Moreover, as shown Figure 1, the gain is higher when the absolute input relative level (RLin) is high, and this non-linear relationship can be empirically fitted by a tangent function, and then quantified by the T slope. We also observe that the spectral and the relative level gains have an inverse variation according to the input relative level (see [2]). The T slope of the non-linear tan(.) fit is a more invariant measure, because this is not biased by the distribution of input relative levels, which depends on the particular distribution of overlap between the two sources.

3. RUN TIME PROCESSING

The so-called 'H-J unit' has two channels, it operates in the temporal domain and it has a feedback architecture:

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

where, X are inputs, Y are outputs, and c,c' are the two channels (also noted 1,2 in Figure 2).

In the original simulations of the H-J model [8,9], the weights were updated online according a learning rate η, and the choice of the two functions f(.) and g(.) was critical for controlling the separation criterion. It is admitted that g(.) can be the identity function and f(.) a non linear function depending on the statistics of the sources [1]. Moreover, it has been shown in [4] that, for speech sources, the sign(.) function is optimal for f(.). Consistently, in our simulations, the sign(.) function also appears to be more efficient than the smoother arctan(.) function. Then, the updating step of the H-J unit is:

$$\Delta W_{cc',p}(t) = -\eta\, sign(Y_c(t))\, Y_{c'}(t-p)$$

Practically, the choice of a small value for the learning rate η is important for getting a good convergence, but this requires a rather long adaptation period. To compensate, the block computation method consists in applying a number of sweeps over the same part of the signal. A great advantage of the block computation is the absence of adaptation period. For this, t is simply reset at the beginning of the block, without other re-initialisation. So, we assume this is computationally intensive because we add a loop of sweeps, but, on the other hand, the choice of sign(.) and identity functions allows a reduction of the updating step to a sign evaluation, an addition and a multiplication by a constant.

In order to have a run time processing of the signal, we overlap and add the successive blocks. The input signal is divided in long rectangular time frames, and the output is windowed for smoothing the discontinuities. The window length could be minimised to cope with dynamic mixture conditions, but this degrades the performances. The algorithm inherently requires long windows, partly because the high number of parameters to

estimate. We have fixed this duration at 1 second, in accordance with our previous study, and because the mixture conditions are quasi-static. However, with this intermediate duration, the problem of slowly moving speakers can also be addressed more explicitly in the future.

The shape of the window is manipulated in order to minimise the overlap and to save computation: this is 75% flat and 25% hanning, to be quasi-rectangular. The de-mixing filters W could also be saved for initialising the next block estimation. However, for preserving independent statistics in successive blocks better for the evaluation of the model, we reset the filters. This generates a time series of independent estimates of the W filters, which can be averaged, or analysed separately. We also can extract features from these estimates, as the time delay of arrival (TDOA), which are useful for the post-processing, as we will see later.

The common settings of the simulations are:

- The learning rate is fixed at $\eta=5.10^{-4}$
- The filters W are initialised at 0 for each block of 1s
- These are not re-initialised between each pass
- The number of sweeps is limited at 40
- p = 0…L and the length L+1 of W depends on the mixing condition

As shown in Figure 1 and Table 1, the segregation of two voices (driver and co-driver, symmetrically placed) by the H-J unit in the RM condition (inter mic. dist.=24cm) is satisfactory, and this validates this first kind of improvement, which consists in tuning the parameters of the model.

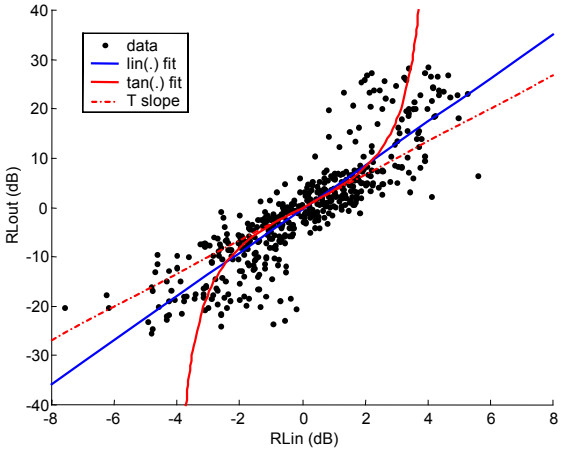

Figure 1: Simulation and evaluation of the H-J unit (L=201), in the RM condition and with two speakers in silence (Clean).

Input	Output	RLin	RLgain	L slope	T slope
Clean	1|2	1.82	6.93	4.43	3.35
Noisy	1|2	1.50	2.75	2.60	2.61

Table 1: Quantification of the H-J unit performance, in silence (Clean) and in car noise (Noisy).

4. USE OF NON-CAUSAL FILTERS

The previous simulation illustrates the favourable symmetric case, and this form of the H-J unit fails to segregate two speeches when the two speakers are located on the same side. Let analyse the reason of this failure. Following [12], the block diagram of the asymmetric case for delayed mixtures (DM) has two branches (Figure 2).

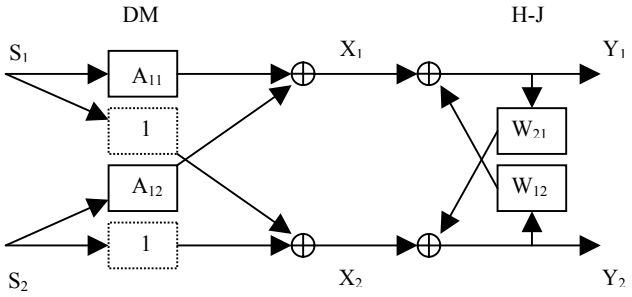

Figure 2: Asymmetric setting in the DM condition.

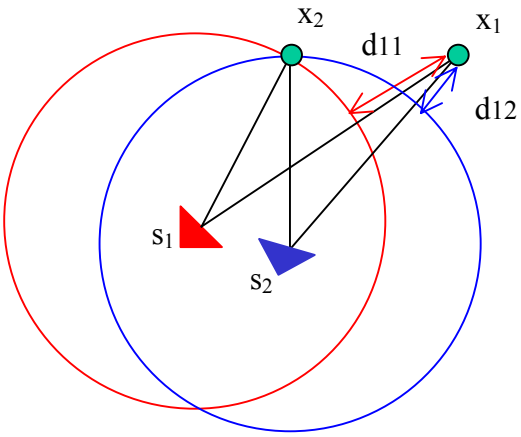

Figure 3: Geometry of the asymmetric setting.

The equations in z of the DM mixing condition are:

$$\begin{cases} X_1(z)=A_{11}(z)S_1(z)+A_{12}(z)S_2(z) \\ X_2(z)=S_1(z)+S_2(z) \end{cases}$$

Then, the de-mixing system to evaluate is:

$$\begin{cases} Y_1(z)=X_1(z)+W_{12}(z)Y_2(z) \\ Y_2(z)=X_2(z)+W_{21}(z)Y_1(z) \end{cases}$$

This system has a solution when the denominator is non zero (a zero denominator might occur only in rare, maybe unrealistic cases [13]):

$$\begin{cases} Y_1(z)=\dfrac{(A_{11}(z)+W_{12}(z))S_1(z)+(A_{12}(z)+W_{12}(z))S_2(z)}{1-W_{12}(z)W_{21}(z)} \\ Y_2(z)=\dfrac{(1+W_{21}(z)A_{11}(z))S_1(z)+(1+W_{21}(z)A_{12}(z))S_2(z)}{1-W_{12}(z)W_{21}(z)} \end{cases}$$

With $\begin{cases} W_{12}(z) = -A_{12}(z) \\ W_{21}(z) = -A_{11}(z)^{-1} \end{cases}$, we have $\begin{cases} Y_1(z) = A_{11}(z) S_1(z) \\ Y_2(z) = S_2(z) \end{cases}$

In this solution, one of the output signals remains a filtered version of the input signal, so separation and de-convolution must be distinguished, and the system is not able to de-convolve the sources. We also have a dual, permuted, solution. Finally, one of the two filters is the inverse of one of the input mixing filters.

This led Nguyen Thi and Jutten [12] to consider that the most general solution of the asymmetric case is the evaluation of IIR filters in B(z)/(1+C(z)). To propose a simpler design, we remark that, under the DM condition (in which we omit here the relative amplitude), one of the two delays is compensated by a non causal shift of one of the de-mixing FIR filters (here, W_{21}):

$$\begin{cases} A_{12}(z) = z^{-d_{12}} \\ A_{11}(z) = z^{-d_{11}} \end{cases} \Rightarrow \begin{cases} W_{12}(z) = -z^{-d_{12}} \\ W_{21}(z) = -z^{+d_{11}} \end{cases}$$

Then, a H-J unit having the same structure and following the same convergence rule is suitable, but we have to incorporate at least one non causal FIR filter (in practice, two) instead of causal ones. Here, when the non causality is only related to the delays of arrival of the sound to microphones, the solution is trivial. More generally, the convolutive effects have inherent causal characteristics, but, when they are associated with delays in the non symmetrical case, these are shifted and become partly non causal. Then, we assume, as Lee et al. [10] that the denominator is generally invertible, and could be expressed by non causal FIR terms instead of IIR terms, which are identified separately in [12]. We adapt this property for the H-J unit, i.e., in the temporal domain, by centring of the L taps:

$$Y_c(t) = X_c(t) + \sum_{p=-L/2}^{L/2} W_{cc',p} Y_{c'}(t-p)$$

One main interest of having non causal FIR filters is to estimate all the terms jointly, and then to preserve the homogeneity of their computation. This is expected to be easier to achieve and more robust.

At first, we compare non causal and causal FIR designs in an asymmetric CM condition, and to complete the demonstration, the other simulations are based on non causal FIR filters. All the parameters remain the same and the evaluation of W follows the same updating rule, with $p \in [-L/2, L/2]$. In the CM condition, each mono signal of a single speaker recording is convolved at 24 KHz with the two 128 samples impulse functions (left and right) of the dummy-head HRTF measured by [7], at a chosen azimuth location (and 0° of elevation), and the two stereo signals are added together (see [15] for details). This convolution generates the delays as well as the intensity differences simulated in the DM condition, and the input signal is at 24KHz for this simulation. We show that, for speakers located on the same side (azimuth at 40° and 20°), the non causal H-J unit converges well (Figure 4) whereas the causal one fails to segregate the speeches. As expected, we observe Figure 5 that one of the de-mixing filters (W_{21}) is not identified.

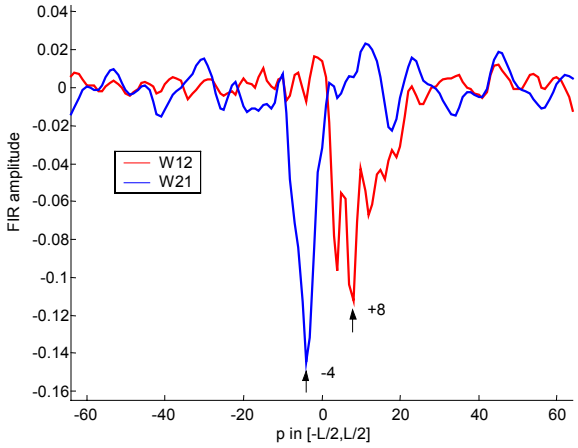

Figure 4: Averaged non causal de-mixing FIR filters (L=129) obtained in the CM condition. Gain1=4.05 dB; Gain2=4.22 dB.

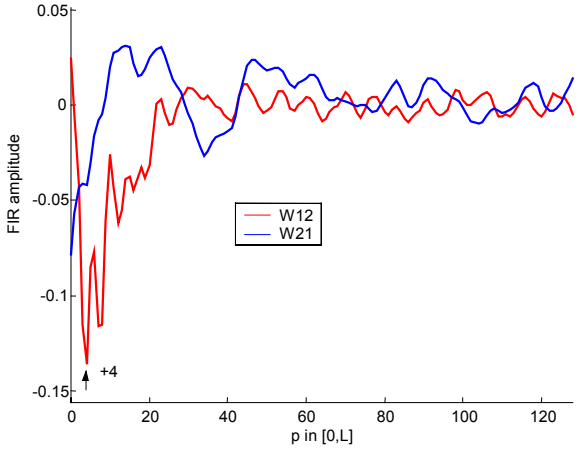

Figure 5: Averaged causal de-mixing FIR filters (L=129) obtained in the CM condition. Gain1=0.64 dB; Gain2=-4.29 dB.

5. SUBBAND PROCESSING

The H-J separation process is temporal and full-band, whereas other BSS methods work in the frequency domain [17]. One way to combine these two approaches is to introduce some frequency decomposition. We feed N H-J sub-units independently with 2N temporal waves obtained after filtering of the two input signals with a Barkscaled filterbank:

$$Y_c^{(i)}(t) = X_c^{(i)}(t) + \sum_{p=-L/2}^{L/2} W_{cc',p}^{(i)} Y_{c'}^{(i)}(t-p)$$

where (i) is the subband index. Then, the computational complexity is multiple of N (Table 2). The FFT-based filterbank groups the FFT bins in wide and quasi-rectangular subbands, along a perceptual frequency scale. This division of the frequency domain is related to the informational content. For speech and with four bandpass filters, one subband almost carries a formant trajectory, and [3] have shown that independent recognition

processes can be applied in each subband and their outputs fused. Since this filterbank is unity gain and composed of N quasi-rectangular filters, the outputs of these N processes are simply added to generate Yc and Yc'. For a small N, the H-J segregation process remains wideband and temporal. Thus, there is no permutation problem arising for the case of two sources having a similar spectral distribution, as two speeches, and this main drawback of the frequency approach is prevented. In a previous study using loudspeakers and real mixtures [2], we found a small peak of gain for N=2. We retrieve this result in the DM condition and with non causal filters, after varying N from 1 (i.e., the algorithm is applied in fullband) to 8. We observe a small maximum of spectral gain for N=4 (Table 2).

input	N=1	N=2	N=4	N=8
CPU (s)	6	12	24	50
Gain1	9.15	10.06	10.30	9.08
Gain2	9.85	10.34	10.57	9.12
RLin	1.22	1.32	1.39	1.42
RLgain	7.79	6.55	6.32	5.67
L slope	7.84	6.21	5.76	5.07
T slope	4.16	2.42	1.47	1.35

Table 2: Gain evaluation for the two sources according to the number of subbands N, in an asymmetric DM condition (d_{11}=6,d_{12}=1, intensity difference=2dB), and for non causal H-J units (L=21, because there is no echoes).

In order to explain this maximum of spectral gain, we suppose the existence of a trade-off between two effects, which have inverse tendencies, as for the CASA model [2]. As shown Table 2 with indexes calculated per subband, the input local relative level (RLin) increases with N, whereas the RLgain, as well as the L and T slopes decrease. Thus, there is a possible compensation between the supplementary gain allowed by the frequency decomposition via the input RL and the degradation of the segregation performance attested by these relative level gain measures.

6. POST-PROCESSING

The problem of constraining the blind separation process by adding some a priori information is classical. The post-processing solution follows a sequential strategy, which consists in running at first the BSS and then to enhance the output thanks to a priori information.

6.1 Joint enhancement

The principle of this method is to amplify the input-output RLgain of the H-J unit, i.e., to produce a joint enhancement, this in order to cancel the residual cross-talk. For describing this method, we refer to the relative level gain estimation (§2.3). We fix the time-frame length of the RMS relative level estimation at 512 samples in clean and at 2048 samples in noise, and each time-frame corresponds to a point n. At a given time n, among the two channels, the signal having a positive relative level Y(n) expressed in dB RMS, is selected and amplified. Then, to restore the continuity of the signal, the successive frames are overlapped and added.

As shown Figure 6, the input-output variation of the relative level is amplified. The apparent effect is to drift the points far from 0 dB along the y-axis, i.e., to stretch the output relative level distribution. Hence, this produces a direct amplification of the RLgain measure (Table 3). To accurately control this process, we define a stretching function s(.) compatible with the non-linear tan(.) fit of the H-J unit characteristic function. The function arctan(.) allows us to preserve the same input-output non-linearity, and to control the gain by adjusting a slope parameter α :

$$s(Y(n)) = \frac{2}{\pi}\arctan(\alpha|Y(n)|)$$

The value of α is rather flexible, and it is fixed at 0.5 in our simulations. Other stretching functions s(.), as the Laplacian function are also valid, since the goal is to amplify more the gain when it is already high.

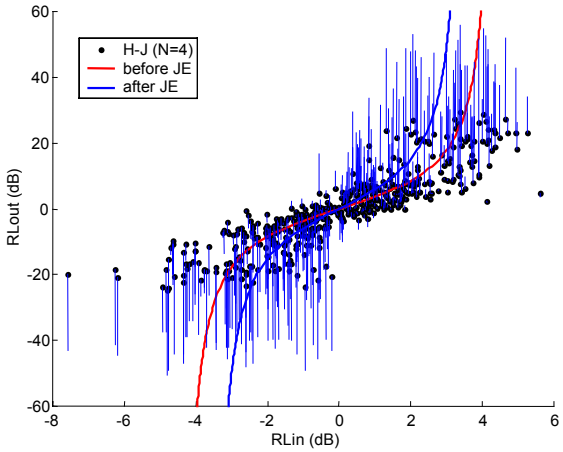

Figure 6: Post-processing by joint enhancement. Simulation of the H-J unit (N=4, non causal, L=201) in the RM condition with 2 speakers in silence (Clean).

input	Clean		Noisy (120Km/h)	
output	1|2 bef.	1|2 after	1|2 bef.	1|2 after
L slope	4.38	8.44	2.68	6.52
T slope	3.49	5.49	2.74	8.75
RLgain	6.96	15.46	2.97	10.39

Table 3: Quantification of the gain of joint enhancement (same inputs as for Figure 1 and Table 1).

This method is preferable to the application of a threshold, relying on the fact that the cross-talk is generally a weak signal. The a priori information is the presence of two sources only. The BSS applications already assume such a constraint, but this has more impact here. Hence, for one speaker only, this method has no sense, and, in presence of a third diffuse source as car noise, the gain is also high (Table 3), but strong audible distortions occur, which are presumably detrimental for ASR. We note that, without post-processing, the H-J unit is transparent for this third stationary source, and that the voice segregation remains satisfactory (Table 3, Noisy). So, to cumulate the advantages of the BSS processing and of the classical robust recognition strategy, a competitive solution is to apply another processing stage after BSS, as the J-RASTA, in order to remove the stationary noise, instead of the joint enhancement.

6.2 Beamforming

The performance of the H-J model implicitly depends on the spatialisation of the sources, but its convergence process is slow and it requires enough data. On the other hand, beamforming, or CASA segregation methods [2] are based on short time TDOA identification. One way to couple these methods is to initialise the de-mixing filters according to a TDOA estimation. But the gain to expect from this approach is small for real mixtures because the delay component is the easiest to identify. As a post-processing, we combine the outputs of several H-J units according to the TDOA, which can be estimated either externally (e.g., by cross-correlation), or by picking the position of the first minimum of each non causal de-mixing filter (Figure 4), this for each channel, and at each time. One property of this approach is to incorporate the fact (evoked §4) that the H-J model generates delayed (filtered) sources, and that the output delays can be appropriated. Knowing the delays, the easiest implementation is the delay and sum beamforming.

This model is tested in the SM condition with 3 microphone pairs (inter mic. dist.=[12,12,24]cm). For each pair, the 2 stereo records of isolated speakers are added together without introducing a delay or an intensity difference. Then, the stereo signal of each pair is an input for a H-J unit (N=4, L=201, non causal) which produces an output pair (1|2). Then, the three output pairs are adequately delayed and summed together to produce a final output pair (1|2). We observe a supplementary spectral gain of resp. 3.40 dB and 2.66 dB, which is calculated relatively to the reference signal obtained by averaging the three microphone signals, but the audible gain is not very convincing.

7. CONCLUSION

In this paper, we have tested the H-J model with realistic mixture conditions in order to propose different improvements, having each a degree of novelty. We conclude that the run-time processing, the use of non causal filters and the frequency decomposition are effective solutions, whereas the post-processing methods are less convincing, but interesting to consider in future works.

Acknowledgements: These results were established during a 2002 summer stay in the POSTECH, of the first author, who thanks Pr. Sung-Yang Bang for welcome in the Intelligent Media Lab. (IML). This French-Korean collaboration was supported by an ARIEL-KOSEF project. We thank Dr. Fritz Class (Daimler-Chrysler) for providing to us the inside car Database [11].

8. REFERENCES

[1] Amari, S., Cichocki, A. & Yang, H. (1995) Recurrent neural networks for blind separation of sources, in Proc. NOLTA-95, Las Vegas, pp. 37-42.

[2] Berthommier, F. & Choi, S. (2001), Evaluation of CASA and BSS models for subband cocktail-party speech separation, in Proc. ICA'01, San Diego, pp. 301-306.

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

[4] Charkani, N. & Deville, Y. (1999) Self-adaptive separation of convolutively mixed signals with a recursive structure. Part II: Theoretical extensions and application to synthetic and real signals, Signal Processing, 75(2):117-140.

[5] Choi, S., Hong, H., Glotin, H. & Berthommier, F. (2002) Multichannel signal separation for cocktail party speech recognition: A dynamic recurrent network, Neurocomputing, 49(1-4):299-314.

[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, France, pp. 177-212.

[7] Gardner, B. & Martin K. (1994) HRTF Measurements of a KEMAR Dummy-Head Microphone, MIT Media Lab.

[8] Hérault, J., Jutten, C. & Ans, B. (1985) Détection de grandeurs primitives dans un message composite dans une architecture neuromimétique en apprentissage non supervisé, in Proc. GRETSI, Nice, France, pp. 1017-1020.

[9] Jutten, C. & Hérault, J. (1991) Blind separation of sources, Part 1: An adaptive algorithm based on neuromimetic architecture, Signal processing, 24:1-10.

[10] Lee, T.-W., Bell, A.J. & Lambert R.H. (1997) Blind separation of delayed and convolved sources, in "*Advances in Neural Information Processing Systems"*, 9:758-764, MIT Press.

[11] Linhard, K., Kolano, G. & Kienzler, C. (2002) Speech+noise inside car, Part I| II DaimlerChrysler research, Information and communication, dept. RIC/AD.

[12] Nguyen Thi, H.-L. (1993) Séparation aveugle de sources à large bande dans un mélange convolutif, PhD thesis, INPG, France (in French).

[13] Nguyen Thi, H.-L. & Jutten, C. (1995) Blind separation for convolutive mixtures, Signal Processing, 45(2):209-229.

[14] Platt, J. C. & Faggin, F. (1991) Networks for the separation of sources that are superimposed and delayed, In J. E. Moody et al. (Eds), Advances in Neural Information Processing Systems, 4:730-737.

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

[16] 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.

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