Evaluation and comparison of sound/speech segregation methods

Frédéric Berthommier

Institut de la Communication Parlée
INPG, France

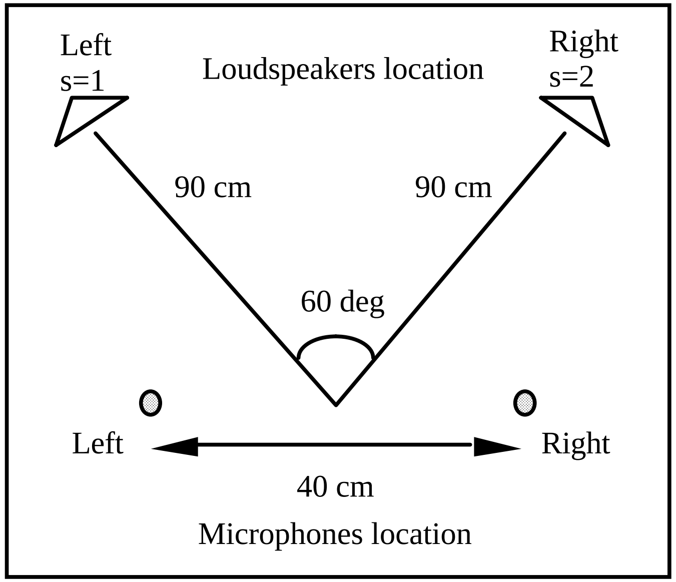
- Comparison between CASA and BSS for binaural segregation of speech

Collaboration with Seungjin Choi (POSTECH, South Korea)

This work was presented during ICSP ’2001 (Pohang)
and ICA ’2001 (San Diego)

The CASA model was presented during the CRAC workshop
(Tessier & Berthommier, 2001)

Number95 Stereo Database



ST-Numbers95 Database

ICP/INP Grenoble

Authors: E.Tessier and F. Berthommier

Left source

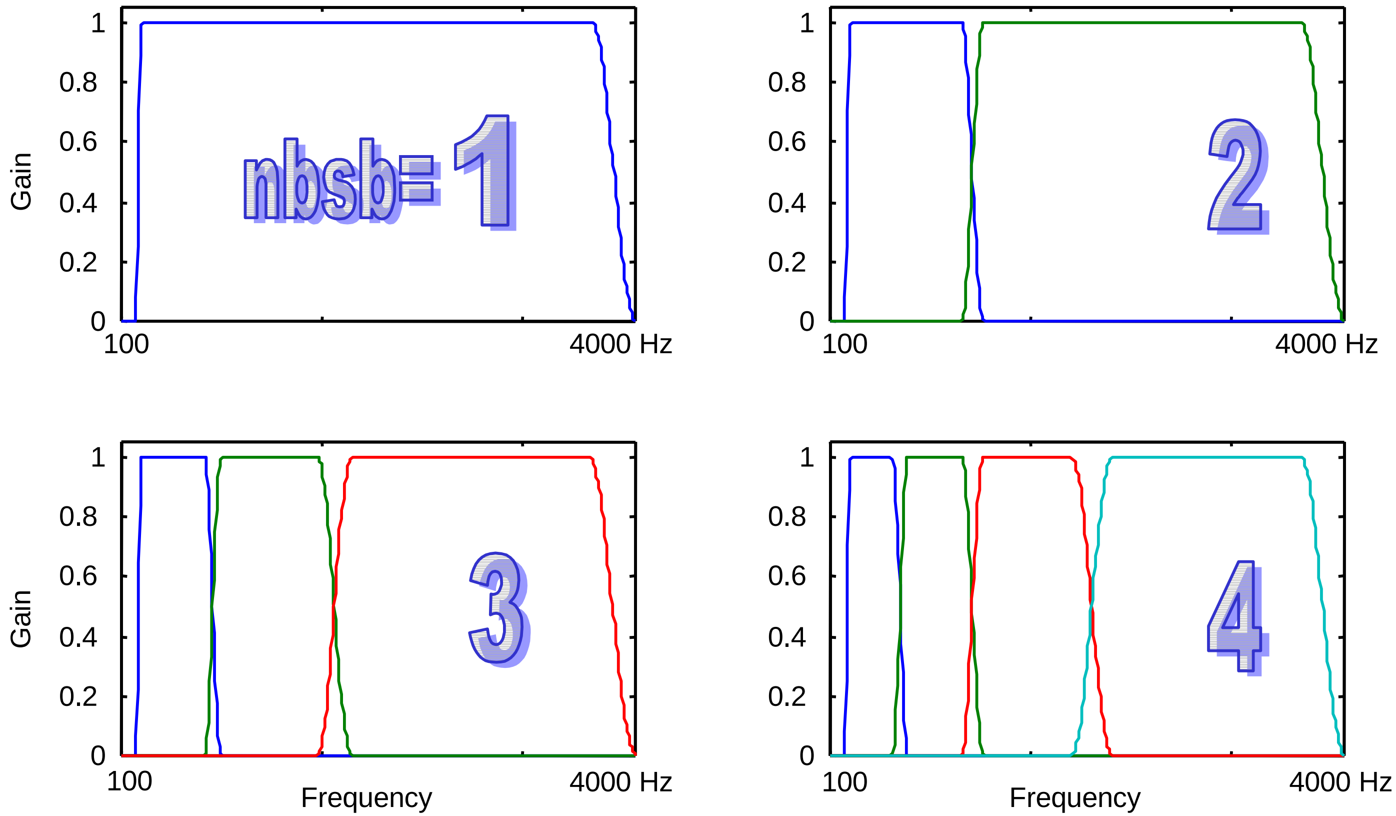
Right source

Reference

Mixture

A large database of binary mixtures of sentences (n=613) has been recorded by [Tessier and Berthommier, 1999]. The signal of Numbers95 is played by loudspeakers and recorded. The temporal overlap between words is about 75% and the relative level is 0dB. The setup is static. Only 332 mixture sentences truncated at 1 s are used in the present study.

Filterbank decomposition

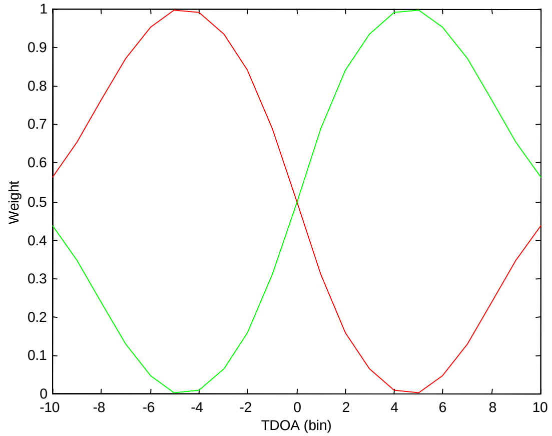


Subband processing

The CASA Model

Filterbank decomposition

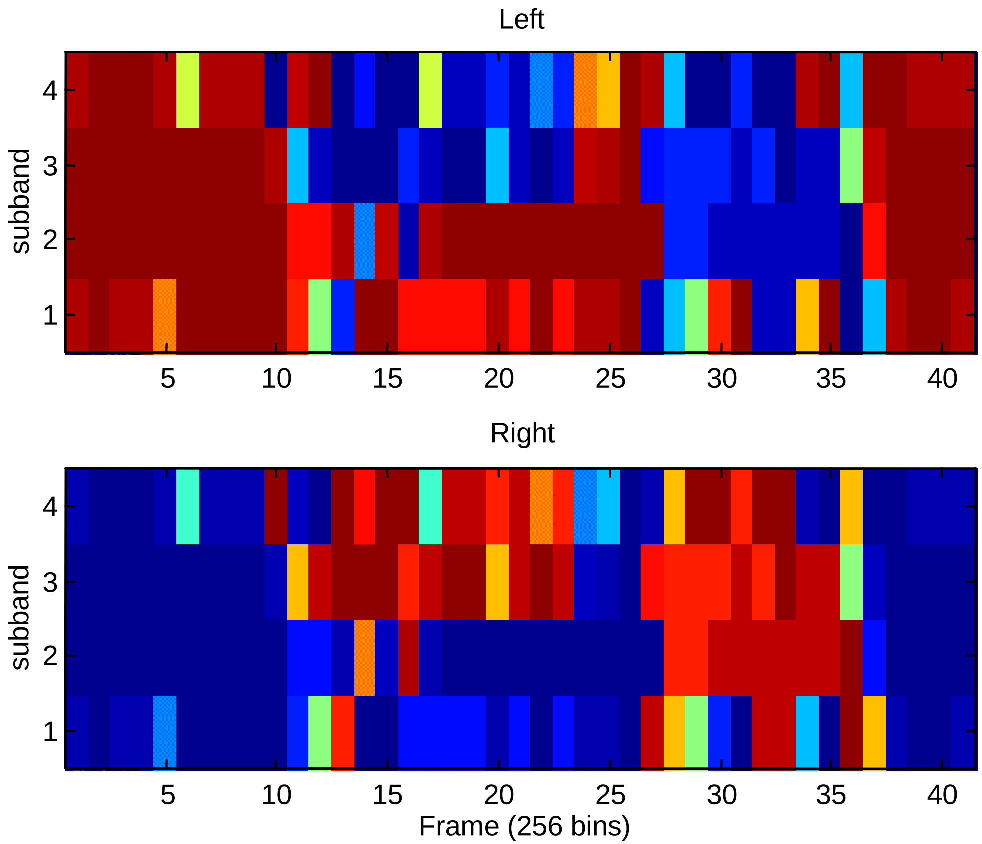
$$|X_i(\omega)| = F_i(\omega)|X_{left}(\omega)|$$



TDOA estimation and weighting

$$W_{s,i}(TDOA_i) = (1 - W_{s',i}(TDOA_i))$$

$$|Y_{s,i}(\omega)| = W_{s,i}(TDOA_i)|X_i(\omega)|$$

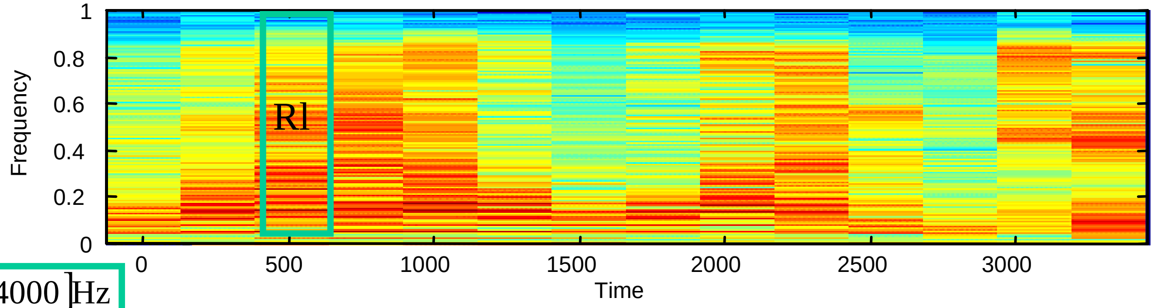


Resynthesis $|Y_s(\omega)| = \sum_{i=1}^{nbsb} |Y_{s,i}(\omega)|$

RESPITE

ARIEL-KOSEF

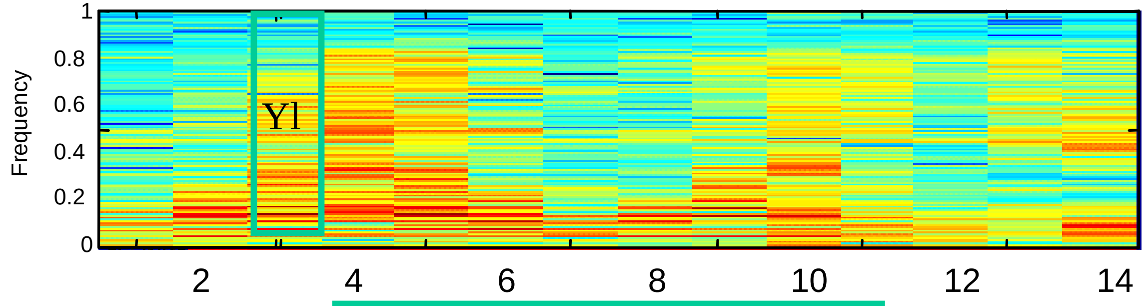
Reconstruction Acuracy



Reference

Left source

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



Left output

Frame of 1024 bins with half overlap

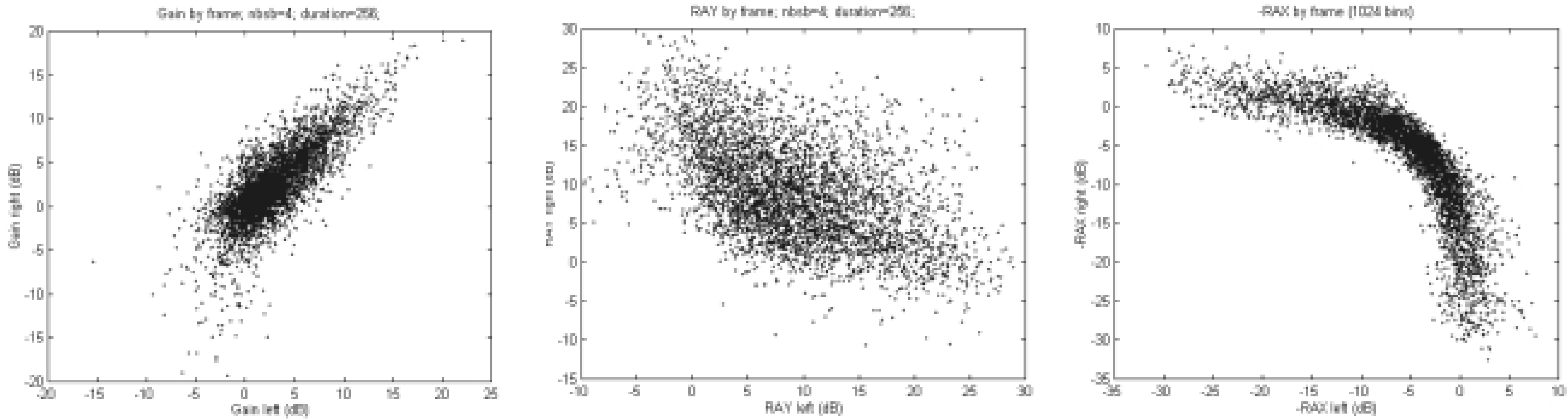
RA (output)

$$RAY_S = 10 \log \frac{\int_{\Omega} \left|R_S(\omega)\right|^2}{\int_{\Omega} \left(\left|R_S(\omega)\right| - \left|Y_S(\omega)\right|\right)^2}$$

RA (mixture)

$$RAX_S = 10 \log \frac{\int_{\Omega} \left|R_S(\omega)\right|^2}{\int_{\Omega} \left(\left|R_S(\omega)\right| - \left|X_S(\omega)\right|\right)^2}$$

Gain of CASA

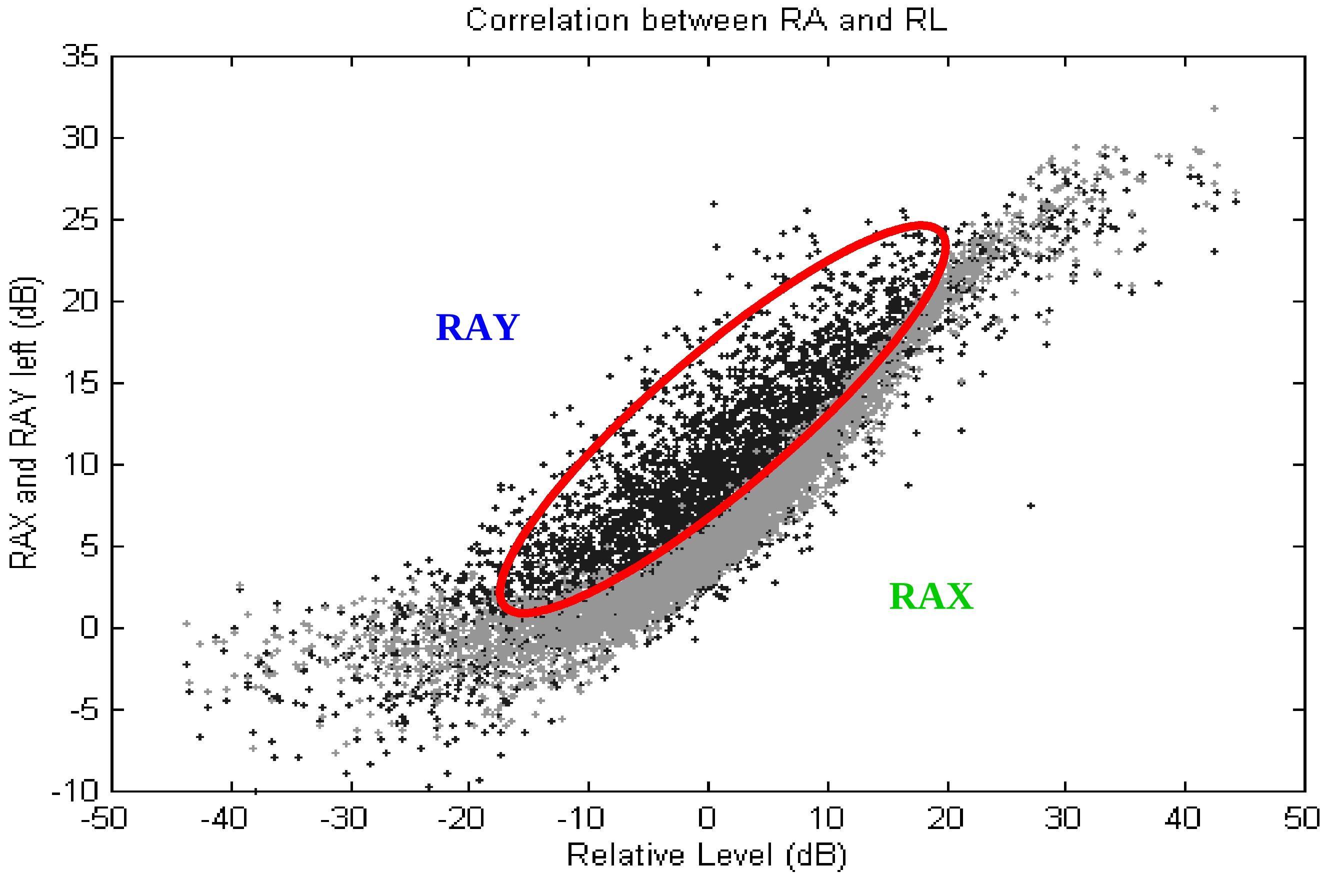


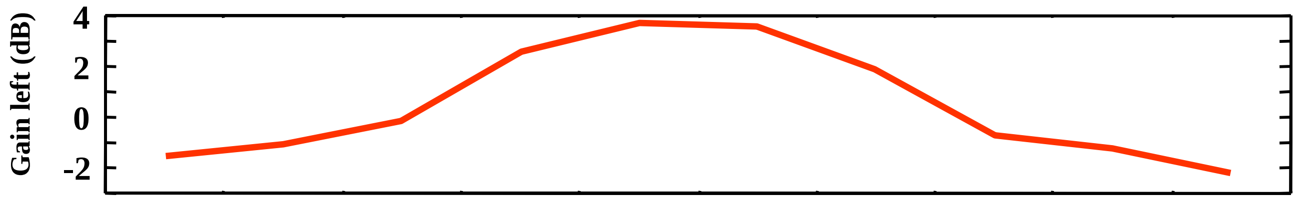
$$\mathrm{Gain}_s = \mathrm{RAY}_s - \mathrm{RAX}_s = \mathrm{RA(output)} - \mathrm{RA(mixture)}$$

$$s = \text{left or right}$$

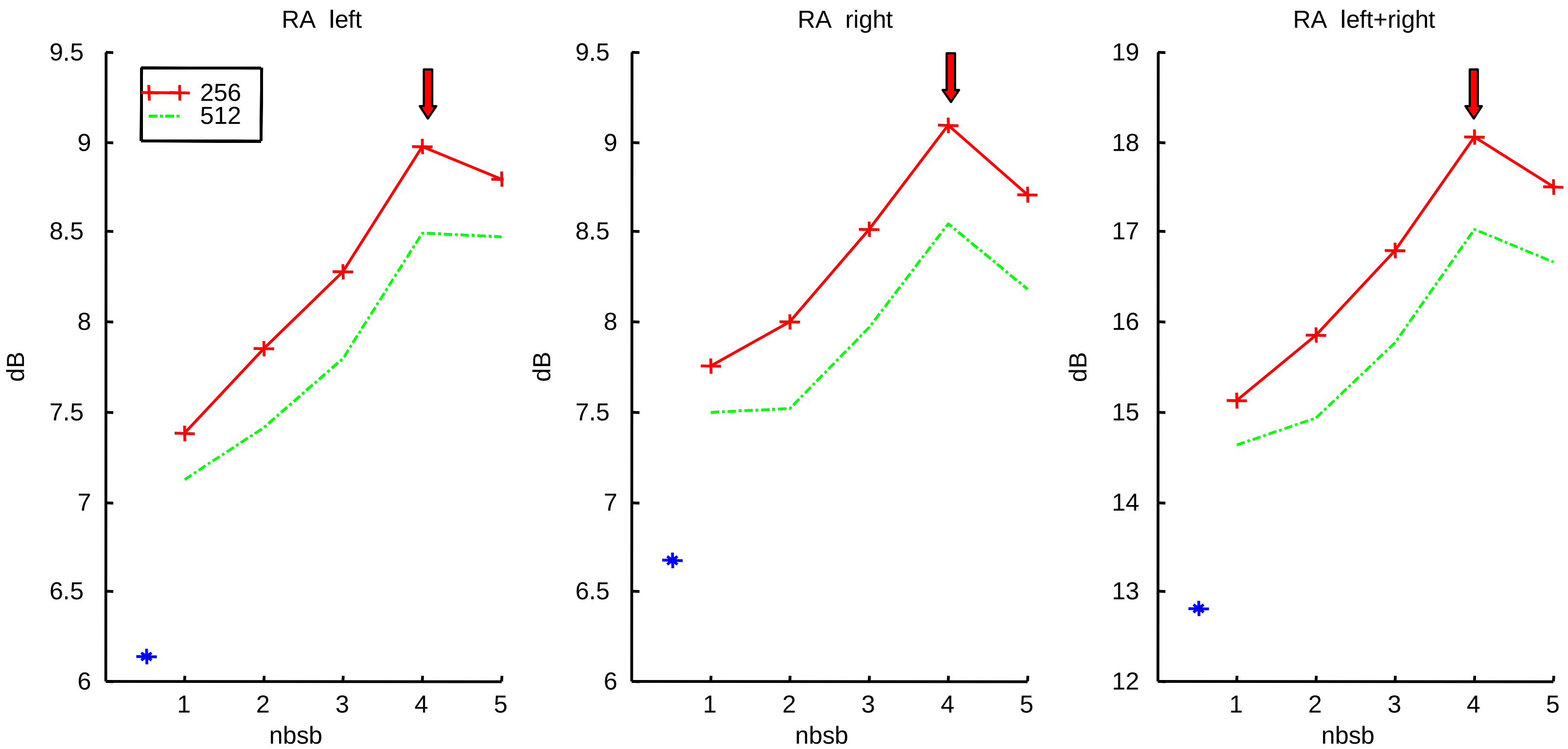
$$\mathrm{Gain}_s = 10\log\frac{\int_\Omega |R_s(\omega)|^2}{\int_\Omega (|R_s(\omega)|-|Y_s(\omega)|)^2} - 10\log\frac{\int_\Omega |R_s(\omega)|^2}{\int_\Omega (|R_s(\omega)|-|X_s(\omega)|)^2} = 10\log\frac{\int_\Omega (|R_s(\omega)|-|X_s(\omega)|)^2}{\int_\Omega (|R_s(\omega)|-|Y_s(\omega)|)^2} \quad (=\mathrm{SNRI}_s)$$

Gain of CASA : Relative Level



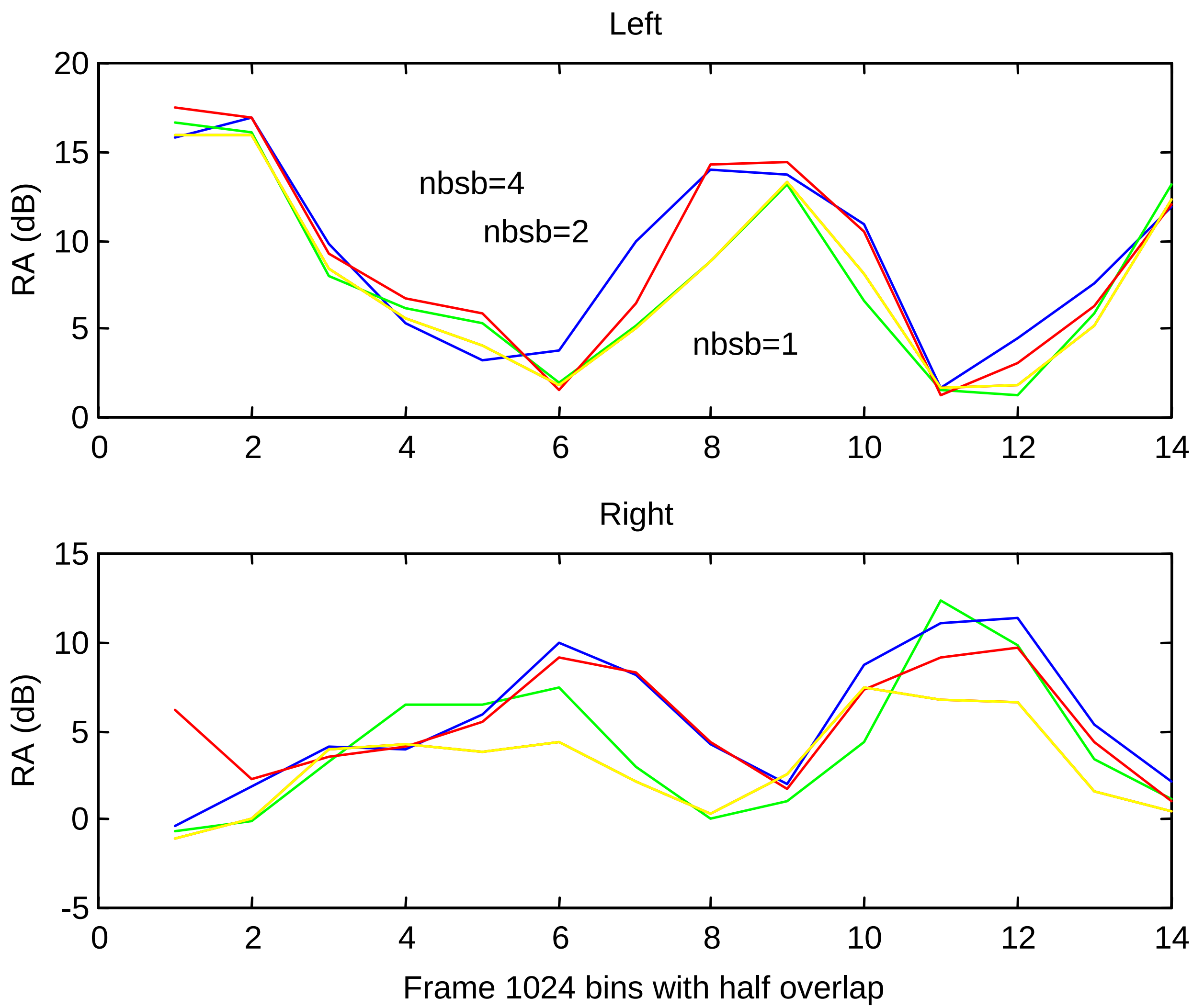


Subband effect for CASA



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 (*).

Effect of nbsb : RA



Mixt.

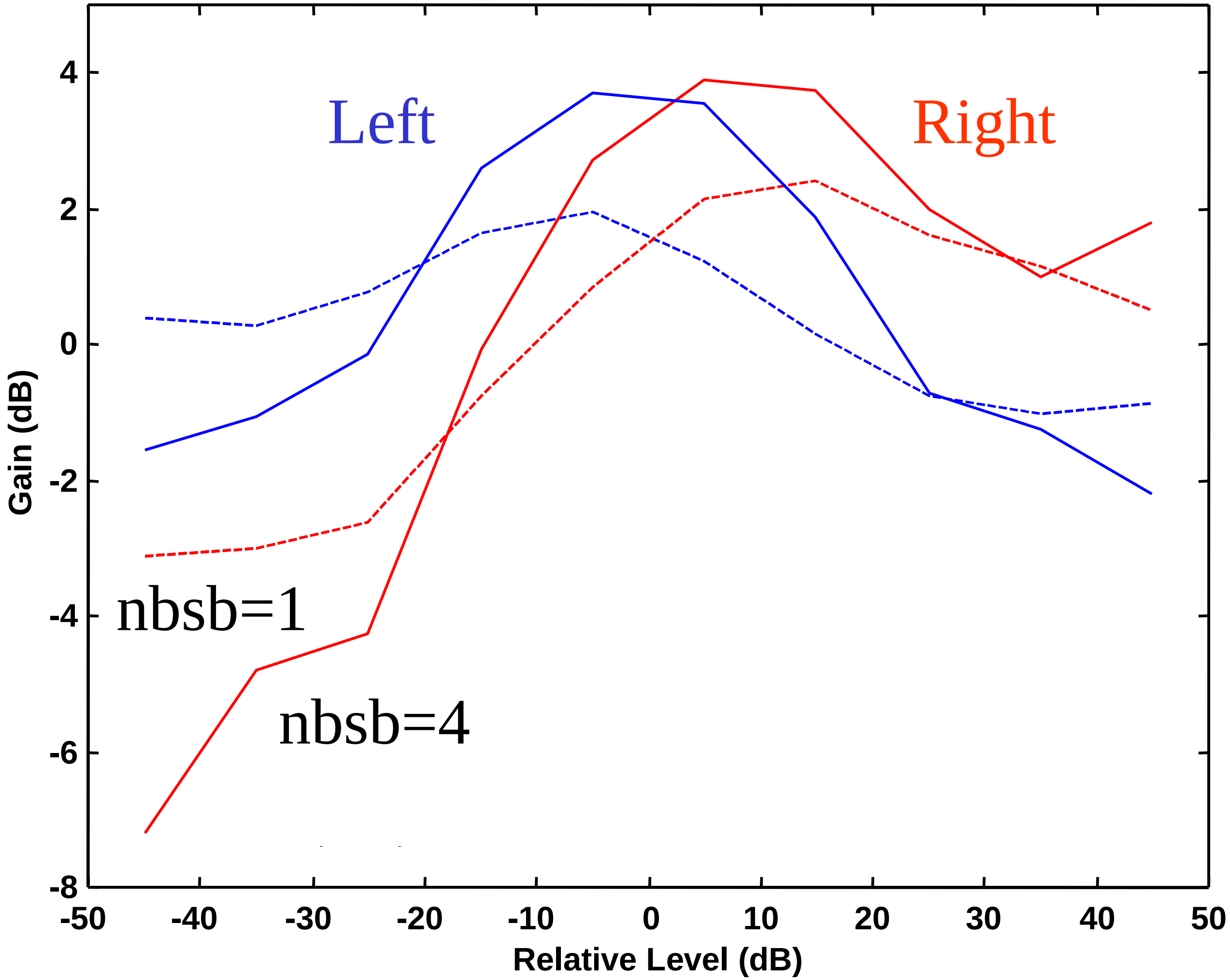
Left

Right

2

4

Subband effect for CASA: Gain

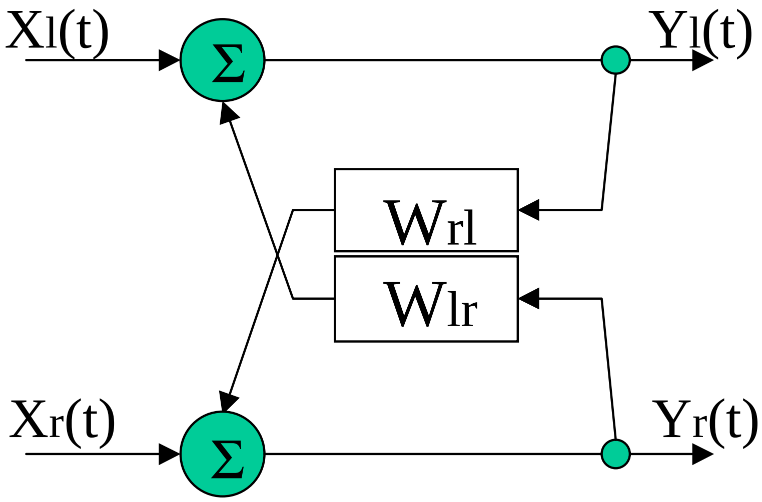


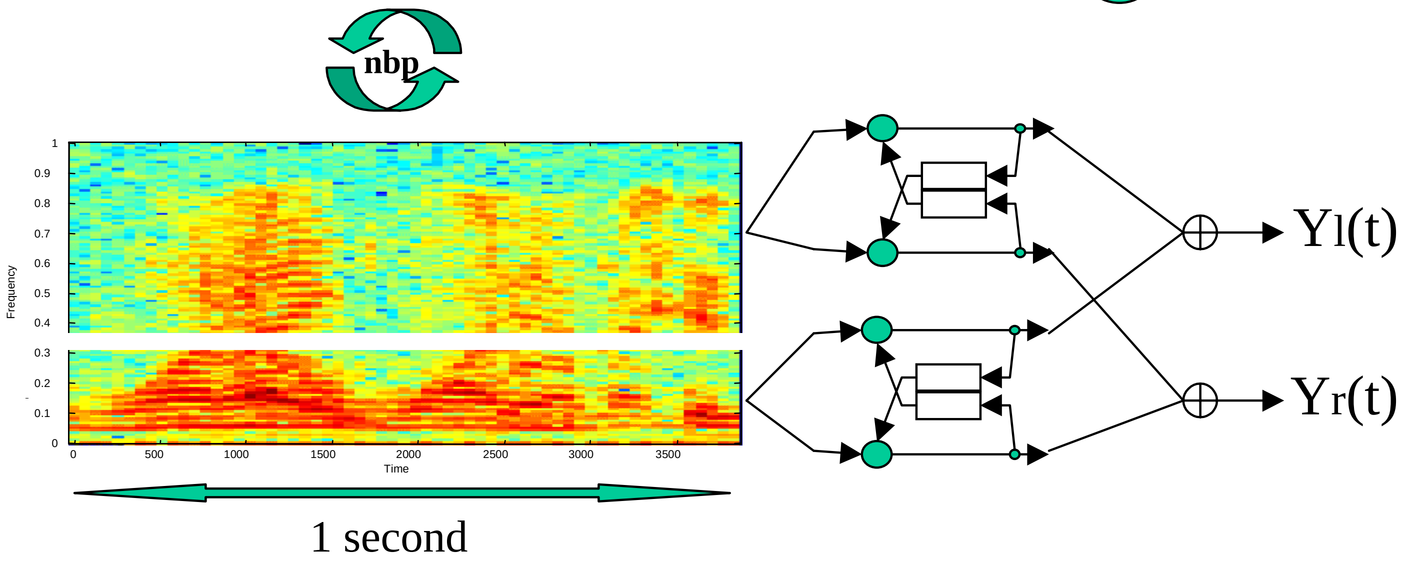
The BSS Model

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

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

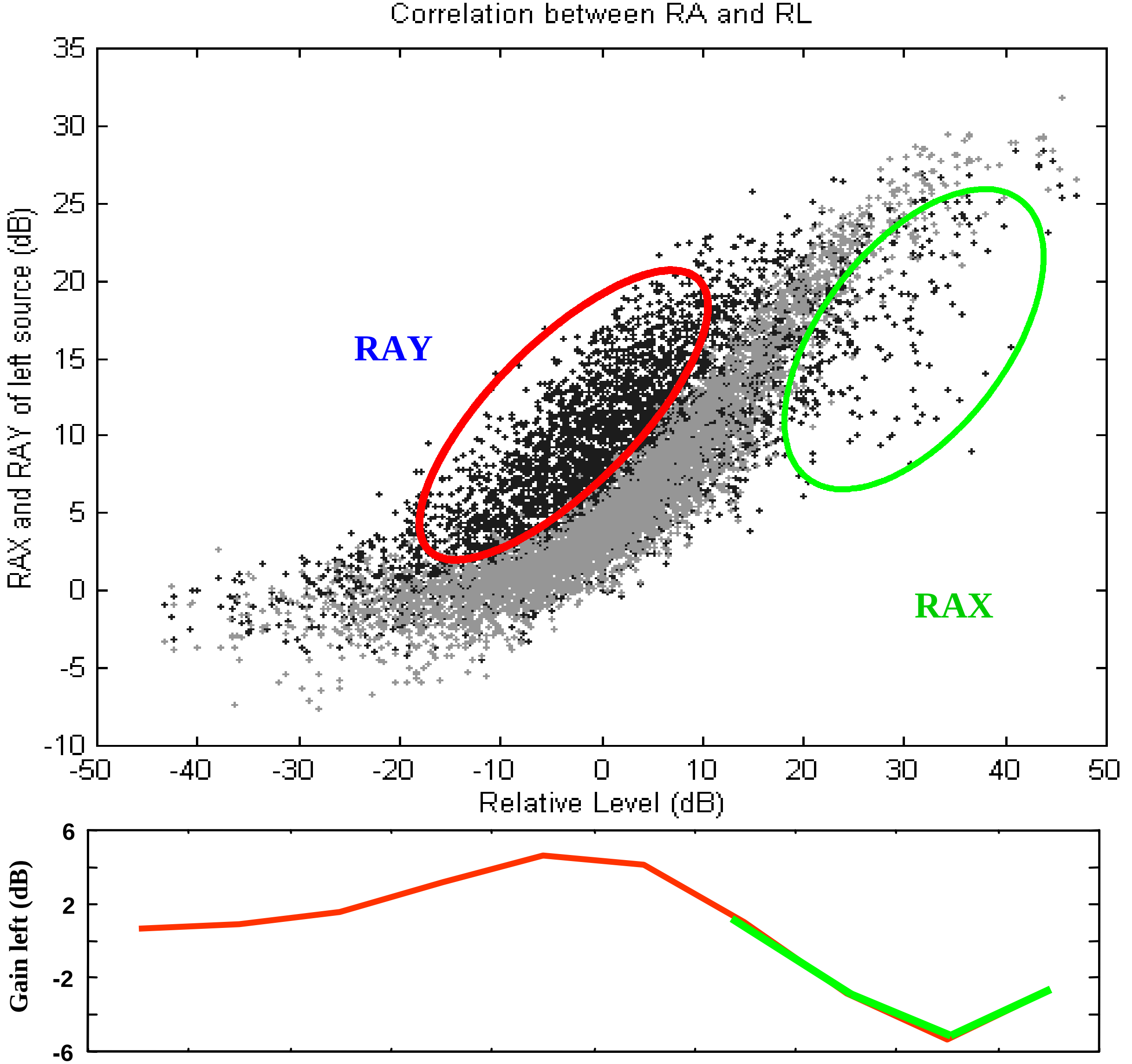
Gain | Non linear function | Delayed output



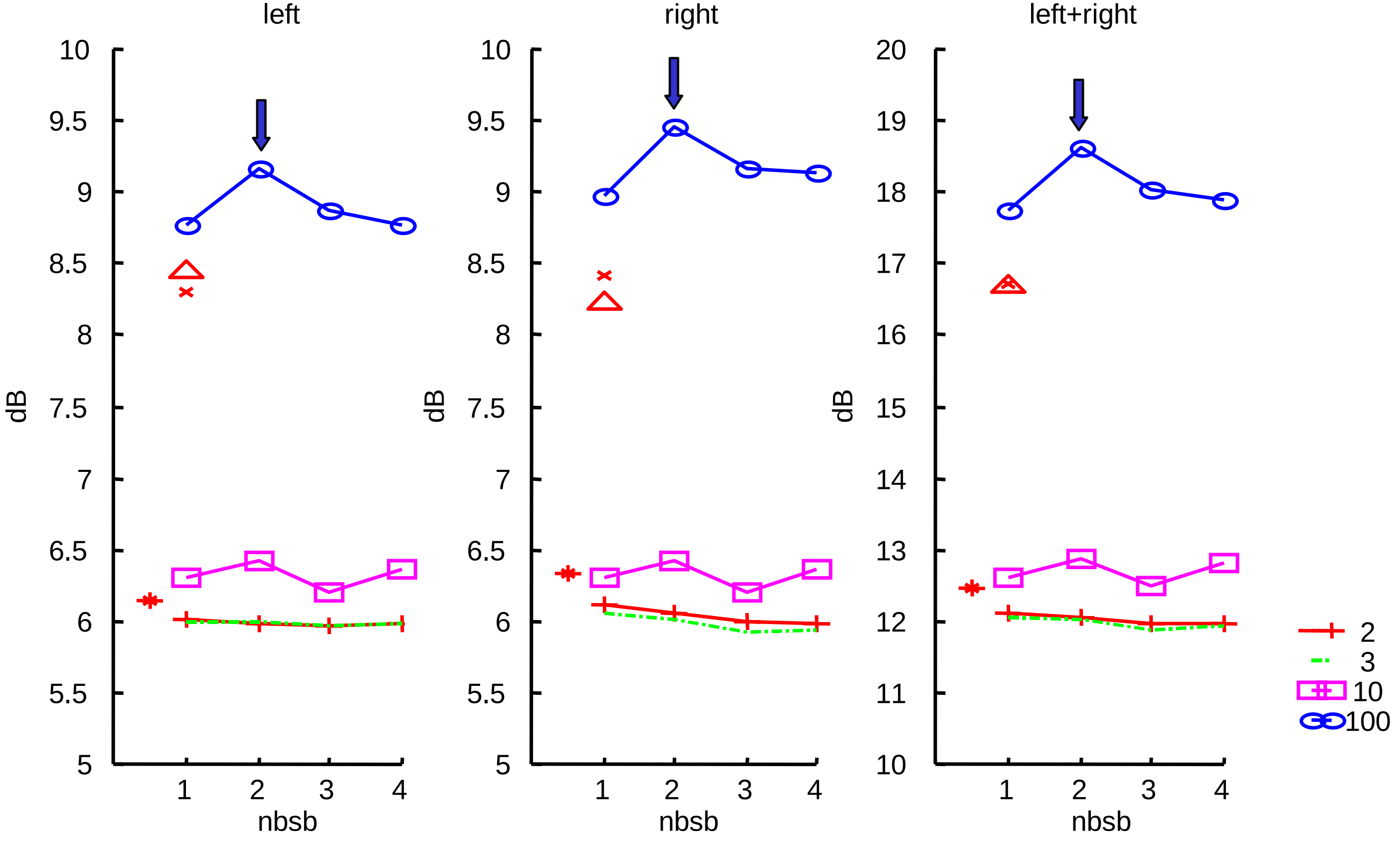


RESPITE ARIEL-KOSEF

Gain of BSS :Relative Level



Subband effect for BSS



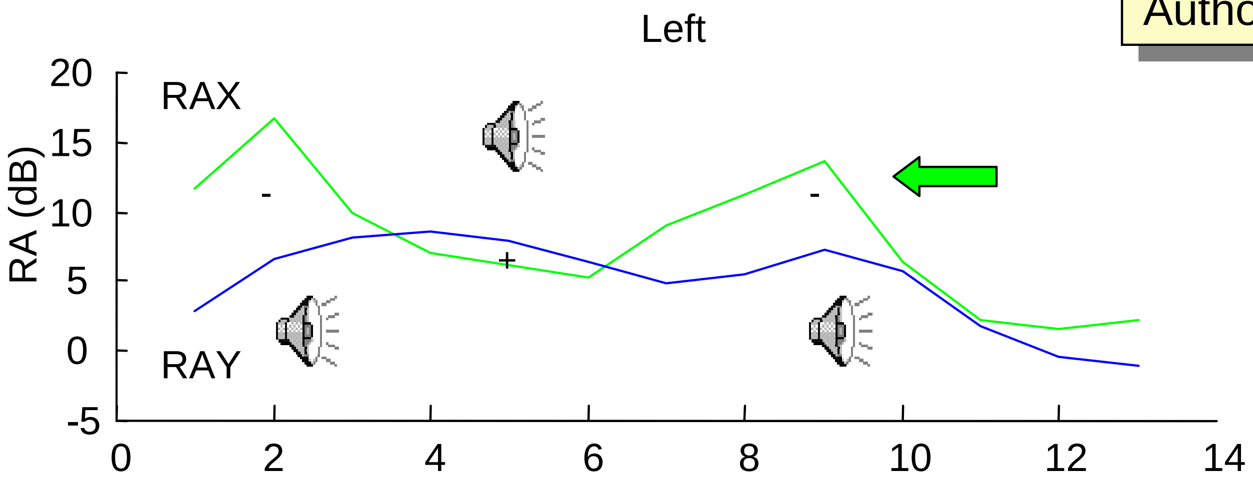
Effect of the number of subbands (nbsb) for the BSS model on the RA (in dB). From left to right: av. left source RA, av. right source RA, av. left+right RA over all frames. The number of subbands varies from 1 to 4 and the three curves correspond to nbp= 2,3,10, 100. The RA of the mixture is labelled (*). In each figures, two points are added at nbsb=1 for the "BSS giv" condition (×) and for "BSS ori" data (Δ).

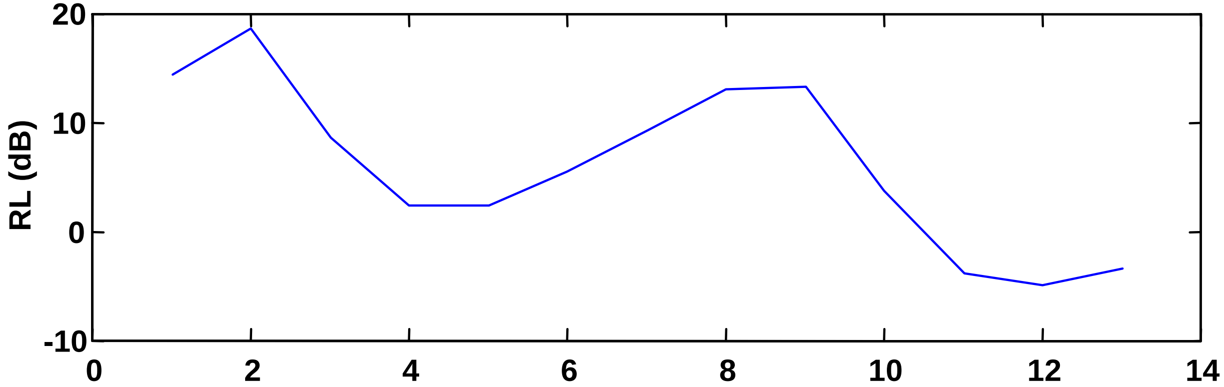
RA and Gain for BSS

Speech Separation Program (C++)

Chungbuk University & POSTECH

Authors: H. Hong and S. Choi



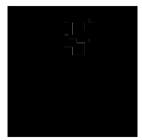


Frame 1024 bins with half overlap

Mixt.

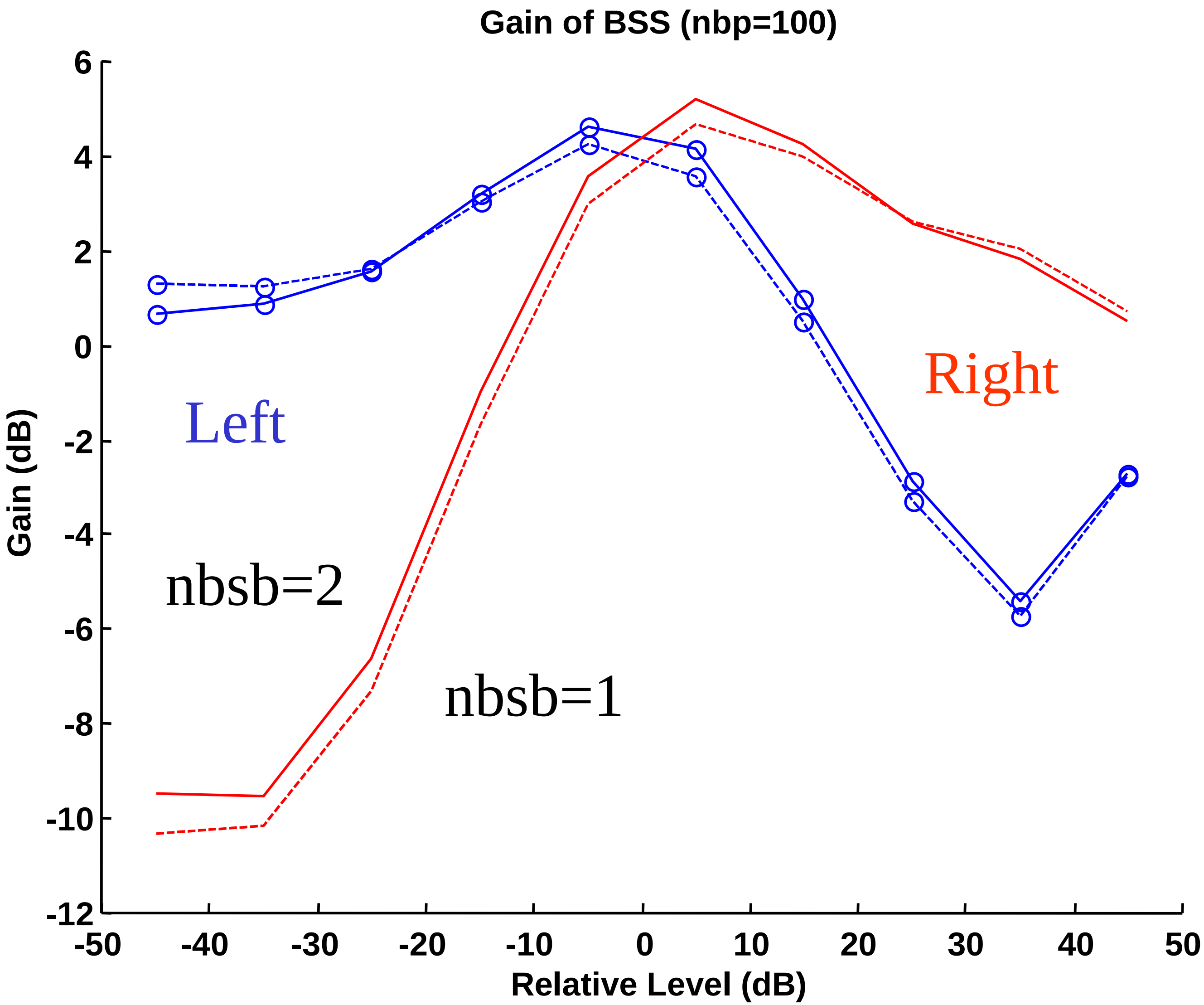
Left

Right

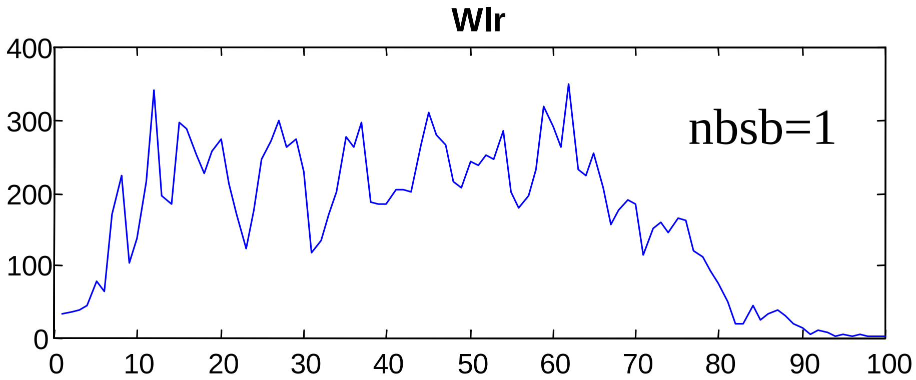


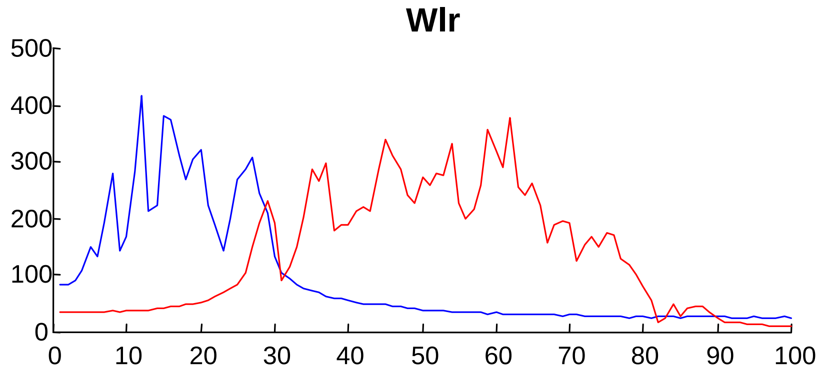
LISA.exe

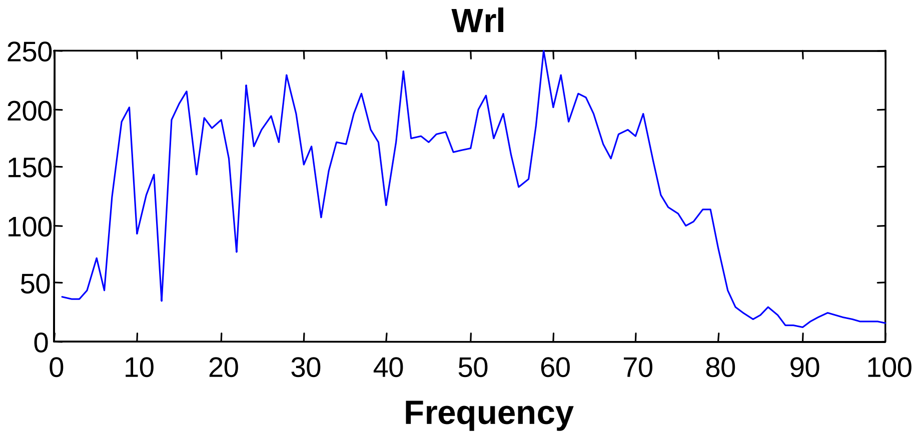
Subband effect for BSS: Gain

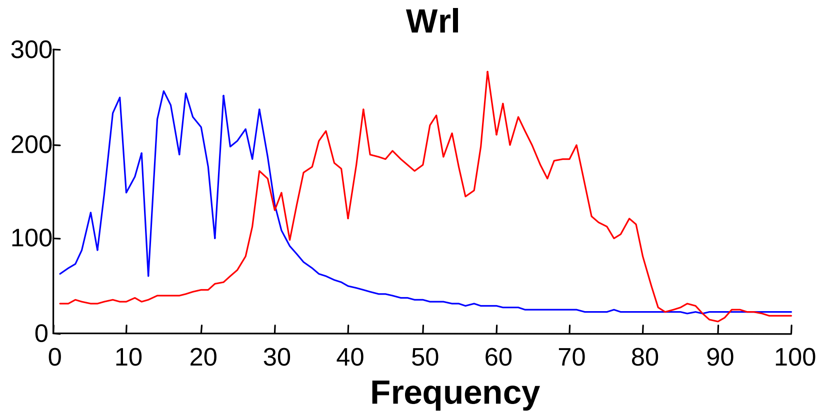


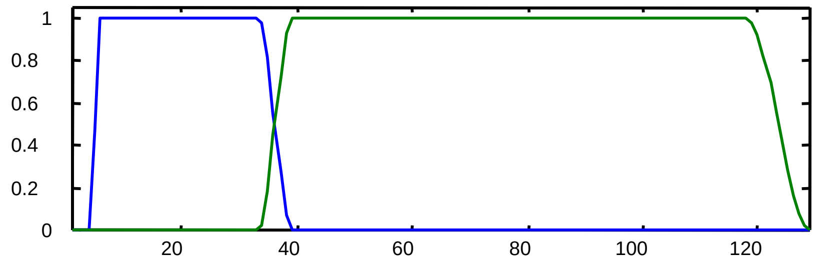
Demixing filters



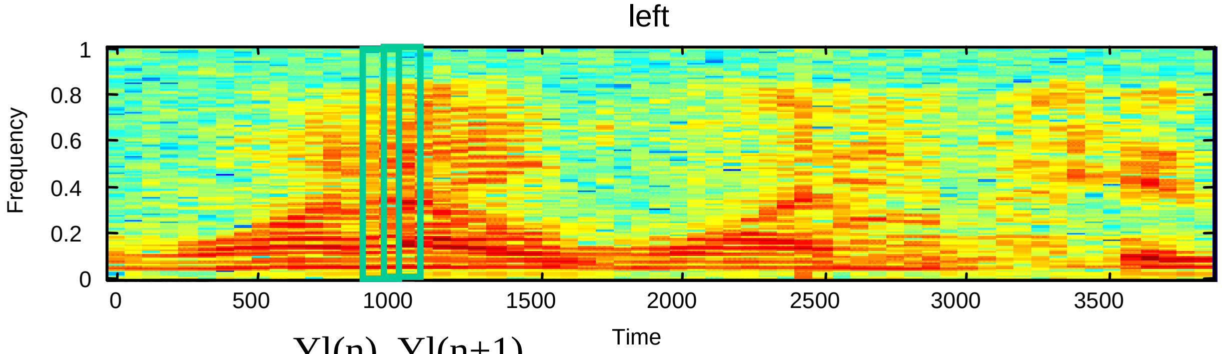








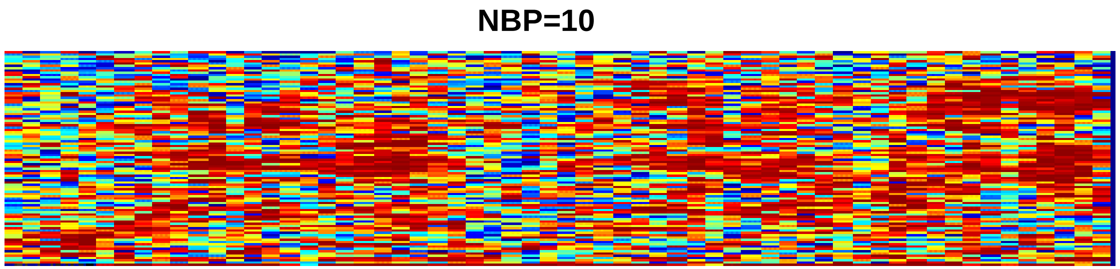
Coherence spectrograms



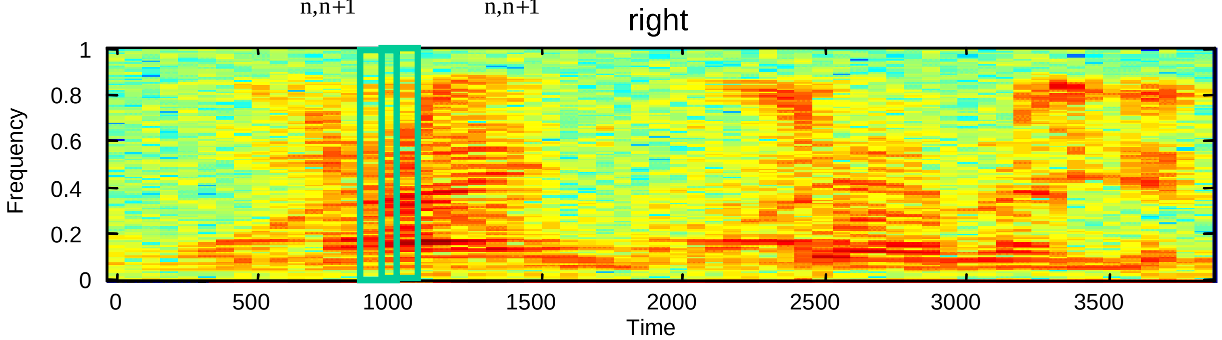
Yl(n), Yl(n+1)

Frames of 256 bins with half overlap

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

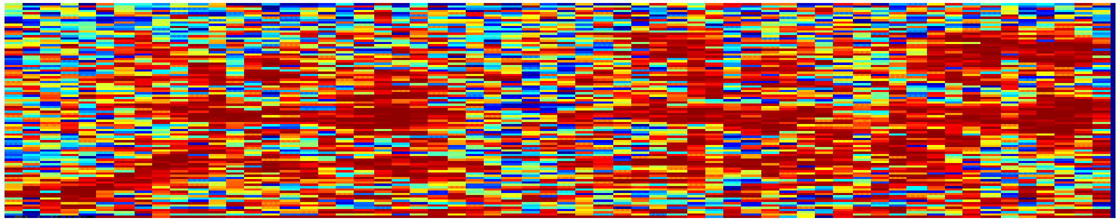
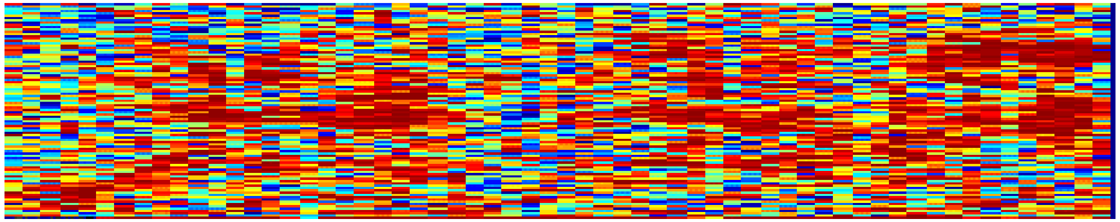
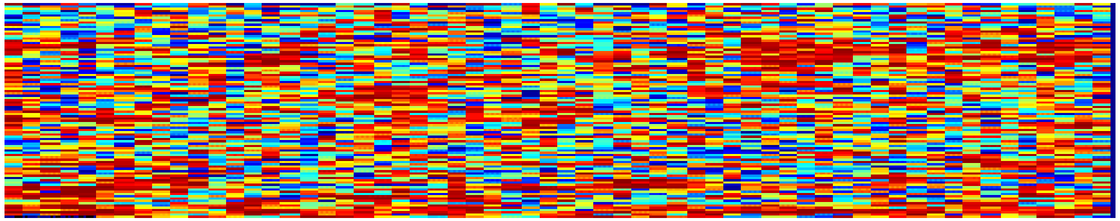


Mean(Coh)=0.65

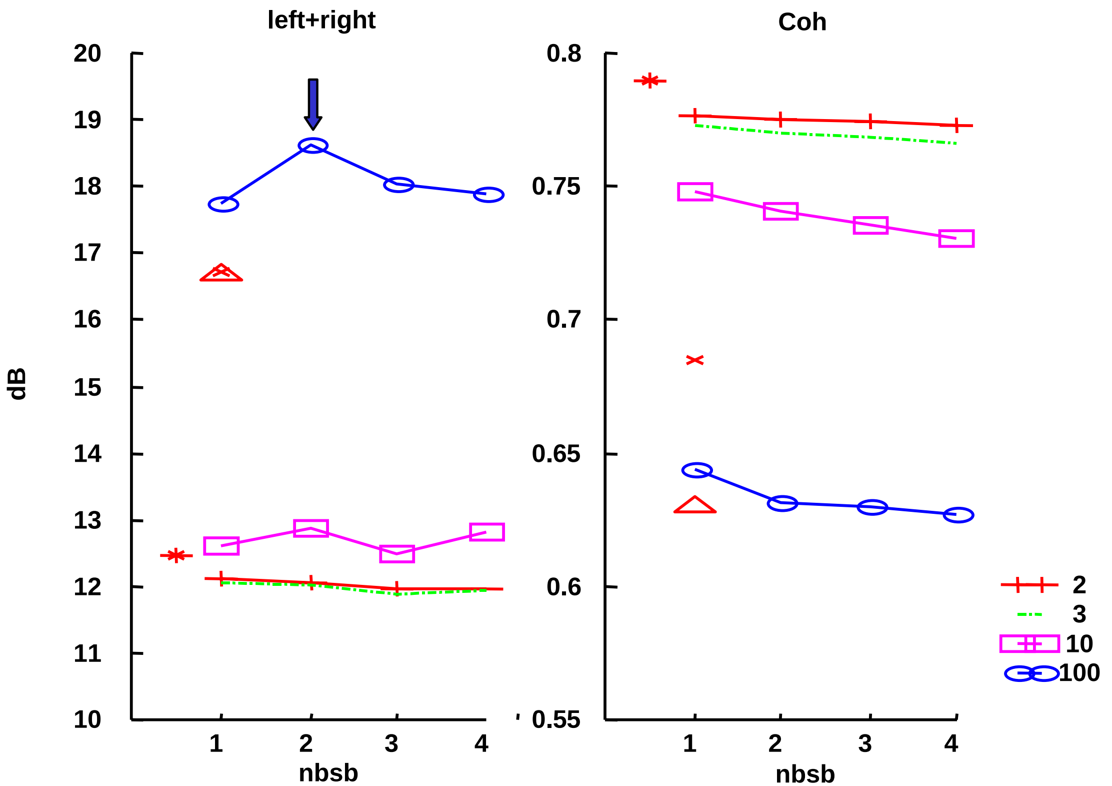


Yr(n), Yr(n+1)

Effect of nbp: Coherence spectrograms

Coh		Left		Right
0.68	NBP=3 		3	
0.65	NBP=10 		10	
0.60	NBP=100 		100	

Coherence statistic



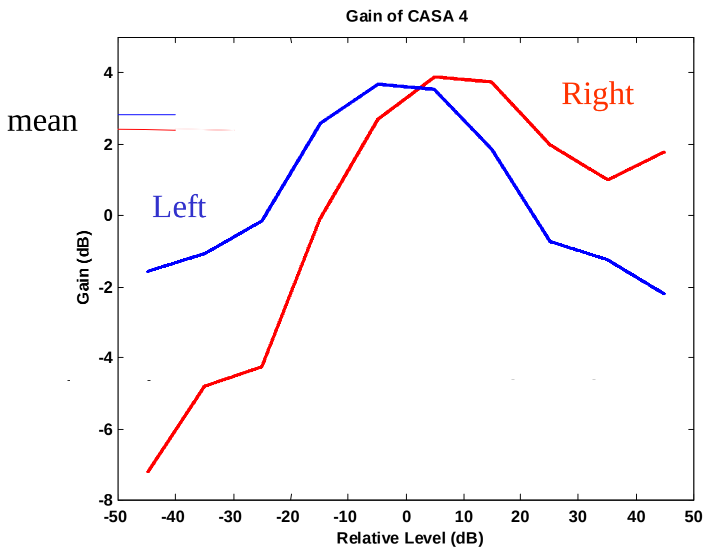
Effect of the number of subbands (nbsb) on the coherence index for the BSS model. Left: average left+right RA over all frames. Right: coherence defined as the mean of the coherence spectrogram. The number of subbands varies from 1 to 4 and the three curves correspond to nbp= 2,3,10, 100. The RA of the mixture 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).

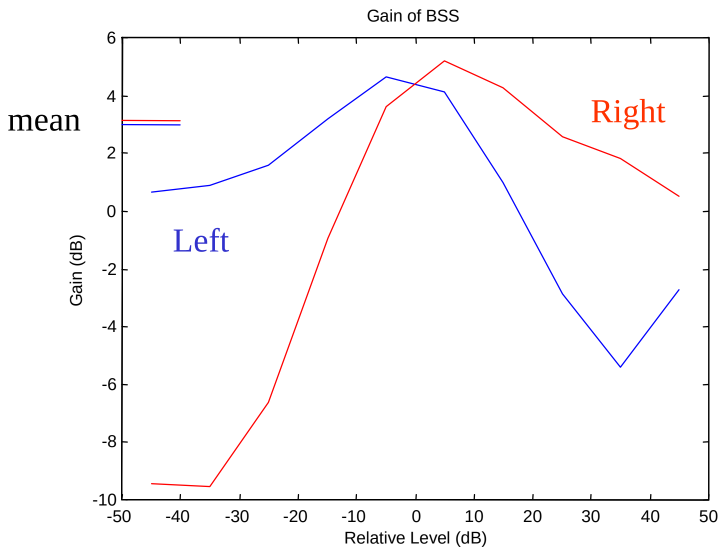
Summary results

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	18.62	9.16	9.46	3.02	3.13
BSS ori	16.67	8.44	8.22	2.29	2.13
BSS giv	16.71	8.29	8.42	2.15	2.09

	REF	CASA	BSS
Left			
Right			

… Hearing

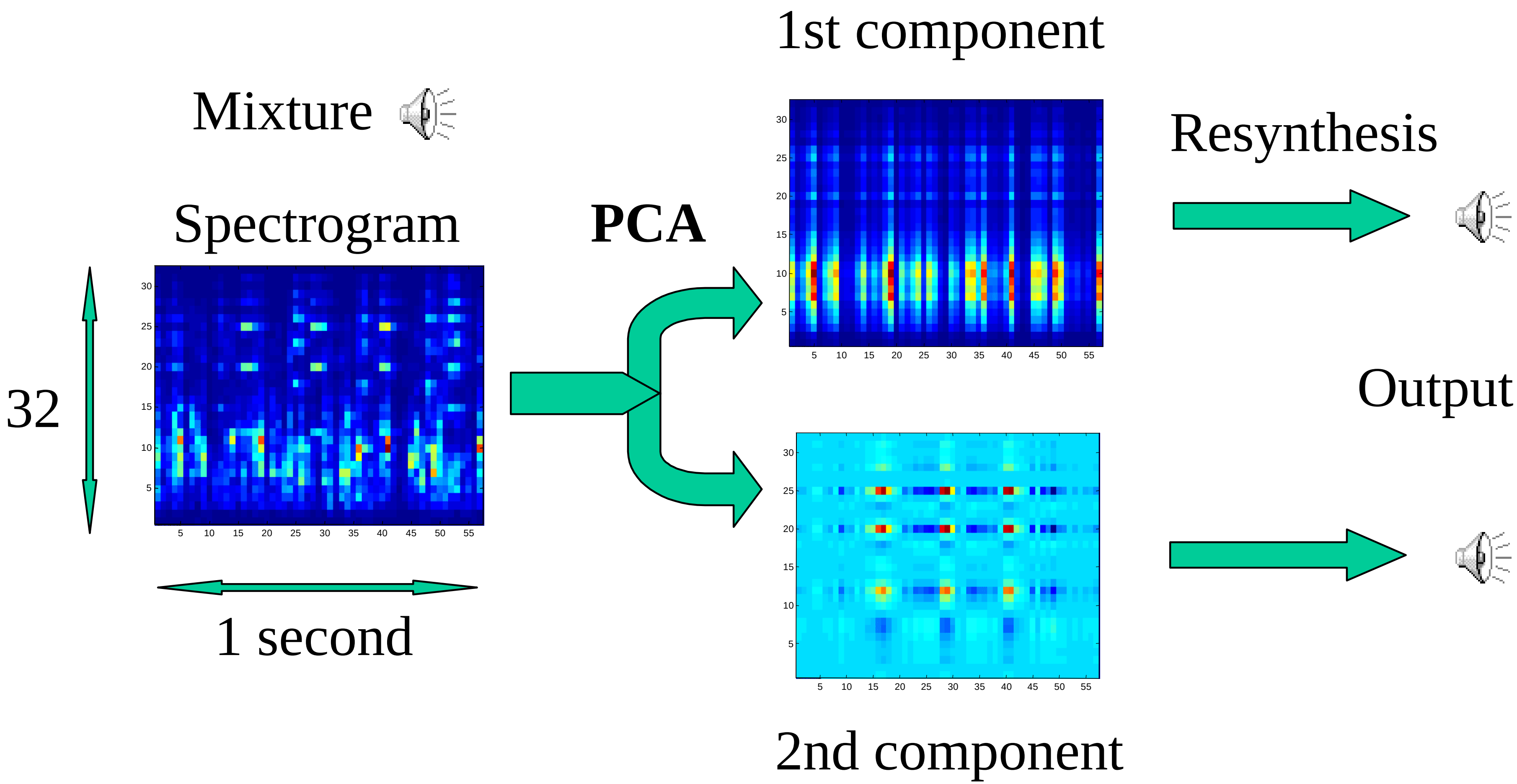




- Comparison between CASA and BSS for binaural segregation of speech

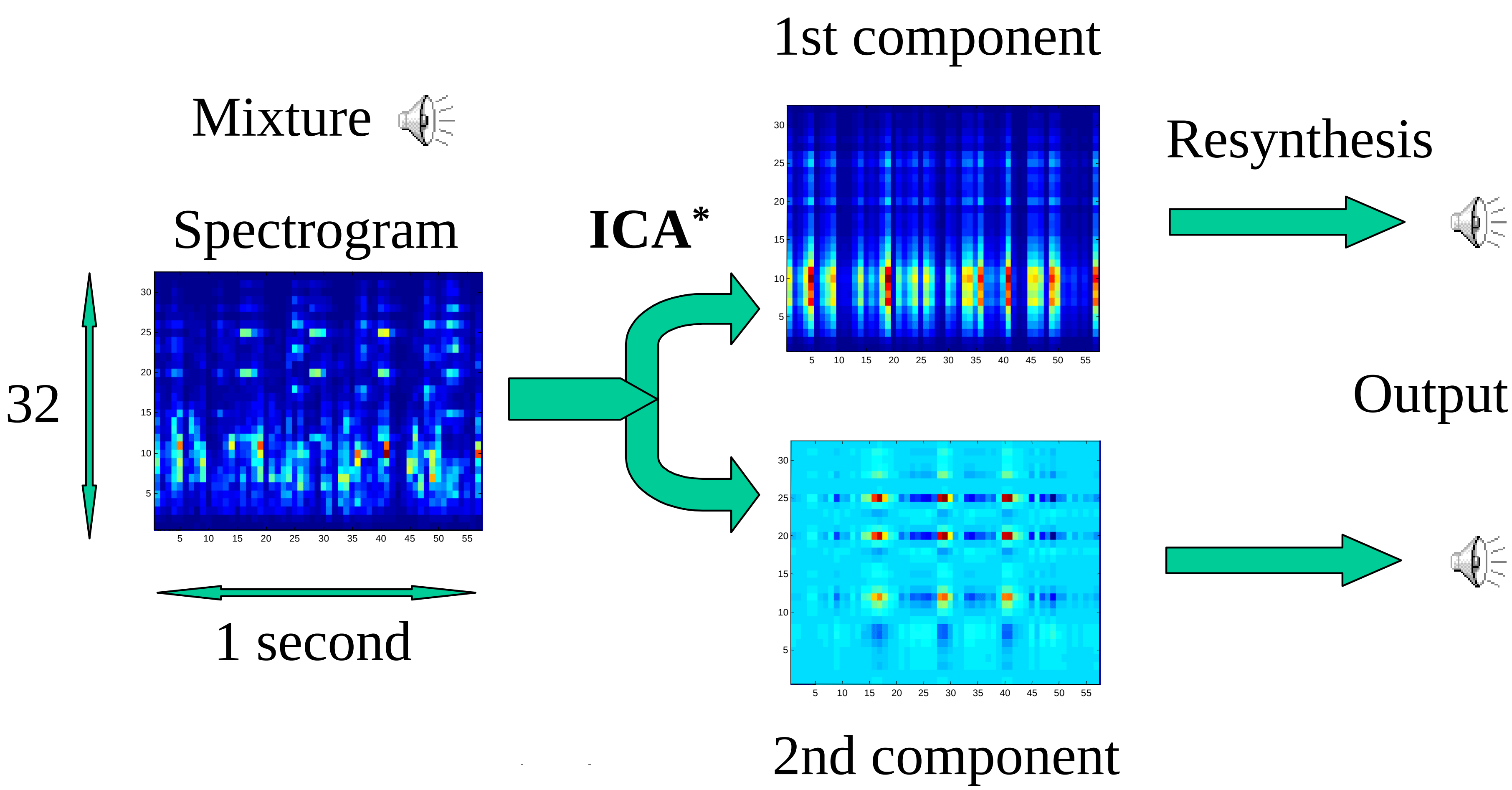
- Monaural segregation of sounds and speeches

Principle



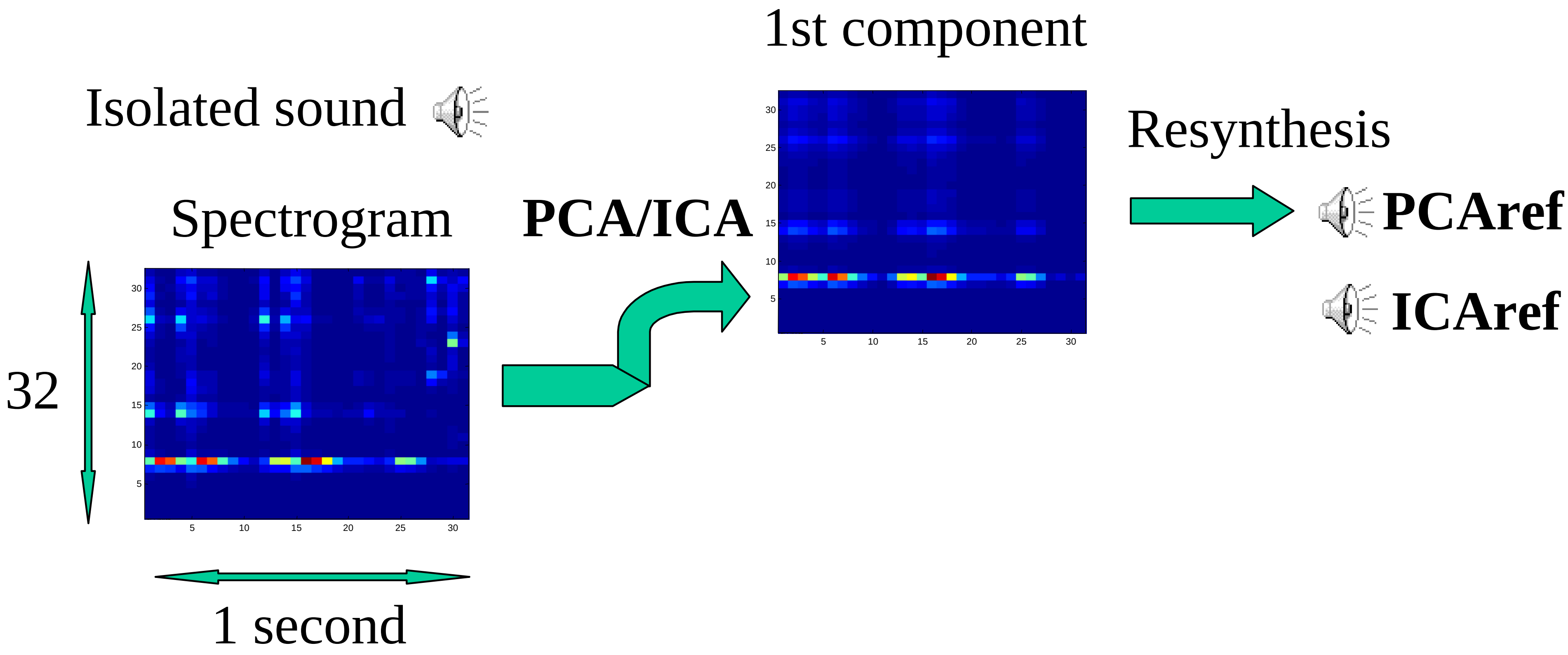
Extraction of components which are
« vertically » or « horizontally » correlated

Principle (II)



ICA*: Independent component analysis by JADE (Cardoso, 1997)

References



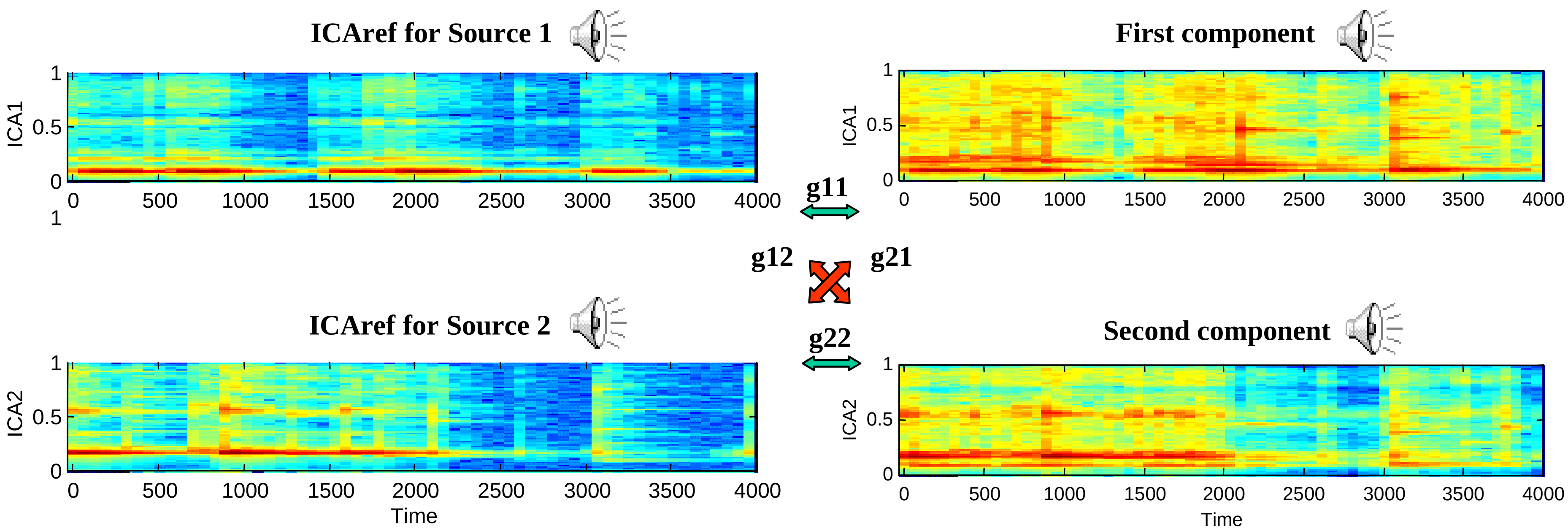
Isolated sound
Spectrogram
32
1 second
PCA/ICA
1st component
Resynthesis
PCAref
ICAref

Gain evaluation

Mixture

ANGK-BALO (66)

REFERENCES **OUTPUT**



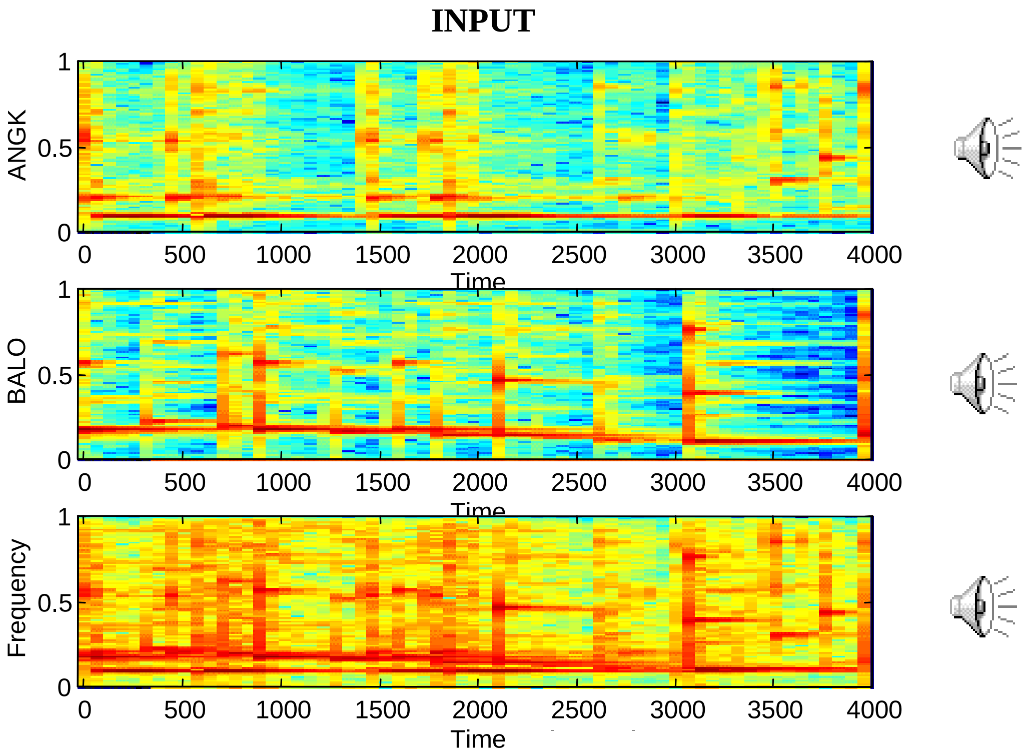
Gain= max(g11+g22 , g12+g21)

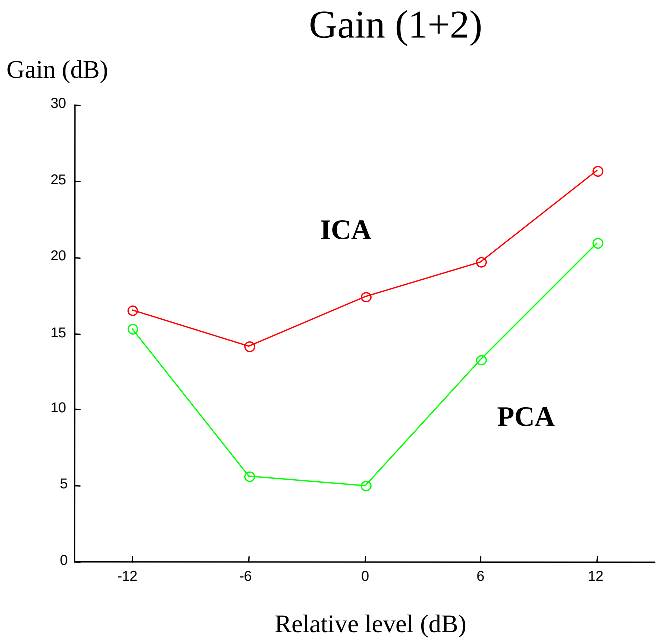
Partial gain = g11 or g12, g22 or g21

Detailed example I (Music)

Mixture

ANGK-BALO (66)



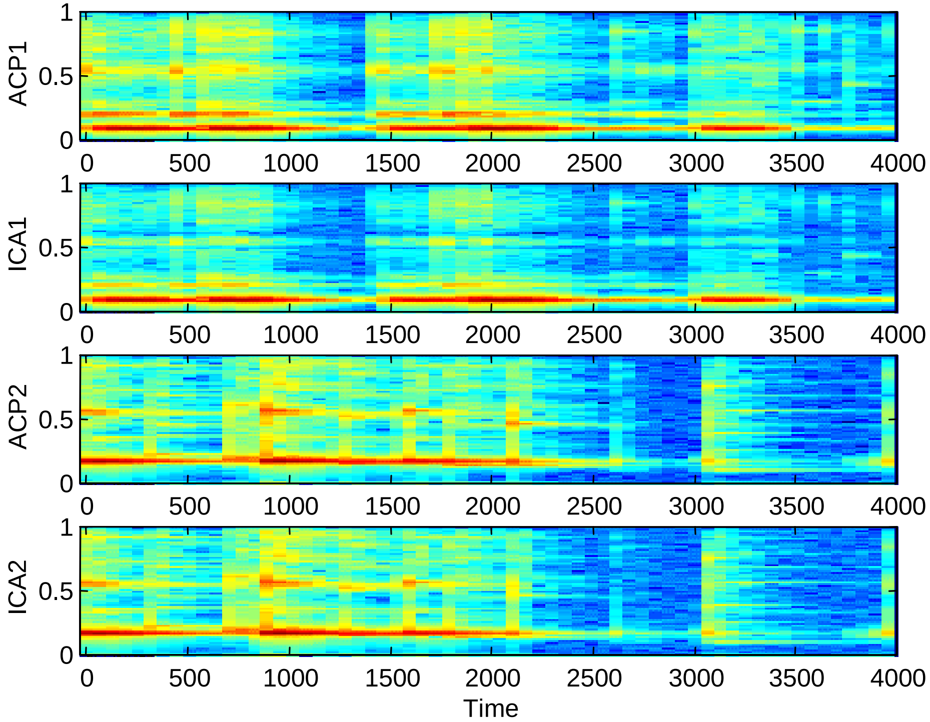


Detailed example II (Music)

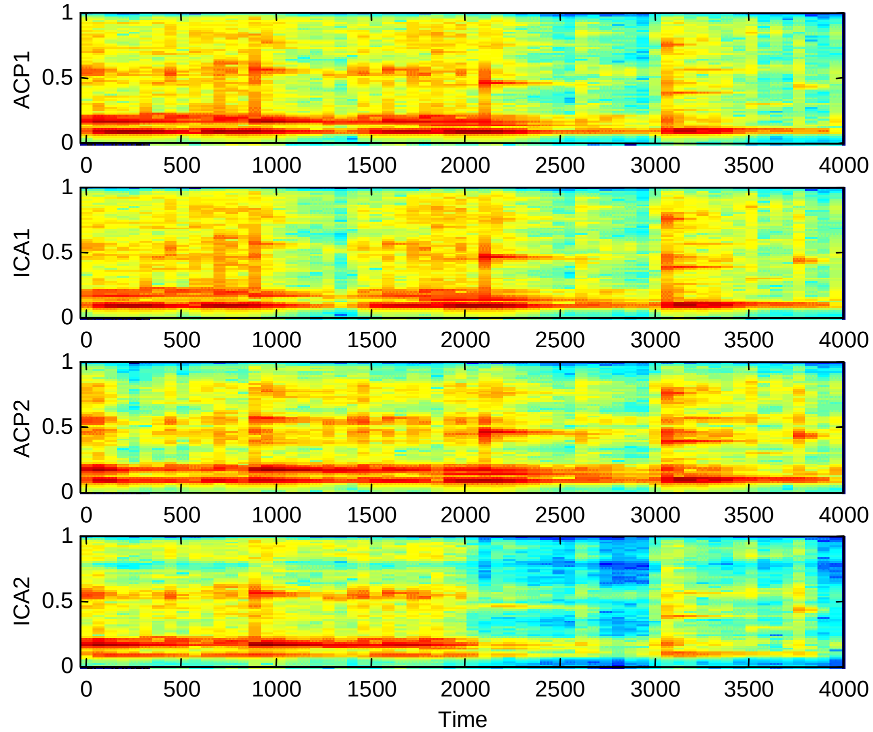
Mixture

ANGK-BALO (66)

REFERENCES



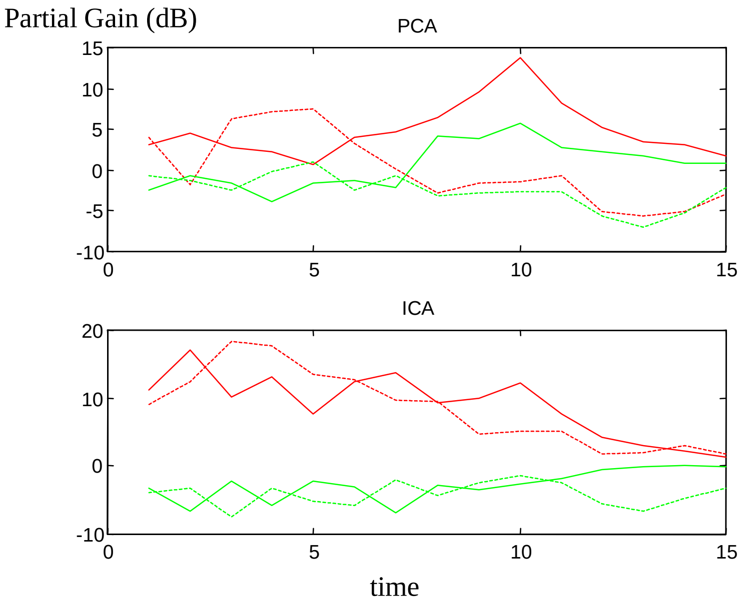
OUTPUT



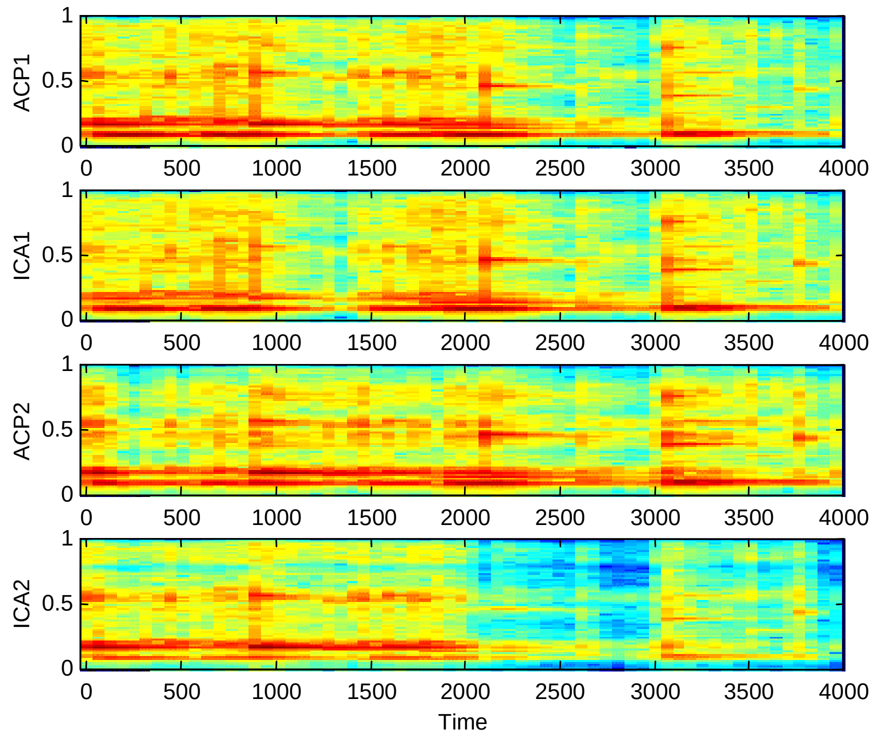
Detailed example III (Music)

Mixture

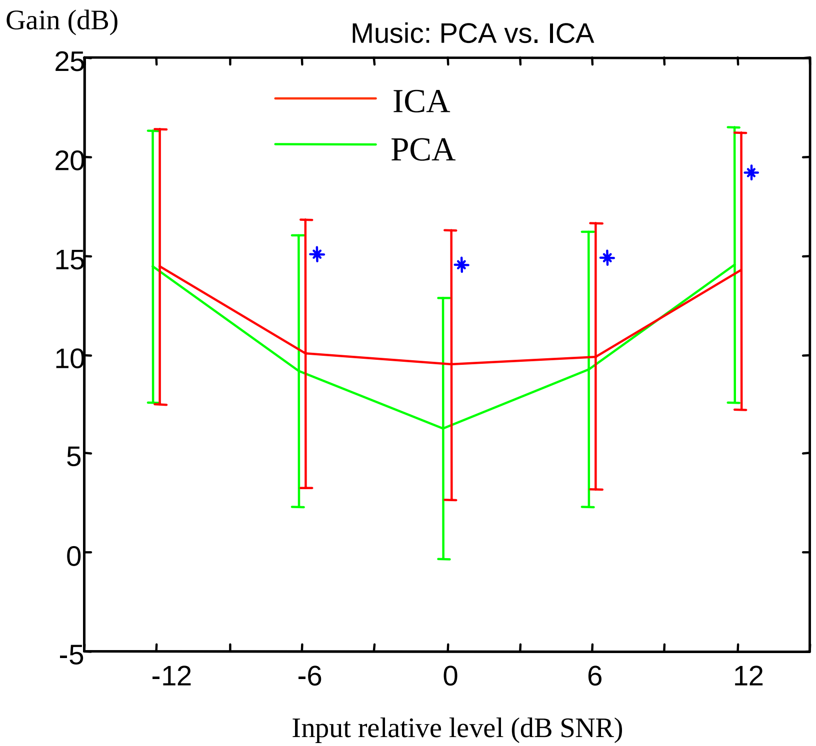
ANGK-BALO (66)

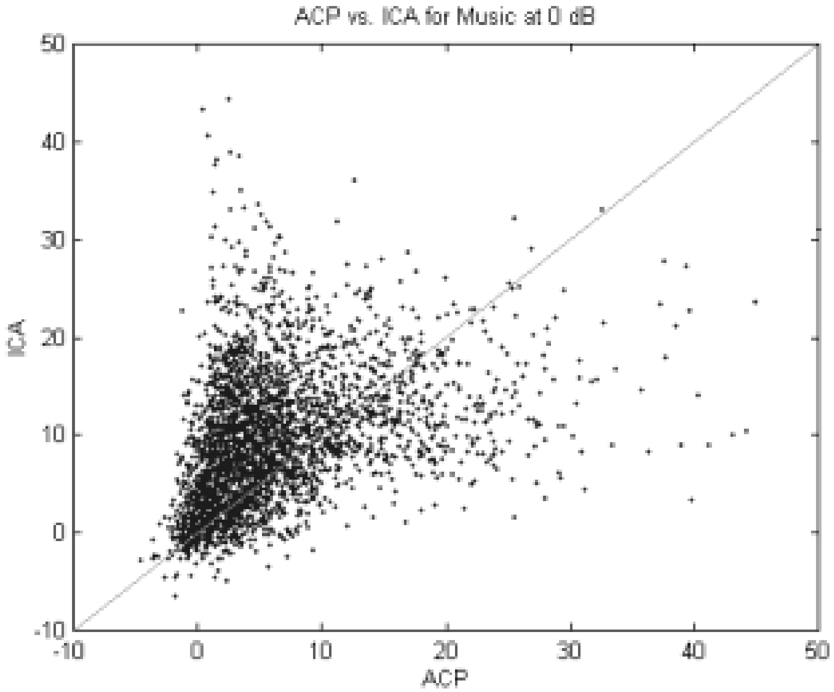


OUTPUT



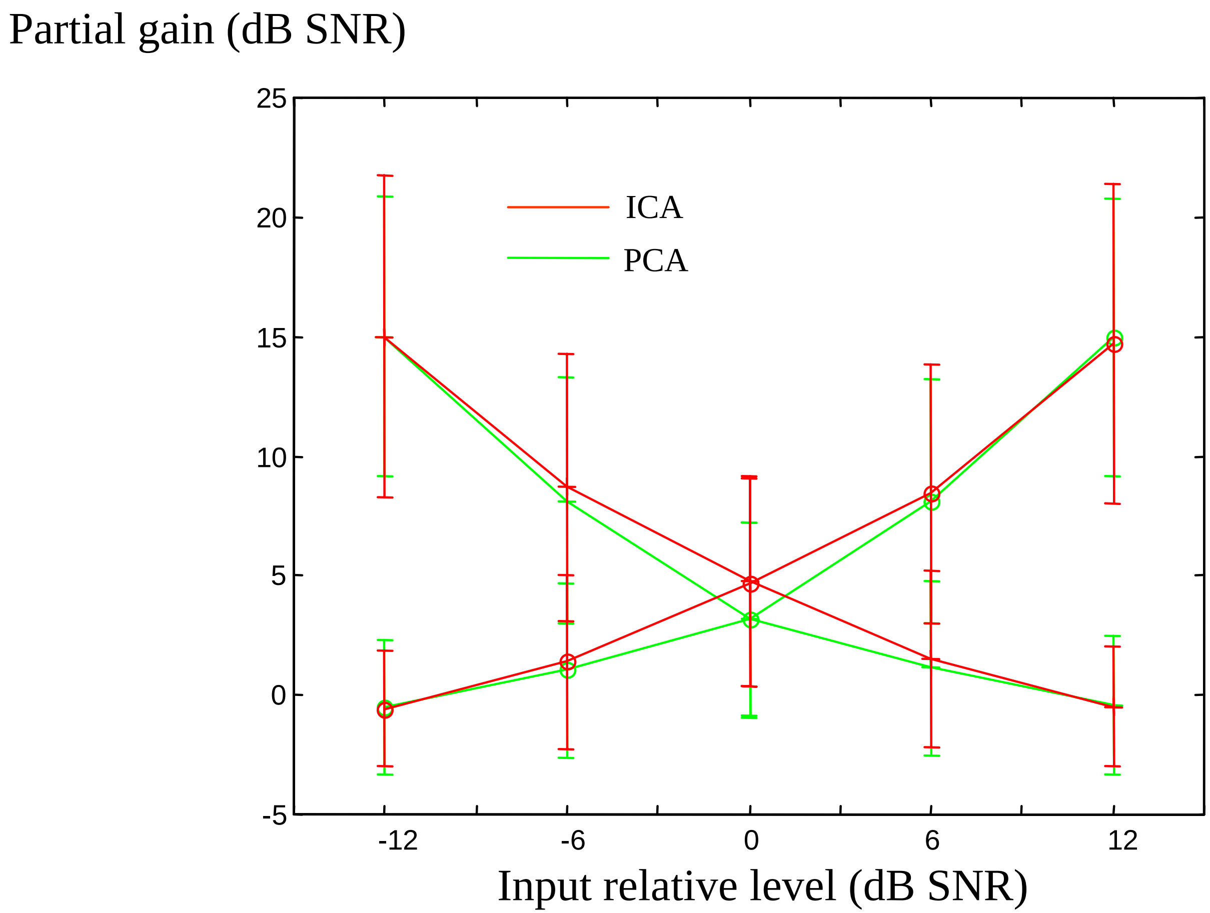
Gain (Music)





3244 mixtures * 5 Rel. Levels

Gain (Music)

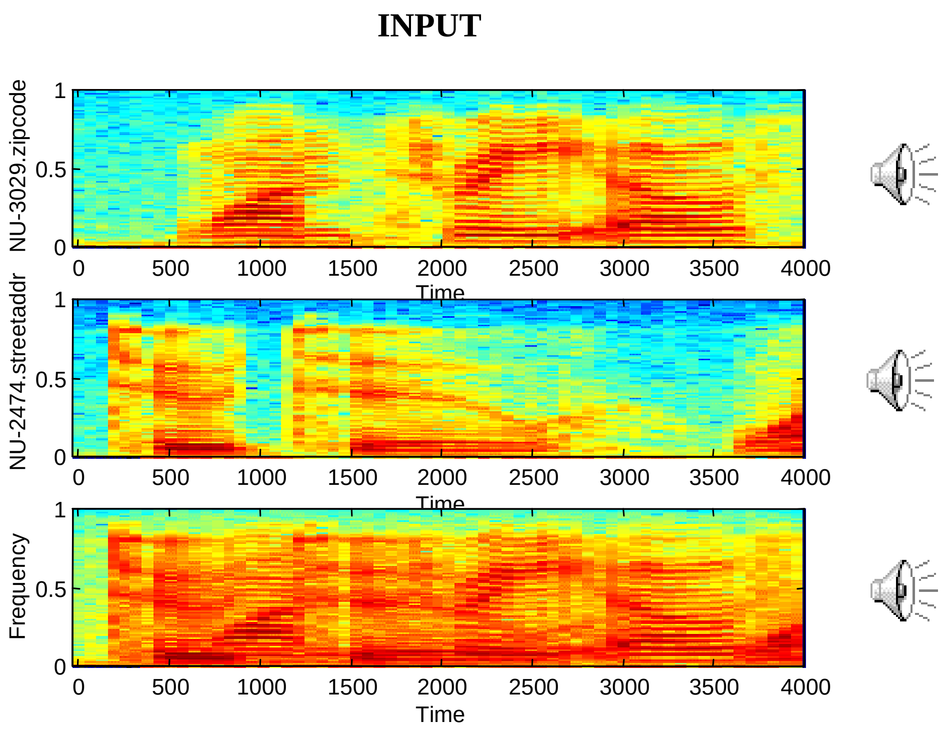


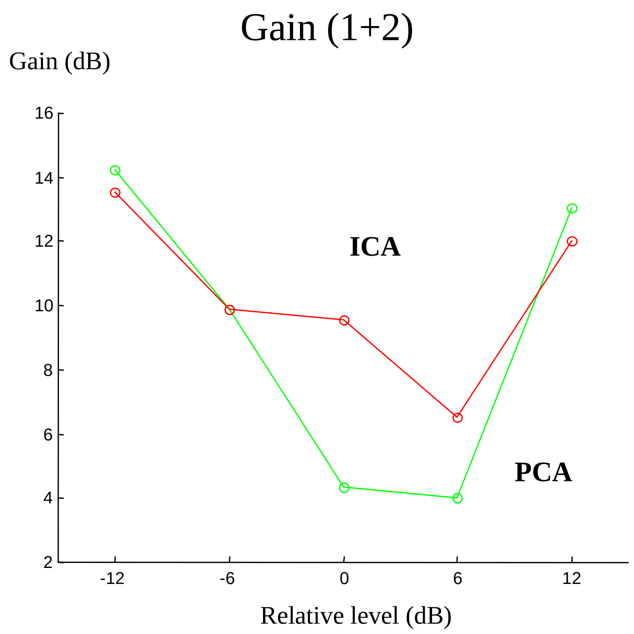
3244 mixtures * 5 Rel. Levels

Detailed example I (Speech)

Mixture

ST-NB95 (1620)



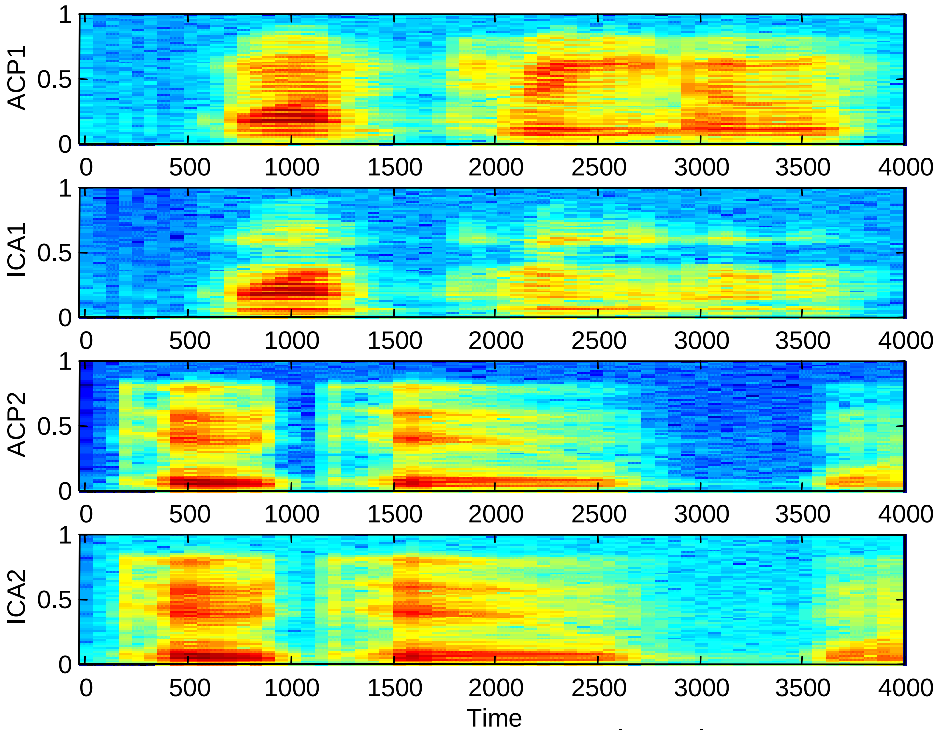


Detailed example II (Speech)

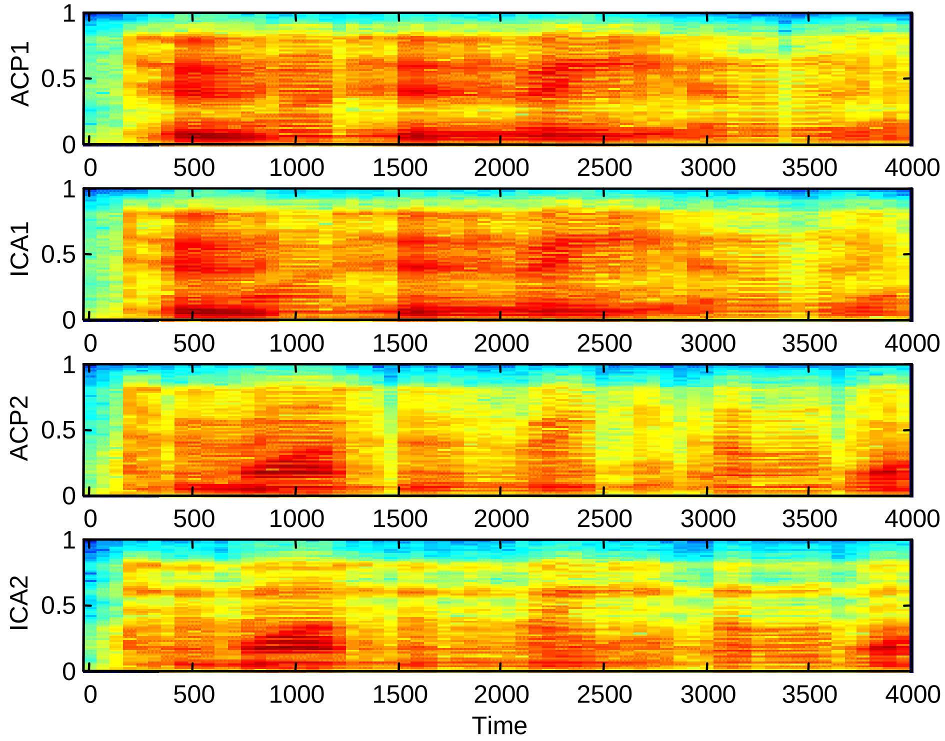
Mixture

ST-NB95 (1620)

REFERENCES



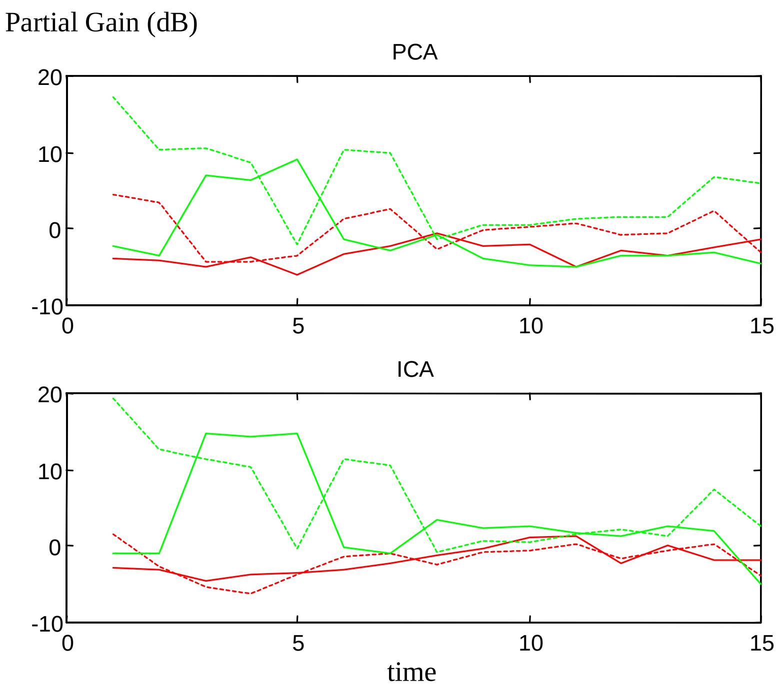
OUTPUT

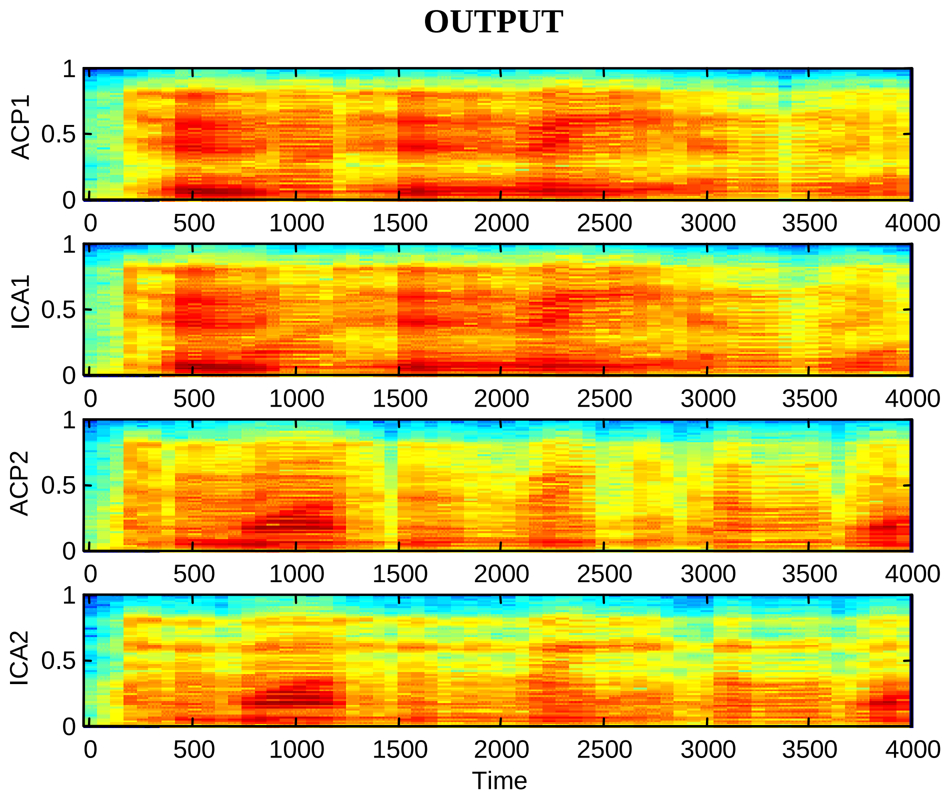


Detailed example III (Speech)

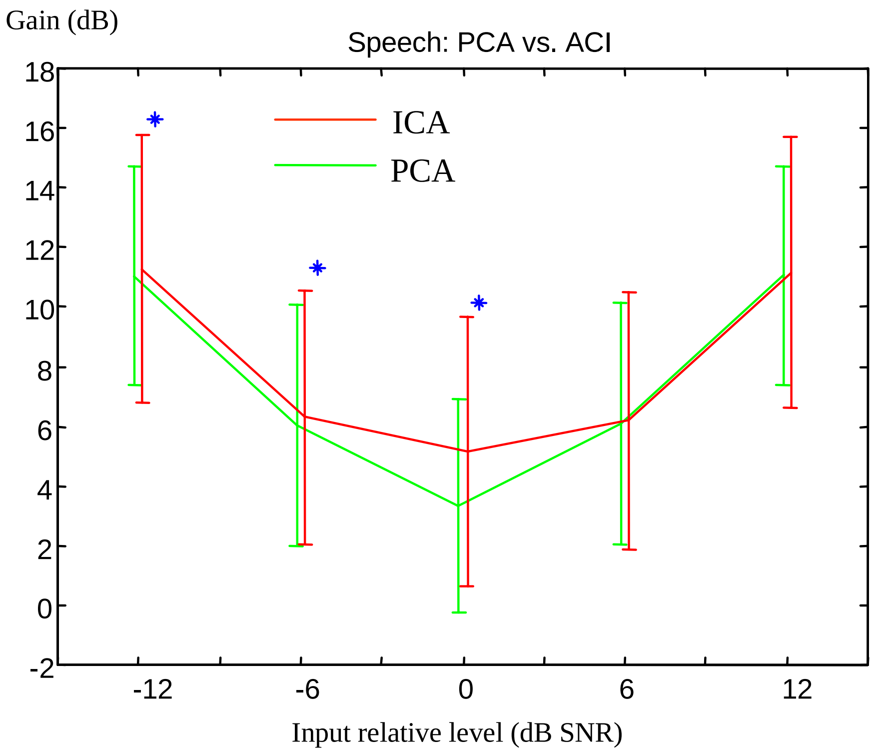
Mixture

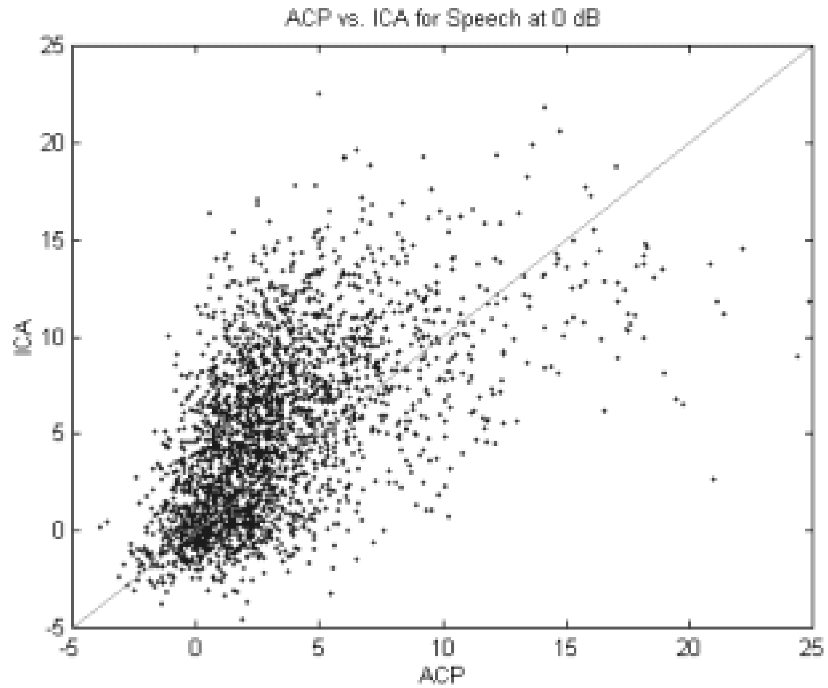
ST-NB95 (1620)





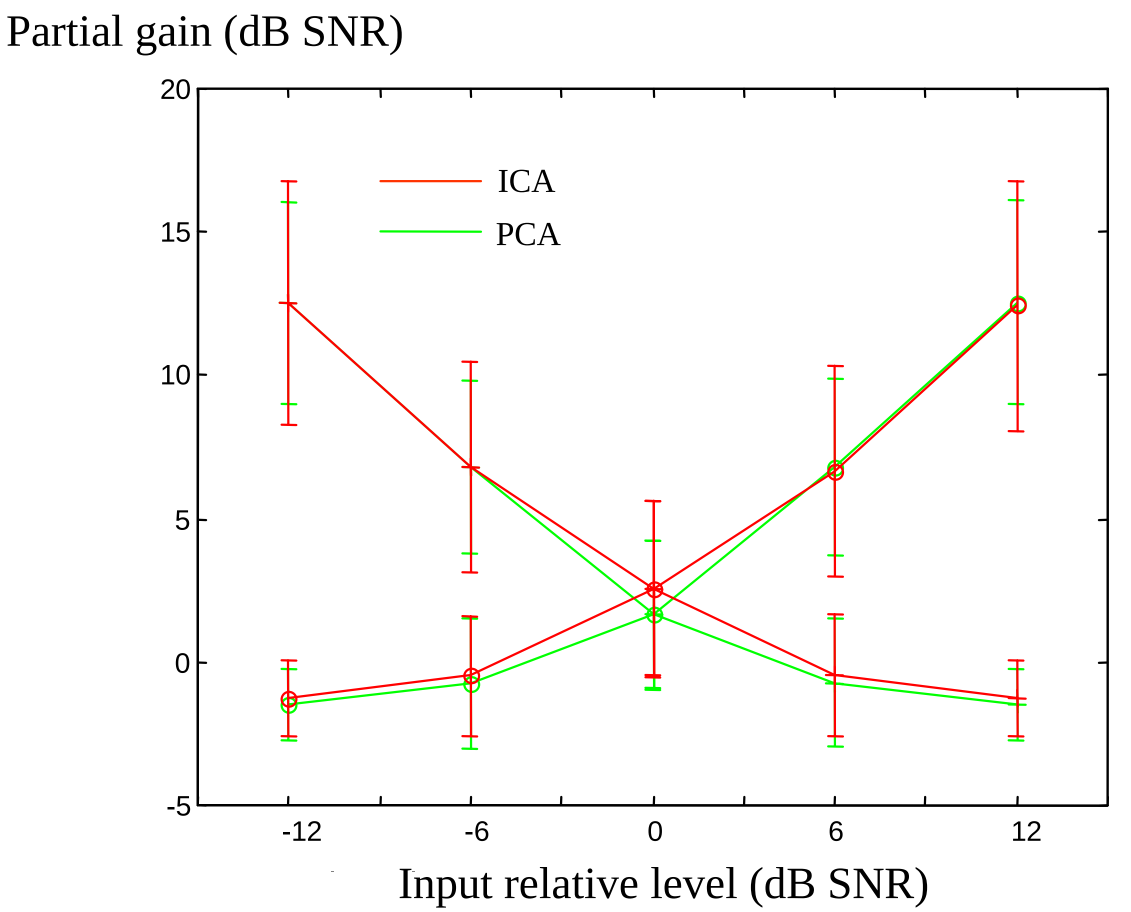
Gain (Speech)





2633 mixtures * 5 Rel. Levels

Gain (Speech)



2633 mixtures * 5 Rel. Levels

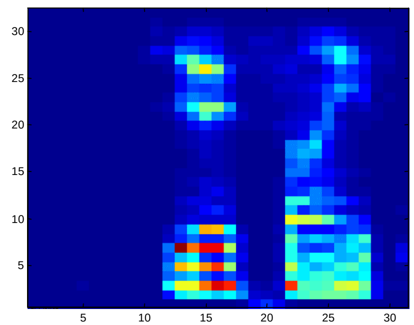
Perspective: set-up of a front-end for robust recognition

First step: adaptation (AV-NB95)

Isolated speech

Spectrogram

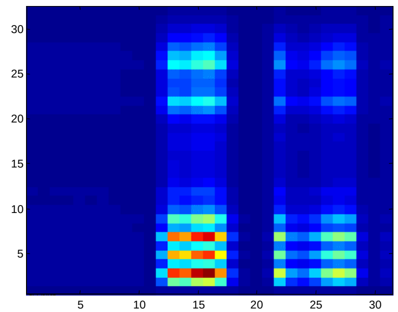
32



1 second

PCA/ICA

1st component



Resynthesis

PCAref

ICAref

TRAINING SET

Perspective: set-up of a front-end for robust recognition (II)

Second step: recognition