Test of several external posterior weighting functions for multiband Full Combination ASR

Hervé Glotin[1,2]
Frédéric Berthommier[1]
Laurent Varin[1]

[1] Institut de la Communication Parlée /INPG,
46 Av. Félix Viallet, 38031 Grenoble Cedex, France

[2] IDIAP Institute of Perceptual Artificial Intelligence,
Simplon 4, 1920 Martigny, Switzerland

Introduction

Preliminary experiments lead us to expect that multistream speech recognition can be made more robust with the inclusion of estimates of the stream's reliability (Hermansky, 1996, Bourlard and Dupond, 1996, Berthommier et al., 1998).

Several techniques are available for generating subband speech reliability estimates. Reliability or weighting factors are empirically estimated, or true probability values $P(SNR_i > T_i \mid R_i)$.

We will study the integration of these measures in the FC formalism, when the estimation is performed before recognition, using the acoustic data only.

The full combination ASR

The full combination (FC) approach allows the weighting of the posterior values estimated locally in the time frequency representation, according a speech reliability measure:

$$P(q_k|X) = \sum_{j=0}^{2^d-1} P(q_k, L_j|X) \simeq \sum_{j=0}^{2^d-1} P(q_k|x_j).P(L_j|X).$$

Using the Bayes' rule we estimate the posteriors of each stream x_j:

$$P(x_j|q_k) \simeq \prod_{i\in J} P(x_i|q_k), \text{ then } :$$

$$P(q_k|x_j)\frac{p(x_j)}{p(q_k)} \simeq \prod_{i\in J} P(q_k|x_i)\frac{p(x_i)}{p(q_k)}$$

$$P(q_k|x_j) \simeq \frac{\prod_{i\in J} P(q_k|x_i)}{p^{|J|-1}(q_k)} \cdot \Theta$$

If we assume that the best estimates of posteriors is produced by the cleanest data, then the weighting factor for the stream x_j corresponds to the probability that the "data of x_j better matches the data of the training set" (this event is called L_j):

$$P(L_j|X) = \prod_{i\in J} P(C_i|x_i). \prod_{i\notin J}(1 - P(C_i|x_i)).$$

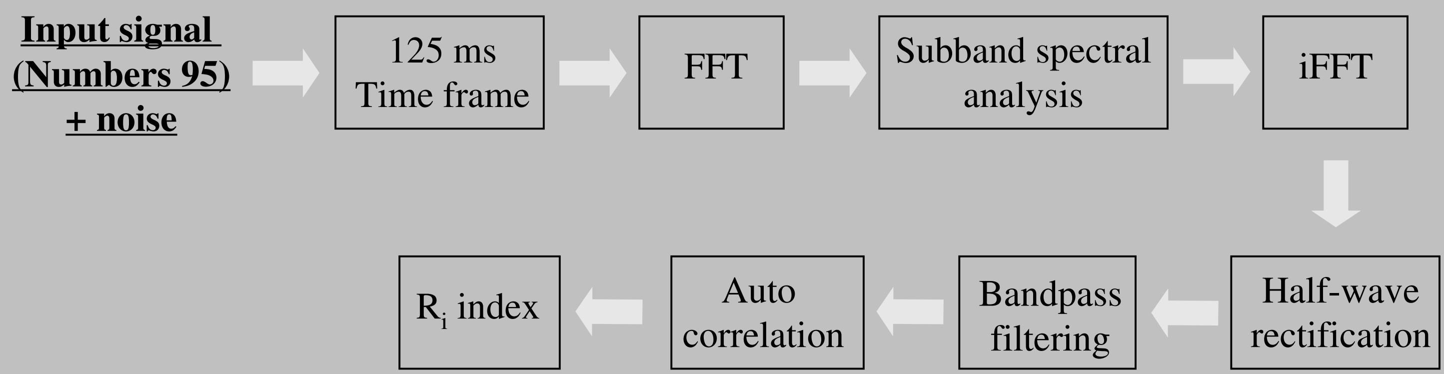
Then, the term $P(C_i)$ is a local speech reliability measure which is determined from the acoustic data x_i (i.e., this is external). Using binary values for $P(C_i)$ (e.g., after a hard decision process about speech reliability), the effect is, for each time frame, to collapse the set of combinations of sub-bands into a single combination and to perform a marginal partial recognition.

RESPITE

SPEAR

Probabilistic estimation of speech reliability from the harmonicity cue (I)

Since most of the speech segments are voiced, we use a method exploiting the harmonicity of speech to derive a probabilistic weight $P(C_i)=P(SNR_i>T_i \mid R_i)$ from the harmonicity index R_i.

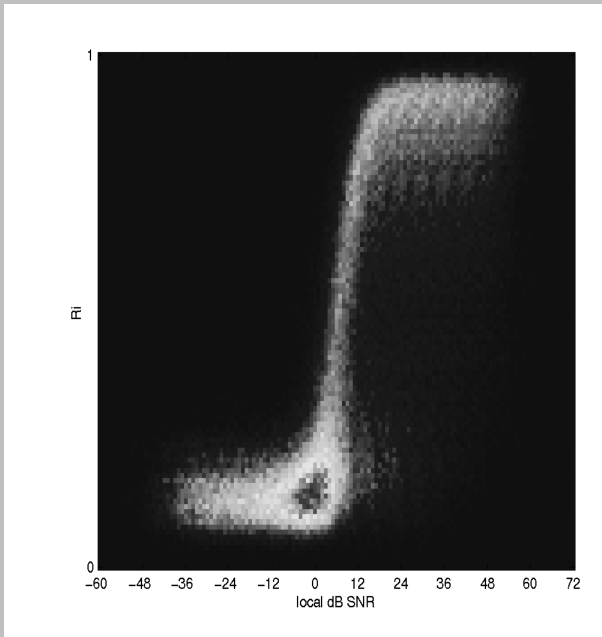


RESPITE

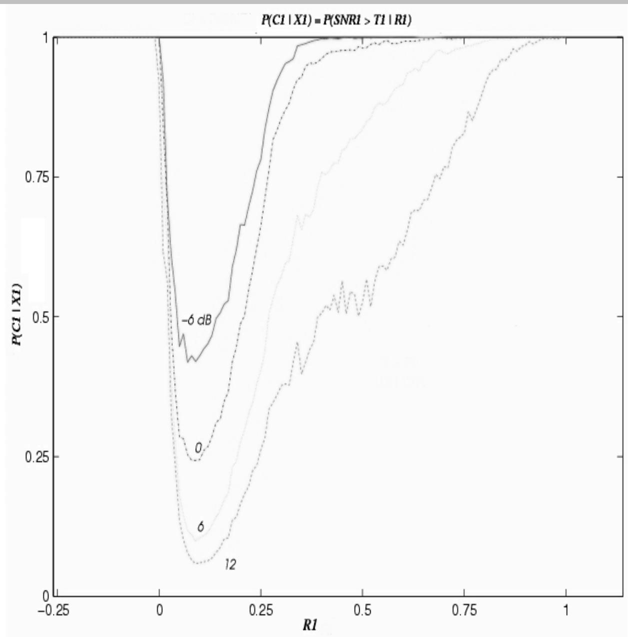
SPEAR

Probabilistic estimation of speech reliability from the harmonicity cue (II)

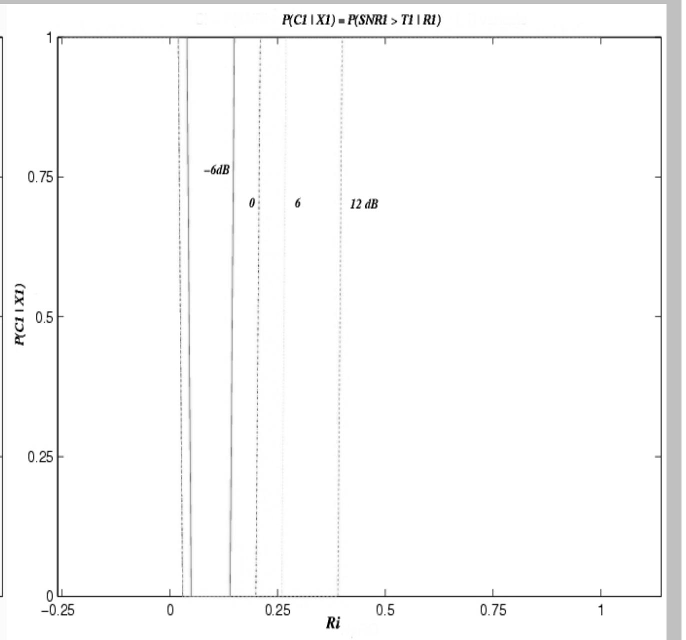
2D HISTO



SOFT (Spro)



HARD (Bpro)



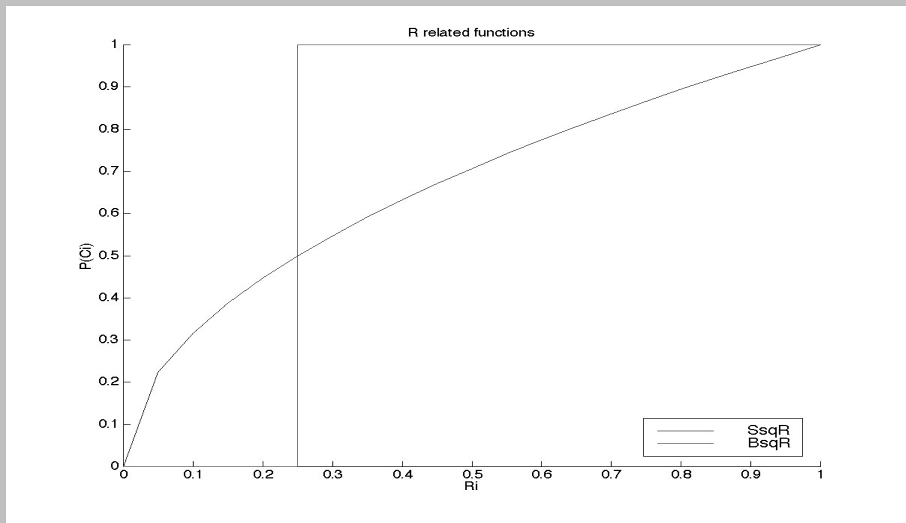
$$\text{if } P(C_i) > 1/2 \text{ then } P(C_i) = 1 \text{ else } 0.$$

Empirical weighting functions (I)

The term $P(C_i)$ is estimated locally from acoustic data x_i, as an empirical weighting factor normalised between 0 and 1, since it is correlated with the speech reliability.

Estimated from the harmonicity index R_i:

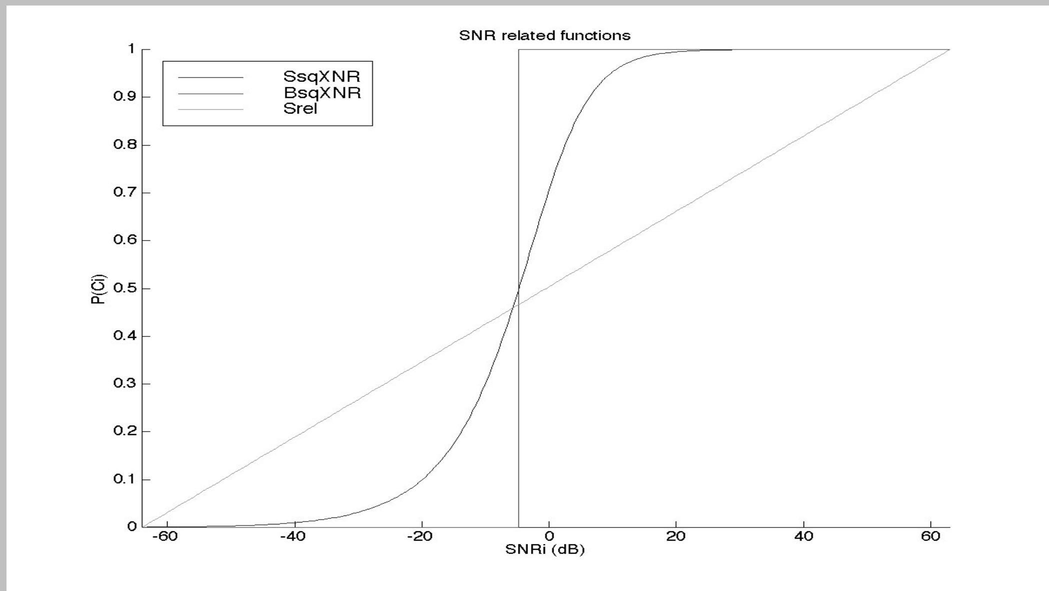
SsqR: square root of R_i
BsqR: thresholding of SsqR at R_i=0.25



Empirical weighting functions (II)

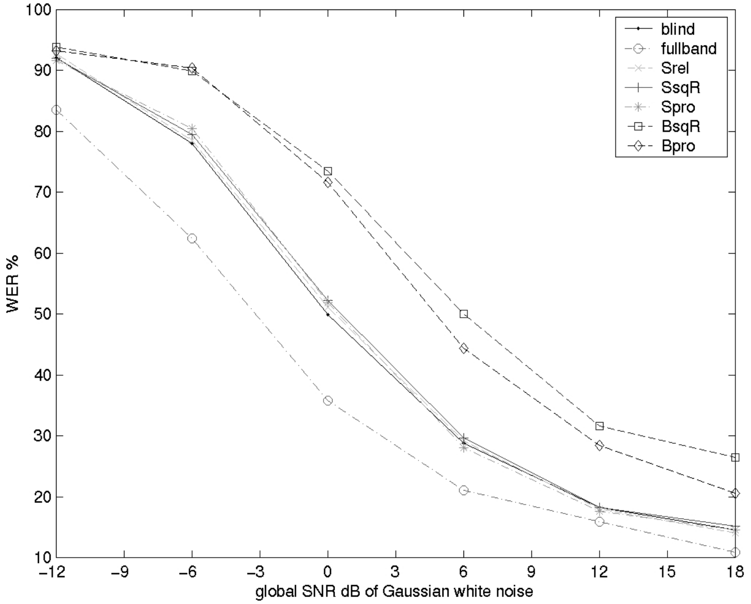
SNRi given:

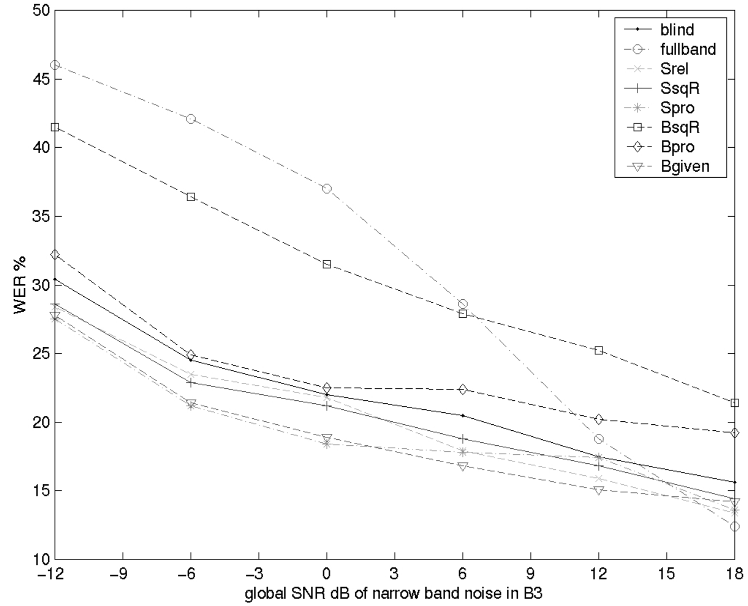
SsqXNR: sqrt of the local $XNR_i=1/(1+1/(10\text{^}(SNR_i/10)))$
BsqXNR: thresholding of SsqXNR at about SNR_i=-5dB
Srel: relative SNR_i= $(SNR_i - SNR_{imin})/(SNRimax - SNR_{imin})$ where
SNR_{imin} and SNR_{imax} are determined from SNR_i statistics (Hirsch, 1995).



Recognition test (I)

We tested two narrow band noises (narb1 and narb3) 300 Hz wide and centred in different subbands (1 and 3) and a non stationary narrow noise (siren) composed of periodic sequences [1,2,3,4,4,3,2,1] of respective noisy subband number.
As broadband noises, we used Gaussian White Noise (GWN), Factory noise from Noisex and a Daimler car noise (DB-car).





Recognition test (II)

%WER were estimated by averaging scores obtained with 200 sentences repeated at 6 different SNRs from -12 to 18 dB by step of 6 dB (silence included).

All weighting functions are found to be sub-optimal comparatively to the constant weighting (blind), but a robustness of the FC for narrow-band noises is observed comparatively to the fullband.

NOISE	GWN	factory	DB-car	narb1	narb3	siren
fullband	38.2	37.8	33.7	26.7	30.8	90.6
blind	46.9	45.6	44.2	24.5	21.7	50.0
		S	O	F	T	
Srel	47.3	45.1	43.9	21.4	20.1	51.9
SsqXNR	60.1	57.4	55.9	23.6	20.4	62.8
SsqR	47.8	45.6	44.9	28.9	20.4	49.6
Spro	47.4	45.0	45.0	27.1	19.3	59.8
		H	A	R	D	
BsqXNR	61.1	58.1	57.2	24.4	20.6	64.7
BsqR	60.9	57.6	53.6	45.9	30.6	67.1
Bpro	58.1	54.9	51.8	34.8	23.6	66.1

Conclusion

We tested various approaches to weight the posterior values estimated locally in the time frequency representation.

The main observation is, whatever the weighting function, the result is not better than the use of a constant weighting value (blind). This effect does not depend on the nature (probabilist or empirical), neither on the support (estimated speech reliability or SNR given) of the variable which is introduced. Moreover, step functions, as binary decision functions, are worse than continuous functions, due to the largest difference with the constant function. So we conclude that the fusion of an external source of information cannot be realised well with the FC model in this condition, despite the apparent compatibility of the formalism with the introduction of such an information. This might be due to the J-rasta pre-processing which removes a lot of noise so SNR of the input signal and posteriors reliability might no longer be strongly correlated.

On the other hand, intrinsic weighting factors compatible with the FC formalism could improve the model. These factors could be derived from outputs of the recognition system as well as from a priori knowledge about the streams' reliability (e.g. the weighting of each subband or each stream by its reliability for speech recognition, knowing a priori that lower frequency subbands carry more phonetic information, or knowing that the larger streams are more reliable).