Speech enhancement and segregation based on the localisation cue for cocktail-party processing

Emmanuel Tessier and Frédéric Berthommier

Institut National Polytechnique de Grenoble, Grenoble
tessier@icp.inpg.fr

Abstract

This paper describes a method of using localisation information for separation of concurrent speech signals. In such a condition, although speech sounds overlap in time and frequency, their localisation is a specific cue which can be exploit. The study includes design and analysis of a double speech corpus of stereophonic recordings. We examine the statistical relation between the estimated TDOA in time/frequency regions, and the local relative level between the two sources *(*known *a priori*), varying the size of each time/frequency region. Using this observation, we propose a model of local estimation of the signal/noise ratio based on this cue, with the aim of reconstructing the components of the mixture by weighting the time/frequency domain.

1. Introduction

We know that the auditory system uses the localisation cue mainly thanks to interaural time and level differences. For humans, this helps the auditory scene analysis to discriminate one sound source to increase speech intelligibility in the cocktail party condition [Bronkhorst, 2000]. To model this effect, one way is to assign to each time-frequency region of the spectrogram a label, which is specific of the source, and then to group the regions belonging to each source, according to these labels (this is a segmentation of the spectrogram). In our model, for each time frequency region, we estimate as a label the Time Delay of Arrival (TDOA) which is related to the azimuth of the source. According to the properties of this estimation process, the delay which is retrieved is a function of the SNR: this is non-linear and close to the delay of the energetically dominant source [Tessier et al., 1999]. So, knowing the current localisation of the target sources, this is a way to estimate the SNR, or the relative level (RL) of two sources, locally in the time-frequency representation. This continuous information is then used to reconstruct a target source and to segregate different sources by weighting. This method is similar to a Wiener filtering [Bodden, 1993] and this is an improvement of the discrete segmentation of the spectrogram provided by the labelling. Comparatively to the model proposed by Bodden, we have a simplified and motivated pre-processing stage, composed of a few frequency channels. This is in order to improve the SNR estimation and to study the trade-off between the accuracy of the SNR estimation and the filterbank resolution required for the decomposition of a binary mixture of speech sources overlapping in frequency.

In this paper, we first describe the design and the analysis of the StNumbers95 database. This permits us to determine the relationship between estimated TDOA and the local relative level, known *a priori*. Fitting the statistical data obtained, we use the inverse relation to estimate the relative level from an estimated TDOA and a target delay corresponding to the desired source. This one is used in a model of segregation of several speechs signals by weighting the mixture spectrogram.

2. The StNumbers95 database

In order to simulate a cocktail-party situation, we recorded at ICP a stereo corpus (StNumbers95) based on sentences of the OGI Numbers95 database. These sentences were recorded using two loudspeakers and two microphones (static positions left and right), first in isolation (left or right loudspeaker), and then mixed by pairs (one sentence per loudspeaker is played). A more precise description of the set-up can be found in [Tessier et al, 1999]. Thus, we can compute a precise reference of the a priori Relative Level (RL) existing between the two sources.

A preliminary analysis of the corpus consisted in

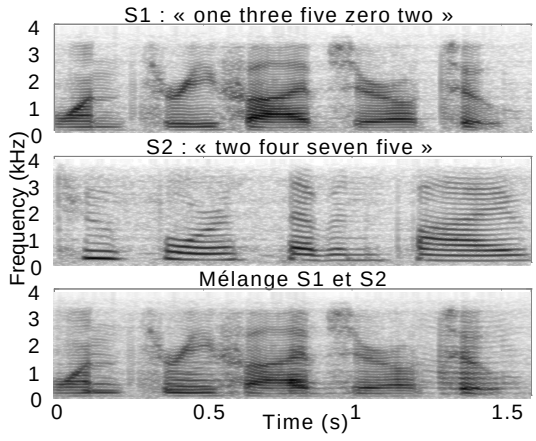

Figure 1: Extract of pair of sentences from the StNumbers corpus. The spectrograms corresponds to the sentences recorded in isolation from loudspeaker left (S1), right (S2) and the mixture (S1 and S2).

Table 1: Recovering duration of mixtures of sentences for different types of signals: silence, voiced and non-voiced speech.

Sentences	Silence	Voiced	Non-voiced	Total
Silence	11	12	4	26
Voiced	12	**33**	**10**	53
Non-voiced	4	**10**	**4**	18
Total	27	55	18	100

quantifying the time overlap between the sentences. Using frames of 25ms (half-overlapped Hanning windows), and the phonetic transcription and labelling of the sentences, we can detect regions of silence, voiced or non-voiced-speech. Table 1 shows the results obtained over the 613 pairs of sentences (representing more than 17mn of recordings). For the speech/speech condition, the overlap ratio (without silences) is relatively important 64% compared to 41% for a more natural database: ShATR [Crawford et al., 1994] (using a similar technique, [Tessier, 2001]).

Figure 1 give an example of two sentences recorded in isolation and simultaneously. Although the time overlap is important, there exist unmasked frequency regions (for example around 0.5s: “three” + “four”).

Although the global relative level between the two sentences is 0dB, the relative level increases while evaluated locally (see figure 2 for an example of time/frequency decomposition, frame of 50ms and 8 sub-bands).

To quantify this increase we have evaluated the mean of each local relative level for different sizes of the time/frequency region. The time decomposition consists in half-overlapped Hanning windows of different duration : 20ms, 40ms, 80ms and 160ms. For the frequency decomposition, we use sub-bands formed by grouping adjacent filter from a filterbank having 24 filters. Each filter is composed of half-windows which are complementary with their adjacent filter. This grouping principle allows us form 1, 2, 3, 4, 6, 8, 12 and 24 comparable filters (see figure 3, for an example of 4 sub-bands).

The figure 8 (diagram A) shows the results of the mean of the local relative level for the 32 different

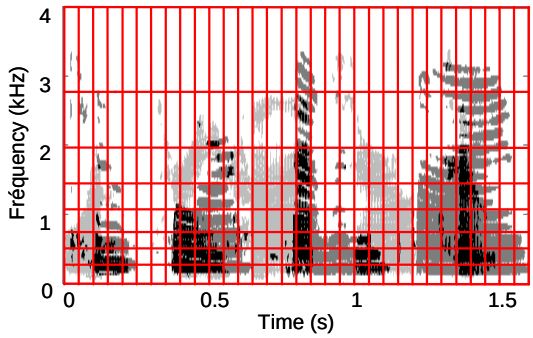

Figure 2: Example of decomposition of the time/frequency domain in 8 sub-bands, with frames of 50ms.The spectrogram of mixture is obtained by thresholding isolated spectrograms to detect components of each sentence (light and dark gray) and interfering regions (black).

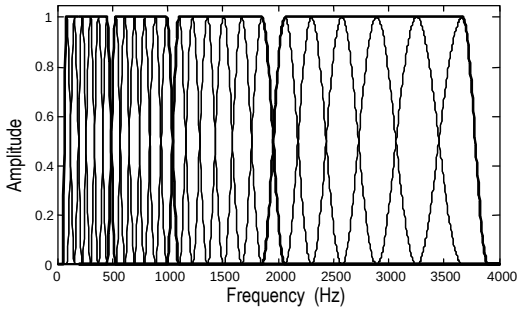

Figure 3: Example of grouping of 4 adjacent filters from a 24 filterbank to form 4 sub-bands. The center frequencies are equally spaced in Bark scale.

configurations over all 613 sentences of StNumbers. The relative level increases with a diminution of the size of the time/frequency region with a plateau for number of more than 4 sub-bands.

3. Time/frequency analysis on the database

Figure 4 shows the schematic diagram of the model. The TDOA estimated in each time/frequency region can be used as source detector. This can be extracted using the cross-correlation between the temporal envelopes of the two microphone signals. Half-rectifying and low-pass

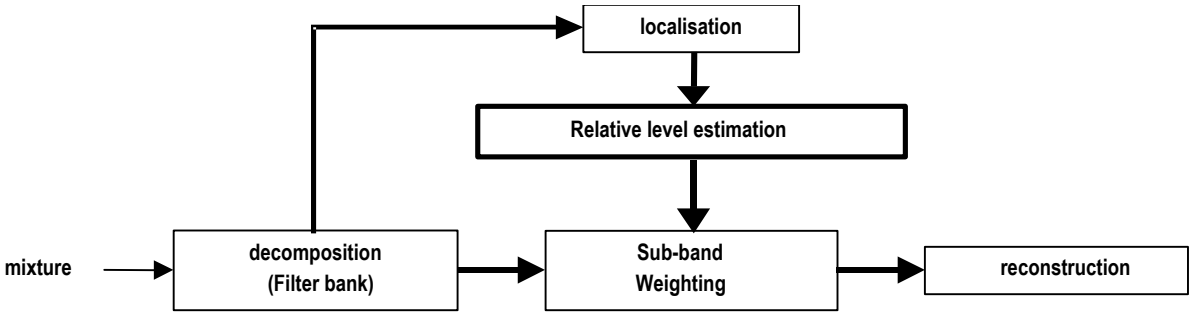

Figure 4: Schematic diagram of the reconstruction model. The signal is first decomposed into sub-bands. Then, in each time/frequency region, we use the localisation cue (through the TDOA detected) to estimate the local relative level. This one is used to reconstruct the weighted spectra.

filtering the sub-band wave performs this envelope detection. The position of the maximum of the cross-correlation function depends on the azimuth of the sources, the type of sound source (bandwidth and coherence), and the relative level between the sound sources (cf. [Tessier, 2001], chp. 3).

The estimation of the TDOA and the knowledge of the *a priori* relative level over all 613 sentences, for each time/frequency resolution, allows us to analyse statistically the relation between the TDOA estimated and the relative level. Figure 5 shows the distribution of the estimated TDOA and the relative level in the case of an analysis with frames of 40ms and 8 sub-bands. At high relative levels, the estimated TDOA corresponds to the target delays (loudspeaker left or right) and TDOA distributions are peaky. Nevertheless, for low relative levels (around ±6dB), the TDOA distributions are flatter (strongly dependent of the number of sub-bands).

In order to compare the different configurations, we consider that the standard deviation of the estimated TDOA for relative levels > 12dB is representative of the global distribution of the TDOA for the considered configuration. Figure 8 (diagram B) shows the comparison results of the mean of the estimated TDOA standard deviation for different time/frequency resolutions. We observe a great increase of this index, and then a loss of accuracy of localisation, for a number of sub-bands superior to 4.

The TDOA estimates are converted in RL estimates thanks to a non-linear fit of the data (see figure 6). The resultant values are used to enhance the desired signal by a linear weight, which is applied in the same regions of the FFT spectrogram. Finally, the signal is re-

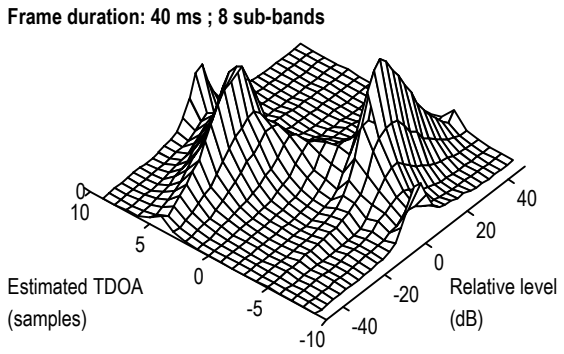

Figure 5: Bi-dimensional distribution of TDOA as afunction of the local relative level known a priori.

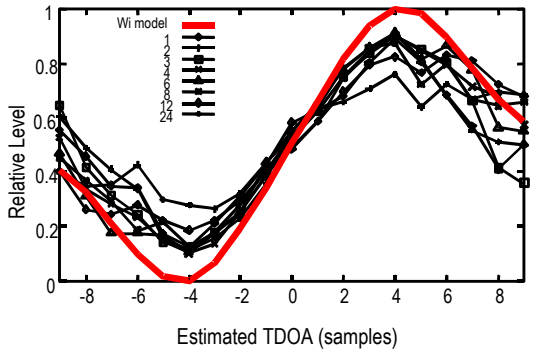

Figure 6: Mean curves of relative level (values $\hat{I}$ [0; 1]) between the two sentences for different numbers of sub-bands as a function of the estimated TDOA. The weighting curves (in bold) fit the statistical data.

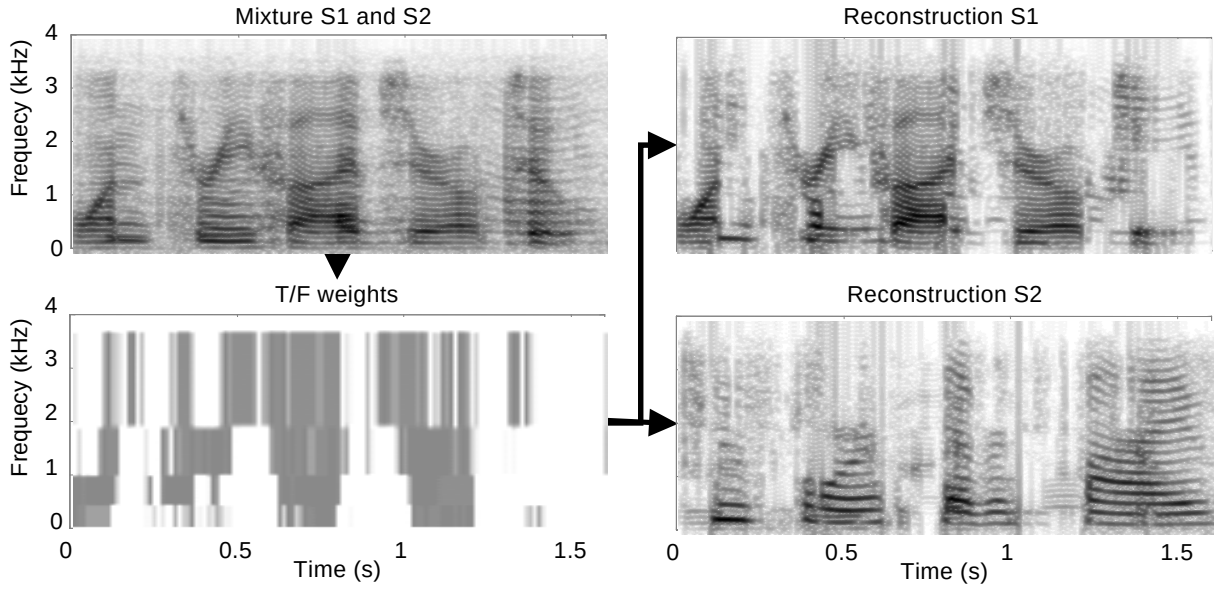

Figure 7: Example of separation of two sentences from a mixture using our model. The lower bottom image represents the weights of each time/frequency (T/F) region. The weight is derived from the local TDOA estimation, as a Relative level estimation. This representation is use to mask the mixture spectrogram to reconstruct the sentences S1 and S2 (using the complement weights).

synthesized by inverse FFT. Figure 7 shows different stages of the process of segregation of two sentences.
With a small number of sub-bands, we sharply decrease the frequency resolution of the SNR estimation, and the possibility to weight differently close frequency components belonging to different sources. But, on the other hand, we improve the dominance effect and we increase the accuracy of the SNR estimation. Then we improve globally the gain of the filtering. In order to find a good compromise, we vary the size of the time/frequency regions and we apply a quantification of the enhancement process.

4. Results

In order to establish the efficiency of the enhancement method, we use two indices. The first one is the distance between the output spectrogram and the reference recorded in isolation, which we name Reconstruction Accuracy (RA). This index corresponds to the energy ratio between the original signal and the difference between the original and the reconstructed signal.
Results of the comparisons between configurations are shown in figure 8 (diagram C). We observe a bell shaped relationship between the number of sub-bands, with a maximum at 4 sub-bands. Increasing the number of sub-bands (i.e. the frequency resolution) leads to an increase in the standard deviation of the TDOA estimation and then a lower accuracy of the RL estimation. The best performances are obtained with a low frequency resolution (4 sub-bands). So, a Wiener technique using such an acoustic estimate is improved when the weighting is applied in wide sub-bands, and not at the (fine-grain) spectral component level (e.g., at the FFT bin level, or for each channel of a 32-channel filterbank).

5. Conclusions

This technique of enhancement and its evaluation could not be realised without the use of specific features of our database StNumbers95. This allows us to compare the enhanced sources with the same sources recorded in isolation. In our simulations, we varied the main factors of the enhancement process. Our experimental set-up allows ideal conditions for such an analysis: the recordings were carried out in an Antioch room and the sources and microphones were static and spatially distinct. For different conditions (echoic room, moving sources), we could expect some degradation to be analysed better in further experiments. Compared to a blind separation technique applied on the same database [Choi et al., 2001], this technique has lower performances, but it theoretically allows us to separate more than two sound sources in dynamical conditions because it operates at the short time-frame level.
This work is supported by the EEC contract LTR RESPITE.

6. References

[1] Bodden, M., ``Modeling human sound-source localisation and the cocktail-party effect", *Acta Acustica*, 1(1), 1993, p.43-55.
[2] Bronkhorst, A., ``The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker condition", *Acustica*, 86; 2000, p 117-128.
[3] Choi, S., Hong, H., Glotin, H., Berthommier, H., ``Multichannel Signal Separation for cocktail party speech recognition : A dynamic recurrent network", submitted in Neurocomputing, special issue on Blind Signal Separation and ICA.
[4] Crawford, M. D., Brown, G. K., Cooke, M. P., Green, P. D., Design, collection and analysis of a multi-simultaneous-speaker corpus, Proc. of the Institute of Acoustics, tm. 16, 1994, p. 183-190.
[5] Lehn, K., ``Modeling binaural auditory scene analysis by a temporal fuzzy cluster analysis approach", *IEEE WASPAAs*, *Mohonk Mountain House*, New York, 1997.
[6] Tessier, E., *Study of the variability of the localisation cue to characterize concurrent speech sources*, PhD thesis (in French), Institut National Polytechnique de Grenoble, 2001.
[7] Tessier, E., Berthommier, F., Glotin, H., Choi, S., ``A CASA front-end using the localisation cue for segregation and then cocktail-party speech recognition", *ICSP'99*, Seoul, Korea, 1999, p. 97-102.

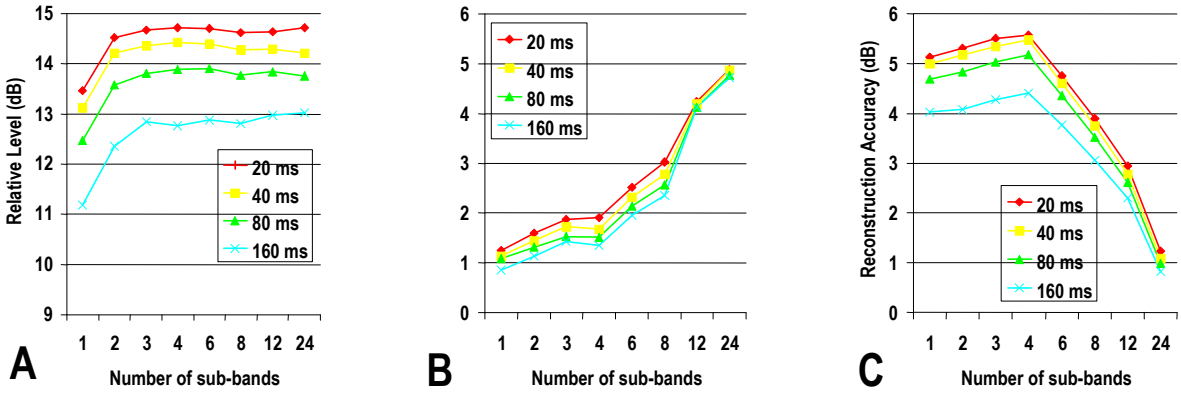

Figure 8: Results of the analysis of the 613 pairs of sentences from StNumbers, varying the size of the time/frequency regions. A) Mean of the local relative level (dB); B) Mean of the standard deviation of the estimated delay (samples); C) Mean of the reconstruction accuracy (dB).