

A CASA FRONT-END USING THE LOCALISATION CUE FOR SEGREGATION AND THEN COCKTAIL-PARTY SPEECH RECOGNITION

Emmanuel TESSIER^{*}, Frédéric BERTHOMMIER^{*}, Hervé GLOTIN^{*,#}, Seungjin CHOI⁺

^{*}ICP/INPG
46, Av. Félix Viallet
38031 Grenoble CEDEX, FRANCE
(tessier, bertho)@icp.inpg.fr

[#]IDIAP
Rue du Simplon, 4
1920 Martigny, Switzerland
glotin@idiap.ch

⁺Chungbuk Nat. Univ.
48, Gaesindong
Cheongju, Chungbuk, KOREA
schoi@engine.chungbuk.ac.kr

ABSTRACT

We propose and test a cocktail-party recognition technique based on segregation applied before recognition. This CASA front-end uses the TDOA (Time Delay Of Arrival) evaluated within subbands in order to determine the Relative Level (RL) of two competing speech sources. To perform the evaluation of the model, we have recorded a stereo database ST-NB95 from the mono Numbers95. This is composed of binary mixtures of sentences at 0dB, spatially placed left and right. With variation of frame duration and bandwidth, we quantify the accuracy of reconstruction of the original sources and we evaluate the recognition score on this new database. This work is a part of a triangular comparative study based on the processing of this database, and addressing the robust cocktail-party speech recognition paradigm.

1. INTRODUCTION

Humans are able to well recognise speech mixtures produced by two speakers simultaneously, and they are also able to identify speech in loud noise, this in a wide range of noisy conditions (stationary or not). Psycho-acoustical experiments have well characterised the "streaming effect" which is the perception of separated sources as an organised set of isolated "auditory objects". This motivates CASA (Computational Auditory Scene Analysis). A dominant hypothesis is this separation is the resultant of an auditory processing of the complex sounds (i.e., not only speech) based on their primitive characteristics: speech and background interference; possibly another speech source; are isolated in order to recognise isolated clean speech. The localisation cue, in our study, the TDOA, is classically used in array of microphones applications in order to enhance or segregate speech. Hence, this is mainly an engineering approach, but a physiologically motivated binaural cocktail-party processor able to segregate concurrent speeches is based on a similar principle ([2], [3]). This model realises a cross-correlation after filterbank decomposition. Then, the spectrum of each candidate source is filtered at the cross-correlation level according a place along the delay axis. To show the segregation and to quantify the recognition gain, these signals can be heard, or an ASR (Automatic Speech Recognition) can be applied separately on each filtered signal [4]. Therefore, one good property of this model is to allow recognition of simultaneous speakers.

We present here a similar model of source segregation, which is designed to be a front-end for cocktail-party ASR. This is a simplification of the binaural cocktail party processor [3]. We propose a new method to estimate the local SNR between the desired source and an interfering source. This estimate allows sharing of energy carried by each frequency channel between multiple sources. This was previously proposed to model double vowel segregation [11]. Then, the candidate

sources are reconstructed in order to feed a fullband HMM/ANN speech recogniser. The development, the adaptation and the evaluation of this model are based on the recording of a new database from an existing one: Number95 (NB95), which is used to evaluate robustness of ASR systems. We study the effect of the variation of important parameters. This is to optimise the model for the simplified-but-hard cocktail-party recognition task proposed together with this database. There is a double difficulty to tackle: the SNR is at 0dB and the interference is produced by another speech source. So, the conventional robust recognition approach, which is mainly based on the pre-processing principle, is not adapted to this task. The principle of the front-end is to segregate the sources before recognition and moreover, before applying any pre-processing other than a frequency decomposition of the signal.

One perspective is to compare properties and performances of different models. This modelling of the "streaming effect" has been adopted by CASA. Using an early segregation process allows a simplification of the architecture: there is apparently no "coupling problem" since ASR is fed sequentially with "clean" speech. The main motivation is to use a primitive information, which is not specific for speech signal, in a level of processing which is also not specific. So the process can be applied as a front-end to speech+noise recognition as well as to recognition of multiple speeches. But we remark that main parameters of the front-end can be tuned in order to optimise the segregation and then the recognition. On the other hand, the recognition level and its associated pre-processing have to be adapted to the segregation process, which is not complete, and which produces distortions and not the "ideal" clean speech. This tuning could depend on the task.

We remark that temporal blind separation methods proposed by [6] operate earlier, before the frequency decomposition, so the processing is fully sequential. Here, the principle is not to use a primitive cue to differentiate components of each interfering signal, but to ground the segregation on the statistical independence of the signals emitted by each source. This model has to be compatible with the mixture condition, which is not simply additive in ST-NB95.

2. RECORDING OF ST-NUMBERS95

The stereo database ST-NB95 is built from the monophonic NB95 in order (1) to spatialise the signal of NB95 in azimuth (2) to introduce a minimal distortion of the original signal and (3) to mix the signals of NB95 with a relative level controlled well. This is done in a soundproof an-echoic room by playing and recording the files of NB95 simultaneously with the same PC. The signal is played with *JBL Control-5* loudspeakers. For the acquisition of the signal, we have used *Panasonic WM-61A* miniature condenser

microphones and a *Soundblaster AWE64 type-1* card. The signal is pre-amplified before acquisition. The geometry of the set-up is shown Fig.1. The 40cm distance between the microphones has been chosen in order to have large arrival time differences. Arbitrarily, the source $s=1$ is the left loudspeaker and has a positive TDOA. The microphones are fixed on a wooden stick. This geometry is static for all the records of the database.

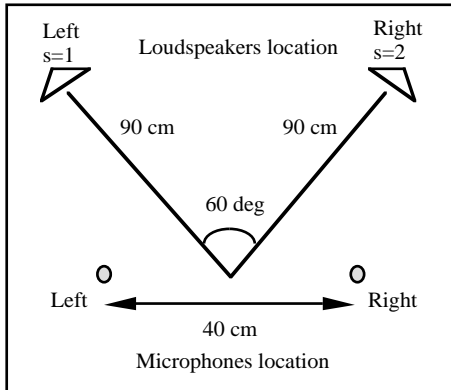


Figure 1: Set-up of the recording of ST-Numbers95. This is symmetric and "play and record" from Numbers95.

The full version of the ST-NB95 database is composed of sentences selected from the test set of NB95: (1) 2*613 sentences are played left or right in isolation. (2) 613 binary mixtures of paired sentences are selected for having the highest speech overlap as possible. (3) 2*613 isolated sentences are played left only added with pink noise. The part of the database we use in the present work includes (1) and (2). The global relative level is tuned at 0dB separately for each pair of sentences. The degree of overlap between paired words is high and this fits well the cocktail-party paradigm. The same sentences have been recorded in isolation, in exactly the same condition, this in order to have in hand a reference signal. So, we can evaluate the local relative levels of the same signals in the mixtures, and we can estimate the accuracy of the segregation methods by direct comparison.

A motivation of this recording is to set the background for a close comparison between three cocktail-party techniques based on different principles: blind separation front-end [6], CASA front-end (this work), CASA labelling [7]. These tackle the recognition problem in a cocktail party condition. This is complementary to the study of robust speech recognition in which a target speech is corrupted with non-speech. The baseline given by a normal recognition system is low, and this allows a sensitive measure of any improvement. This triangular comparative study opposes blind separation with CASA techniques since these are based on representations of the signal and extraction of primitive information, here the localisation cue. Another aspect of this opposition is the extraction of information about SNR or "data reliability", which is achieved with the CASA techniques we propose. A second opposition exists between front-end techniques and CASA labelling. The front-end principle consists to address an enhanced or a segregated signal to the recogniser, which is not deeply modified. On the contrary, the aim of CASA labelling is to inform the recognition level about the SNR existing in the peripheral time frequency representation. This new approach of robust recognition was promoted by partial recognition [8] and multistream [5] which are adaptations of the recognition level. The main property of these two models, and of their further improvement is to merge the acoustic data stream and information about "data reliability". Our CASA labelling

model ([1],[7]) is a proposal to extract and to formalise better the content of this "data reliability" stream. Finally, this properly contrasts the SNR estimation problem we address in the present work with reliability estimation we have done in ([1],[7]). In the present CASA front-end, the SNR is explicitly estimated in order to process the signal before recognition. On the contrary, the CASA labelling, which is addressed to the recogniser in ([1],[7]), is related to the SNR, but it is not a SNR estimation. This is more compatible with the new recognition approach we mentioned before.

The different principles, which are involved in these three models, are not exclusive, and we can expect an additive improvement. Moreover, each of them is more or less adapted to the recognition task (cocktail party, speech+noise) and to the involved cue (localisation, voicing, audio-visual, etc.). In the same way, after [1], we propose to compare these three models for the voicing cue. Hence, this is outside the scope of ST-NB95.

3. MODEL ARCHITECTURE

In classical recognition models, the time-frequency (TF) representation feeds the recognition process after a pre-processing step, which produces acoustic vectors. In these systems, a gain of robustness is expected from a better pre-processing able to filter out interfering signals and to regularise the speech features. An achievement of this point of view is the RASTA-PLP pre-processing [9]. Hence, this stage is highly specific to speech signals. On the other hand, the pattern matching stage can be adapted to a noisy condition. These two methods cannot well work whatever the interference: (1) the pre-processing method is expected to fail when the interference is another speech signal, and (2) the adaptation method is specific for a given noisy condition (e.g., in-car). So, these two kinds of models cannot tackle robust cocktail-party recognition without appeal of another source of information.

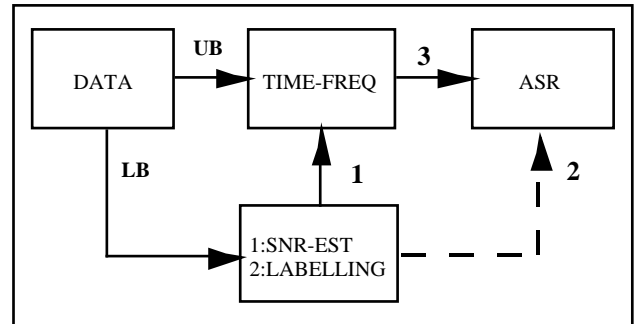


Figure 2: Principle of segregation and recognition. A SNR estimation process operating on the time frequency representation developed in the upper branch (UB) is performed in parallel in the lower branch (LB). This allows estimation (1) of the contribution of each source. The input (3) of the recognition process (ASR: automatic speech recognition) is a reconstructed spectrogram for each source. But practically, a waveform is re-synthesized and recognised separately for each source. In other models ([1],[7]), ASR is fed with CASA-labelling (2).

The pre-processing design is compatible with the incorporation of secondary features such as voicing or spatial localisation if we adopt the "enhance and recognise" scheme (Fig. 2, arrow (1)). We superimpose a parallel pathway to the "main speech processing route", which can include a PLP as

well as well as J-RASTA pre-processing in our application. These are designed to be specific to speech features. An extraction process of the secondary feature is performed in this second branch, which allows to enhance or to segregate the acoustic information, when the contribution of concurrent sources is removed. Remarkably, this operates in a different time-scale. Using the localisation cue, the scale is expressed in ms for the delay estimation. Moreover, this cue is not specific for speech signal. Note that, in comparison with ([1],[7]), (Fig. 2 arrow (2)), the parallelism is weak, since it is inherent in the frequency decomposition performed in the upper branch. The segregation/recognition principle is globally sequential.

Another property of this module is to assign an output TF representation (i.e., a spectrogram) to each candidate source, this before recognition: the two spectrograms are reconstructed. Here, we have two simultaneous sources. This is fixed and a priori known. Moreover, using the localisation cue, this assignment depends on reference delays, and then on the location of the sources which is static and known. We can say that the contributions of each source are "segregated" and "grouped" according to a source-specific information, which is the TDOA.

More precisely, the segregation task consists in recovering from two input channels the spectrograms of two sources, $s=1$, placed left, and $s=2$, placed right. To avoid any confusion, the two segregated representations are labelled $s=1$ and $s=2$, whereas the two input channels are labelled "left" and "right". Hence, the equivalence between the number of sources and number of microphones is not a constraint. To achieve this task, the information, which is extracted in this parallel module, is a local estimate of the RL between the two sources. This is performed in each TF region having a fixed duration and bandwidth. We vary these two parameters in order to optimise the segregation. Two criteria are used to evaluate the performance of the segregation model: (1) similarity between output spectrograms and spectrograms of the signals recorded in isolation (2) recognition performance for segregated signals. Practically, after reconstruction of the two spectrograms, a waveform is re-synthesized in order to feed a recognition module, which runs independently. The evaluation of similar models has been done after evaluating the intelligibility gain for subjects by hearing the re-synthesized signals [2].

4. MODEL DESIGN

4.1. TF representation

We perform a FFT-based spectral analysis with a 44 kHz sampling frequency. In order to divide the TF representation in TF regions we vary the size, this is followed by a filterbank decomposition. In the temporal domain, the time-frame wave is multiplied by a hanning window. This has three levels of duration, 25, 50 and 75ms, to study the influence of this factor. In the spectral domain, the filterbank decomposition is based on the product of the Fourier transform module by the transfer function of the $i=1..nc$ channel filters. We also vary the number of channels with 4 levels $nc=(4, 8, 16, 32)$. In the upper branch of the model (Fig. 2, UB), this is just a segmentation allowing the weighting given by the lower branch (LB) module (arrow (1)). These filters are defined in the Bark scale, and we take the definition proposed by [10]:

$$F_{\text{Bark}} = 13 \operatorname{atan}\left(\frac{0.76 F_{\text{Hz}}}{1000}\right) + 3.5 \operatorname{atan}\left(\frac{F_{\text{Hz}}^2}{7500^2}\right)$$

To build this filterbank, we take into account the three following constraints: (1) It covers the frequency domain $[F_{\text{min}}, F_{\text{max}}]=[50, 4100]\text{Hz}=[0.49, 17.40]\text{Bark}$. (2) The filters are bandpass and their center frequencies F_{ci} are equally spaced in the Bark scale. (3) The global transfer function of the filterbank is flat and unity-gain within a large part of the frequency domain.

So, to define each filter F_i , we have chosen to warp a hanning window defined in the Bark scale. This is centered on F_{ci} , and to derive F_{ci} , we compute in Bark the interval separating two filters $F_{\text{int}}=(F_{\text{max}}-F_{\text{min}})/nc$, i.e., the width of the frequency domain, divided by the number of filters. We have $F_{\text{ci}}=F_{\text{min}}+(i-0.5)*F_{\text{int}}$. The $nc-2$ central hanning windows are symmetric and they cover two intervals. The two extreme ones have their half sides reduced to cover 0.5 interval. The case of $nc=8$ filters is developed Fig. 3. The hanning windows defined in the Bark scale are warped in the FFT linear frequency scale.

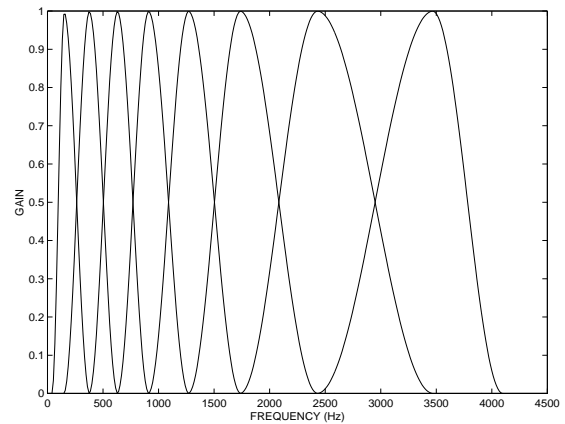


Figure 3: Filterbank having 8 filters. With $F_{\text{int}}=2.11$ Bark, we have $F_{\text{ci}}=[1.55, 3.66, 5.78, 7.89, 10, 12.11, 14.23, 16.34]\text{Bark}$.

In this study, only the "left" input data is taken in the upper branch (UB). For each time frame of the spectrogram, the local upper branch spectrum is:

$$|X_{i,\text{left}}(\omega)| = F_i(\omega) \cdot |X_{\text{left}}(\omega)|$$

4.2. TDOA estimation

In the lower branch (Fig. 2, LB), the TDOA is locally estimated in each TF region allowed by filterbank decomposition, this after cross-correlation. Hence, the cross-correlation function we use is FFT-based. The TDOA estimate is the maximal value picked within an observation window. With a sampling frequency at 44kHz, 40cm distance between microphones, and $c=340\text{m/s}$, the maximal TDOA is 51bin, so the observation window of the cross-correlogram is set at $[-51,51]\text{bin}$. According to the recording set-up shown in Fig. 1, for the left source $s=1$, the difference between right and left pathways is 19.6 cm, and the a priori global TDOA is 25.4bin. Since the set-up is symmetric, this is -25.4bin for the right source $s=2$.

4.3. RL-estimation and sharing function

Then, the local TDOA_i is used together with the global TDOA of the two sources in order to estimate the local relative level RL_i . The knowledge of the global TDOA can be acquired automatically, or it is given. In the first case, the sources are

assumed to be rather stable, and a fullband estimate done in a long time frame (about 200ms) is robust (see [12]). Hence, the number of sources is also known and fixed at two.

A previous study [11] using mixtures of stationary vowels has shown: (1) when a source is dominant in one TF region, the measured delay is near the expected delay of this source, (2) when two sources interfere, the measured delay is shifted towards the delay of the dominant source, according to the relative level between the two sources, this in each frequency channel. So, we proposed to estimate the relative level on the basis of a non-linear function, which is a fit of the statistical relationship between the observable local TDOA and a reference relative level. This is here the ratio between the two local module spectra. This relationship is established having the two isolated sources in hands. The resulting non-linear function is only an overall and qualitative fit of what is observed within all frequency channels. This is retrieved with two static sources which can be various types of signals (non speech, harmonic, non harmonic, narrow or wide-band etc..). Using this fit, in each channel i , from the observed $TDOA_i$, we estimate the relative level in the mixture. Then, this is used as a sharing function. For each source s , we have:

$$W_{s,i}(TDOA_i) = \frac{1}{1 + e^{\beta_s(TDOA_i - \text{bias})}}$$

We adapt the parameters of the two functions, one for each source, to the current set-up (see the result Fig. 4):

- (1) The pair of sources is centred, and $\text{bias}=0$.
- (2) The a priori TDOA of each clean source is 25.4 bin for $s=1$ and -25.4 bin for $s=2$. We set the slope β_1 in order to reach 1-eps, eps is a small value, at $TDOA=25.4\text{bin}$.
- (3) This is used to share the energy carried by each channel i between the two sources, and we verify that:

$$W_{2,i}(TDOA_i) = (1 - W_{1,i}(TDOA_i))$$

$$\Leftrightarrow \beta_2 = -\beta_1, \quad \forall TDOA_i$$

The sharing functions of the two sources are characterised by an inverse slope (Fig. 4), and their sum is one.

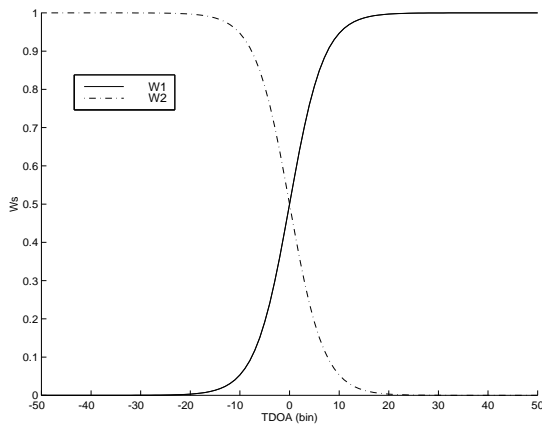


Figure 4: Sharing functions of the two sources $s=1$ and $s=2$ adapted to the current set-up. We have $\beta_1 = 0.2878$, $\beta_2 = -0.2878$ and $\text{bias}=0$.

4.4. Weighting and reconstruction

The sharing function allows us to estimate the spectrum of each source from the spectrum of the mixture data. We preserve the fine details of this spectrum. Again, each source

is reconstructed according an estimation, which is done *via* a parallel representation. The output of this process is applied locally on the upper branch (UB) mixture TF spectrogram in order to estimate the local channel spectra of each source:

$$\left| \hat{X}_{s,i,\text{left}}(\omega) \right| = W_{s,i}(TDOA_i) \left| X_{i,\text{left}}(\omega) \right|$$

Then, the reconstructed spectrum is, for each source, the sum of the local channel spectra. For each time-frame of the spectrogram, we have:

$$\left| \hat{X}_{s,\text{left}}(\omega) \right| = \sum_{i=1}^{nc} \left| \hat{X}_{s,i,\text{left}}(\omega) \right|$$

So, the FFT resolution is preserved and the filterbank analysis only operates subband weighting. The re-synthesis of the signal is easily done by inverse FFT.

5. MODEL EVALUATION

5.1 Measure of the reconstruction accuracy

The recording of isolated sentences allows a reference to estimate the accuracy of the reconstruction from the mixture. We make the direct comparison between references and products of segregation in the spectral domain. Following [13], we define the Recognition Accuracy (RA) measure. This index results of the frame by frame computation of a distance between reference and reconstructed spectra:

$$RA_s = 10 \log \frac{\int_{\Omega} |X_{s,\text{left}}(\omega)|^2}{\int_{\Omega} (|X_{s,\text{left}}(\omega)| - |\hat{X}_{s,\text{left}}(\omega)|)^2}$$

$$\text{where } \Omega/2\pi = [200, 3800]\text{Hz}$$

A complete statistic of RA is established for all time-frames of the 613 pairs of sentences of ST-NB95. This is expressed relatively to the reference RL (Fig. 5, left graphs and Tab. 1), having the two sources recorded in isolation:

$$RL = 10 \log \frac{\sigma_{1,\text{left}}^2}{\sigma_{2,\text{left}}^2}$$

where σ^2 is the mean power and "left" refers to left microphone. Note that for this computation, the time-frame duration is the same as for segregation. In Fig. 5, we show in two conditions that RA closely depends on RL. Moreover, the sum of the two RAs (which is significant of the accuracy for both sources) is not a flat curve (bold curve, Fig. 5 left graphs). In the best condition, when $nc=4$ ((25ms, 4 filters) in Fig. 5), there is a minimum at 0dB. This function becomes bell-shaped when nc increases. The average over all frames shows the best condition is (25ms, 4 filters), with 10.4 dB.

	4	8	16	32
25 ms	10.4	9.9	8.1	5.8
50 ms	10.1	9.5	7.6	5.8
75 ms	9.5	9.0	7.1	5.6

Table 1: Average "summed RA" in dB, over all frames of the database. Row: variation of the number of channels (nc). Column: variation of the time-frame duration.

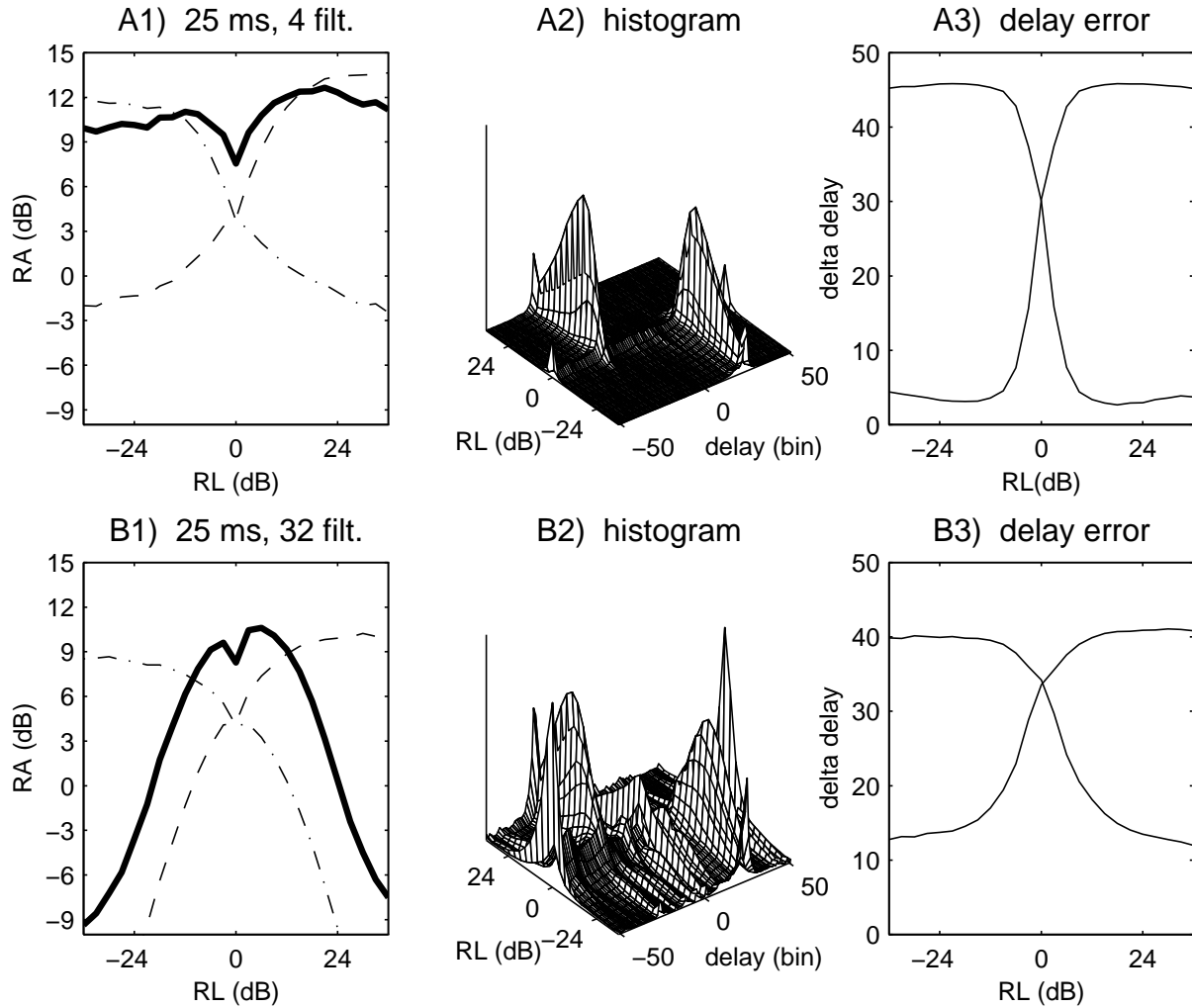


Figure 5: Statistics of recognition accuracy in dB (left, (A1,B1)) and distribution of TDOAi estimate relatively to RL_i (center, (A2,B2), right, (A3,B3)). Two conditions are figured: Ax, up (25ms/4 channels), Bx down (25ms/32 channels). In (A1,B1), the RAs are the dashed curves, whereas the "summed RA" is in bold. In (A2,B2): histograms of TDOAi relatively to RL_i. In (A3,B3): average difference between observed TDOAi and global TDOA for the two sources, relatively to RL_i.

In Fig. 5, center graphs, we show the distribution of the local TDOAi estimate relatively to the local relative level RL_i. In the 4 channels condition, this is well concentrated around the two global TDOA, whereas it is dispersed in the 32 channels condition. So the switch of the TDOA is steeper around 0 dB in the 4 channels condition (Fig 5, center and right graphs): there is a steep dominance effect of one of the two sources. The consequence is a better estimation of the TDOA, which is more suitable for RL estimation thanks to the sigmoid function.

5.2 Recognition experiments

Recognition is implemented with the STRUT software package. The training procedure is carried out using the train part of the original NB95. The whole database is a set of 15,000 sentences produced by 1,132 speakers and transmitted by telephone, only including numbers. This is sampled at 8kHz. A HMM is built for each of the 32 different words, also including probability of transition between the phonetic states, to select the best word candidate within the limited dictionary and to correct it. The sampling frequency of ST-NB95 is about 44 kHz (43,993 Hz), so a decimation factor of 5.5 is applied to resample the re-synthesised signal at 8kHz before recognition.

We used hybrid HMM/ANN, one state HMM with duration modelling. Multilayer Perceptrons (MLP) have been trained on 1,534 utterances of the monophonic Numbers95 digit database, and are used without any adaptation to generate local probabilities for HMMs. Two common types of feature's pre-processing are tested: PLP and JRASTA-PLP. For both of them, the MLP have 10 lpc order analysis, 12 cepstral coefficient and energy. We extracted delta and delta-delta of all previous parameters. This set of parameters is taken for 9 successive frames of 25 ms shifted of 12.5 ms. Then MLP have a total of 351 input units, 1,500 hidden units, and 33 outputs, one for each phoneme class.

A small degradation of the signal is introduced during the record and a re-adaptation of the recognition system is not needed in the context of the present study. We evaluate this by comparison of the fullband, J-RASTA/PLP, recognition WER (Word Error Rate) between "clean" sentences of NB95 (7.1/6.9%; 1,200 sentences) and "isolated" sentences of ST-NB95 (s=1, 9.6/12.5%; s=2, 9.1/11.7%; 613 sentences each). The WER, computed at the sentence level, cumulates three error types: insertion, deletion and substitution of words.

ms/nc	s=1	s=2	Mean
25/4	53.7 / 65.4	54.4 / 65.7	54.0 / 65.5
25/32	62.3 / 65.8	61.2 / 65.7	61.7 / 65.7
75/4	55.8 / 66.3	54.8 / 65.1	55.3 / 65.7
75/32	61.6 / 66.4	60.8 / 64.3	61.2 / 65.4
Mean	58.3 / 66.0	57.8 / 65.2	58.0/65.6

Table 2: WER \pm 1% of the model with comparison between J-RASTA and PLP pre-processing in four conditions (613 sentences). Column: s=1 and s=2 reference sources. Row: the four conditions with varying time-frame duration and varying channel number. a/b: J-RASTA/PLP. Mean WER of J-RASTA is 58 %, whereas PLP WER is at 65.6 %. The best condition is 25 ms/4 channels with a J-RASTA pre-processing (54% WER).

Finally, we establish cocktail-party recognition performances in WER (Tab. 2). We match the two reconstructed sources with their respective word content reference source; i.e., the protocol is adapted for the simultaneous speech recognition task and recognition is tested on the left microphone channel only, to recognise s=(1,2) sources. There are four factors in this experiment: frame duration, channel number, pre-processing type and source. For the recognition experiment, we have limited the number of levels of the two first factors, by taking the extreme values only. The segregation is performed using 2*2 conditions of frame duration (25,75)ms and channel number nc=(4,32). The recognition is applied after two different pre-processing methods J-RASTA or PLP.

First, we verify that the source factor has no effect. On the contrary, the pre-processing factor has a significant effect and the J-RASTA method works better than the PLP. This pre-processing is known to be more robust than the PLP for various noisy conditions, so this is probably not task dependent. The two other factors, duration and channel number, have no effect with the PLP method. On the contrary, the channel number factor has a significant effect using J-RASTA, with no influence of the duration. An improvement is observed when nc is decreased, without improvement when the duration is increased. This is consistent with the RA statistics we have in these four conditions. Finally, the best condition is 25ms/4 channels using J-RASTA with 54% WER. In this condition, the model is able to recognise a little below 50% of the words emitted by the two sources, so we can conclude it recognises the dominant words. After averaging, we show a global improvement 58.0/65.6% against 72.4/73% in comparison with both J-RASTA/PLP WER applied on the mixture (left channel only) without segregation process. Note that our procedure of WER estimation averages the scores established for each reference source separately. So it favors the insertion of words from the competing source into the target source.

6. CONCLUSION

After varying two main parameters of the segregation model which tune the TF window size, we observe that the bandwidth is the main one. We find that four channels is the best tuning within nc=(4, 8, 16, 32) and that is related to the degradation of TDOA estimation (and then of RL estimation) in narrow bandwidth. This is also consistent with the structure of the speech signal. This has four formant trajectories by average, which roughly fall in these four subbands. This is compatible with the principle of multistream recognition [5] we do not apply here, but which is used in [7] after a similar type of processing. We cannot conclude that a 4-channel design is optimal since we have not included the fullband condition in our study. The reason is we

do not perform a local SNR estimation in this condition, and the result can be influenced by the low frequency content only. The time-frame duration has a small effect in the ST-NB95 recording condition. In more natural conditions (echoic, noisy etc..) the duration is expected to have a more significant effect on reconstruction, since the TDOA estimation will be influenced [12]. Finally, the recognition results are slightly worse than those of our CASA labelling method [7], but this comparison must be refined to conclude there is a significant difference.

ACKNOWLEDGEMENTS

This work is a part of EEC projects TMR SPHEAR and LTR RESPITE (Task 2.1).

7. REFERENCES

- [1] Berthommier, F., Glotin, H., A new SNR-feature mapping for multistream speech recognition, Proc. ICPHS'99, San Francisco.
- [2] Blauert J. (1997), "Spatial Hearing: The psychophysics of human sound localisation", MIT Press.
- [3] Bodden, M. (1993) Modelling human source sound localization and the cocktail-party effect, Acta Acoustica, 1:1:43-55.
- [4] Bodden, M., Anderson, T., R. (1994), Improvement of speech recognition in noise, WOS-report, 94-2094.
- [5] Boulard, H., Dupont, S., Hermansky, H., Morgan, N. (1996) Towards subband-based speech recognition, EUSIPCO, 1579-1582.
- [6] Choi, S., Lyu, Y., Berthommier, F., Glotin, H., Cichocki, A. (1999) Blind Separation of Delayed and Superimposed Acoustic Sources: Learning Algorithms and Experimental Study, ICSP'99, Seoul.
- [7] Glotin, H., Berthommier, F., Tessier, E. (1999) A CASA-labelling model using the localisation cue for robust cocktail-party speech recognition, Proc. Eurospeech'99, Budapest.
- [8] Green, P.D., Cooke M.P., Crawford, M.D. (1995) Auditory scene analysis and HMM recognition of speech in noise, ICASSP, 401-404.
- [9] Hermansky, H., Morgan, M. (1994) RASTA processing of speech, IEEE Trans. on Speech and Audio Processing, 2:4:578-589.
- [10] O'Shaughnessy, D. (1987) Speech communication: Human and Machine, Addison-Wesley, New-York.
- [11] Tessier, E., Berthommier, F. (1997) A model of the cumulative effect of pitch and interaural delay differences for double vowel segregation, ICSP'97, pp. 753-758, Seoul.
- [12] Teissier, P., Berthommier, F. (1999), Switching of azimuth and elevation for speaker audio-localisation, ICSP'99, Seoul.
- [13] Yang, X., Wang, K., Shamma, S., A. (1992) Auditory representations of speech signals, IEEE Trans. on Inf. Theory, 38:2:824-839.