A CASA-LABELLING MODEL USING THE LOCALISATION CUE FOR ROBUST COCKTAIL-PARTY SPEECH RECOGNITION

Hervé GLOTIN*+, Frédéric BERTHOMMIER* and Emmanuel TESSIER*

*Institut de la Communication Parlée/INPG
46, Av. Félix Viallet
38031 Grenoble CEDEX, FRANCE
(bertho, tessier)@icp.inpg.fr

+IDIAP
Rue du Simplon, 4
1920 Martigny, Switzerland
glotin@idiap.ch

ABSTRACT

We propose a new cocktail-party recognition technique based on the coupling of a CASA-labelling method using the TDOA (Time Delay Of Arrival) with multistream recognition. This is an alternative to the classical "segregate and recognise" architecture. First, we have recorded a stereo database ST-NB95 from the mono Numbers95. This is composed of binary mixtures of sentences at 0dB, placed left and right. The probability to get the labels "left" and "right" is assigned to the subband time frames thanks to a mapping function. This depends on the relative level. It is established *a priori*, using a reference database composed of isolated words recorded in the same condition. We adapt the recognition paradigm to this particular situation. The model WER of binary mixtures is about 50%. This is a great improvement relatively to the WER (73%) of the fullband PLP. We conclude the model is able to recognise the dominant words of a binary mixture.

1. INTRODUCTION

Humans are able to well recognise speech mixtures produced by two speakers simultaneously, and they also are able to identify speech in loud noise, this in a wide range of noisy conditions (stationary or not). A psycho-acoustical phenomenon, which is related to the robustness of speech identification, is the salience of speech segments against a background of noise. This can be compared to the visual figure-ground segmentation. A more complex situation is an auditory scene where multiple sound sources are present, which can be heard selectively. The cocktail-party effect is a particular case of this situation, when the task is to hear and recognise speech produced by a given speaker. The localisation cue has a central place in this phenomenon, since it can be involved by three ways: **(1)** spatial localisation is a primitive attribute of a sound source required to analyse the spatial organisation of the scene **(2)** audio and visual localisation co-operate to perform speaker localisation and then lip-reading **(3)** it is enough stable in time to provide information about spectro-temporal organisation of mixtures of signals. We will develop this third point.

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 first 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, more precisely, the TDOA, is classically used in order to enhance or segregate speech and there are many applications based on arrays of microphones. 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]. 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. Therefore, one good property of this model is to allow recognition of simultaneous speakers.

We propose an alternative to this well documented "segregate and recognise" scheme. The main motivations are both modelling of human perception and applications: **(1)** improvement of the robustness of speech recognition with the use of secondary features such as localisation and voicing **(2)** showing that phonetic (recognition) processes are involved in the organisation of the "auditory scene", when the task is to hear speech. So the architecture of the recognition system has to be reconsidered. We illustrate this alternative by the use of the localisation cue, in order to perform speech recognition of binary mixtures at the word level. First, we record a database of binary mixtures. Then, we adapt and then combine two recent models we will briefly present: **(1)** the SNR-feature mapping technique [1] (here, it is named "RL-feature mapping") **(2)** the "full combination" multistream recognition [5]. Since these two models are dedicated to robust recognition of one target speech only, the present task is to recognise the dominant words.

2. RECORDING OF ST-NUMBERS95

The stereo database ST-NB95 is built from NB95 (Numbers95) 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.

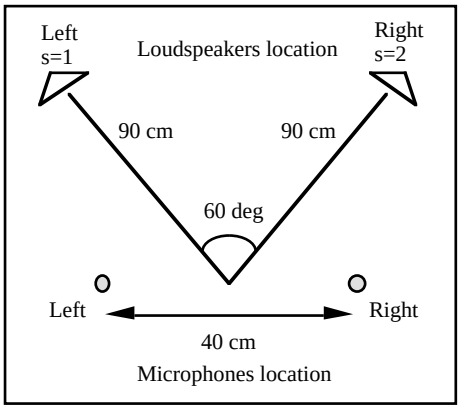

Figure 1: Set-up of the recording of ST-Numbers95

This database is dedicated to establish and compare the performance of robust recognition techniques and cocktail-party segregation methods. The full version of ST-NB95 database is composed of sentences selected from the test set of NB95: **(1)** 2*613 sentences played left or right in isolation **(2)** 613 binary mixtures of paired sentences **(3)** 2*613 isolated sentences played left only added with pink noise. The part of the database we use here 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 (RL) of the same signals in the mixtures, and we build the RL-feature mapping functions.

3. A NEW 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 [6]. 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. 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 (arrow **(1)** in Fig. 2). But these secondary features are not required to perform the acoustic-phonetic decoding.

So, we superimpose to the "main route" a parallel pathway, which can include the previous pre-processing. An extraction process of the secondary feature is performed in parallel, and produces a labelling of the acoustic information. Then, this information is fused at the recognition level (arrow **(2)** in Fig.2). Therefore, the information carried by the main speech-processing route is augmented by labelling, instead of a direct enhancement.

This "CASA labelling" process is compatible with the TF representation feeding the ASR module. Since this is here a multistream model having four subbands [3], the acoustic vectors are localised in time-frequency regions as well as the labelling information. A TF region is simply defined to be a rectangle having frequency bandwidth and duration. In the current implementation, the size of this TF region depends on the size of the target speech features. This is the average phoneme duration and we assume that each frequency subband carries one formant trajectory by average. This lack of frequency resolution is the main reason this model is unable to recognise two words at the same time. The frame duration is 125ms, sliding by steps of 12.5ms. Labelling and recognition are established for the center frame of 25ms.

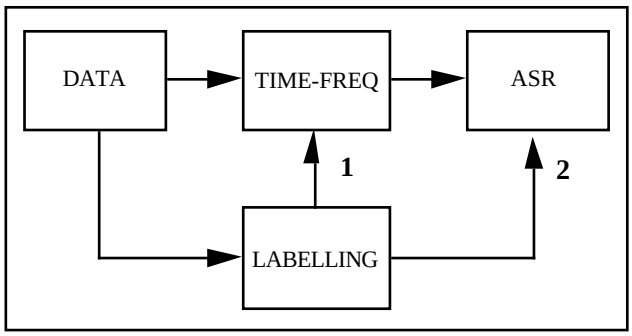

Figure 2: Principle of CASA labelling. The input of the recognition process (ASR: automatic speech recognition) is a time-frequency representation of the speech data. A labelling process compatible with this representation is performed in parallel. These estimates could be used in order to produce an enhancement of the speech directly **(1)** or can be addressed to the recognition step **(2)**, as we propose here.

4. CASA LABELLING

4.1. Extraction of the TDOA

The TDOA is locally estimated in each TF region by cross-correlation after demodulation. The subband waves are half-wave rectified and bandpass filtered in the pitch domain. The bandpass filter is trapezoidal [0, 50, 350, 1000]Hz. The demodulation allows a homogeneous subband estimate. The TDOA estimate is the maximal value taken within an observation window. With a sampling frequency at 44KHz, 40cm distance between microphones, and c=340m/s, the maximal TDOA is 51bin, so the observation window of the cross-correlogram is set at [-51,51]bin.

Then, this local delay is used together with the a priori delay (i.e. the azimuth location) of the different sources in order to label the TF regions. This knowledge can be acquired, or it is given. In the first case, the sources are assumed to be rather stable, and a fullband estimate is robust. A previous study [7] 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 (here, in each subband). So, when there are two static sources, as in ST-NB95, each TF region can get one of the three labels "left", "right", "none". A probability to get these labels is established after choosing a RL threshold and the building of a RL-feature mapping. Remarkably, this probabilistic labelling is compatible with the multistream approach.

4.2. RL-feature mapping

The goal is to acquire a statistical description of the relationship between the local $RL_{s,i}$ of a given source (s=1, "left" and s=2, "right"; i is the subband index) and the observable parameter which is the $TDOA_i$. To achieve the mapping step, the effective $RL_{1,i}$ is established using the two isolated records:

$$RL_{1,i} = 10\log\frac{\sigma^2_{1,i,left} + \sigma^2_{1,i,right}}{\sigma^2_{2,i,left} + \sigma^2_{2,i,right}}$$

where σ^2 is the mean power and "left","right" refer to left, right microphones. Hence, the relationship between the RL of the two sources is simply $RL_{2,i} = - RL_{1,i}$. During the mapping stage, the RL is an observable, which is globally controlled, whereas it is the hidden variable during the test. To build this statistical representation of the relationship between the RL and the TDOA, the frames of the 613 paired sentences have been incorporated. This corresponds to 89305 time frames of 125ms duration sliding by 12.5 ms step. The local $RL_{1,i}$ varies between -36 dB and 36 dB.

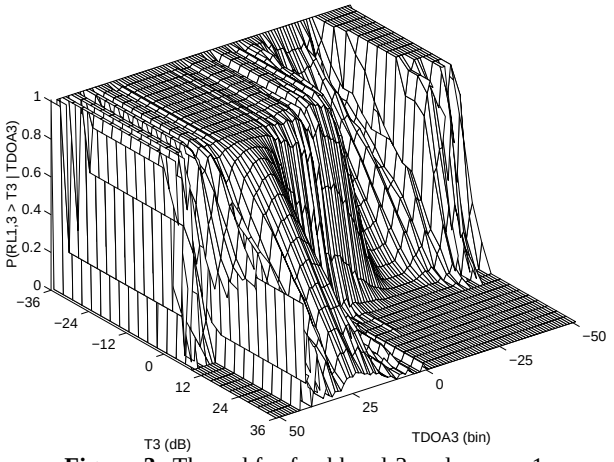

Figure 3: The c.d.f. of subband 3 and source 1.

A 2D counting histogram is built for each subband separately, as in [1]. The two axes of this histogram are: **(1)** the effective $RL_{1,i}$ observed in each TF region, **(2)** the $TDOA_i$ estimate. It counts the number of frames observed for each ($RL_{1,i}$,$TDOA_i$) levels (3dB steps for the RL and 1 bin for the TDOA). After normalisation, a conditional probability distribution $P(RL_{1,i}|TDOA_i)$ is derived from this histogram. A cumulative density function (c.d.f., Fig. 3) is established for each source s=1,2, in which a new "threshold axis" replaces the $RL_{1,i}$ axis of the probability function. The sign of the $RL_{1,i}$ axis is inverted to evaluate the c.d.f. of the second source.

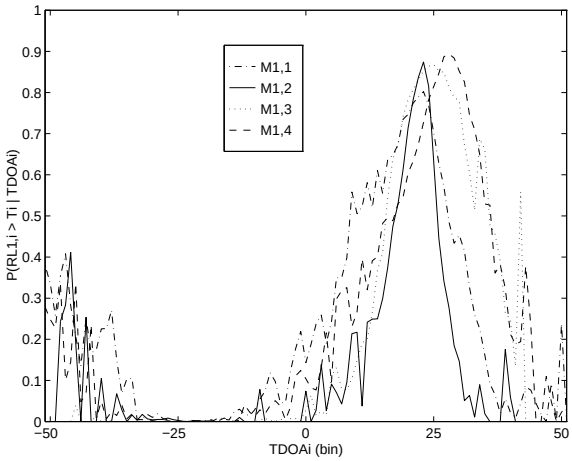

Figure 4: The four $M_{1,i}$ functions of the source 1.

Finally, the 2*4 functions $M_{s,i}$ which are extracted relate the observable $TDOA_i$ and the probability $P_{s,i}$ to carry clean data of the source s. This is the probability to be above a given local $RL_{s,i}$ threshold T_i for each source s:

$$P_{s,i} = M_{s,i}(TDOA_i) = P(RL_{s,i} > T_i | TDOA_i)$$

This function (Fig. 4) is a slice of the previous c.d.f. As in [1], the T_i values are deduced from simulations of subband recognition performance with white noise: this is the point where the degradation of performance reaches 10%. With the current subband recognition model, we found T_i = [12, 9, 9, 9]dB.

5. RECOGNITION EXPERIMENTS

5.1. CASA-ASR coupling

The "full combination" method [5] is a recent development of the multistream theory that uses a criterion for choosing the values of the weights, and then performs an additive fusion of estimates. We adapt this method to the recognition of two simultaneous sources. Let $C_{s,j}$ be the event "j is the stream which carries all the clean data of the source s". Then the probability $P(C_{s,j}|X)$ defines the weight assigned to each stream j=0..15 for the source s. This stream-level probability is derived from the previous subband probability $P_{s,i}$ to carry clean data of the source s:

$$P(C_{s,j}|X) = \prod_{i \in S_j} P_{s,i} \prod_{i' \in \overline{S_j}} (1 - P_{s,i'})$$

in which S_j is the set of subbands included in the stream j. Then, we derive the global multistream posterior for each source. We calculate a weighted average of all streams' posteriors for each source s:

$$P_s(q_k \mid X) = \sum_{j=0}^{15} P(q_k, C_{s,j} \mid X)$$

$$P_s(q_k \mid X) \approx \sum_{j=0}^{15} P(C_{s,j} \mid X)\, P(q_k \mid X_j)$$

Remark that only the term $P(C_{s,j}|X)$ is specific of the sources, whereas the streams' posteriors are not. In the context of robust speech recognition (i.e. speech+noise), this was interpreted by the authors [5] as a posterior weighting favouring the cleanest streams relative to the others. We propose another interpretation related to the ternary labelling of the TF regions as ("left", "right","none") which can be evaluated from the two isolated signals (as in [4]). The labels ("left","right") are assigned when a source is dominant in a TF region, otherwise the label "none" is given. Here, the probability $P_{s,i}$ to get these labels is determined from the mixture *via* the RL-feature mapping. Therefore, the "full combination" weighting takes into account the reliability of the labelling process itself.

5.2. A multistream recogniser

Recognition is implemented with the STRUT software package. We cut the frequency domain into four bands with little overlap [216, 778]Hz, [707, 1631]Hz, [1262, 2709]Hz, [2121, 3769]Hz. Four subband recognisers MLP(i) are trained separately. Their input includes PLP features, energy, 1st and 2nd derivatives. Here, the fusion of the outputs $P(q_k|X_i)$ of each MLP(i) is performed for each time frame in order to evaluate the posteriors $P(q_k|X_j)$ for each of the 15 streams. The posteriors of the void stream j=0 correspond to the priors $P(q_k)$. Since these four subband recognisers work independently, the fusion is effected by a product, which is corrected to take into account the number of subbands included in the stream [5]. This is a reduction of the computationally intensive use of the effective set of $MLP(S_j)$.

5.3. Implementation and testing

The training procedure is carried out using the train part of the original NB95. The whole database is a set of 15000 sentences produced by 1132 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 a limited dictionary and to correct it. The sampling frequency of ST-NB95 is about 44 KHz (43993 Hz), so a decimation factor of 5.5 is applied to resample the signal at 8KHz before recognition. 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 PLP recognition WER (Word Error Rate) between "clean" sentences of NB95 (6.9%; 1200 sentences) and "isolated" sentences of ST-NB95 (s=1, 12.5%; s=2, 11.7%; 613 sentences each).

	Left	Right	Mean
s=1	48.9/74.3	49.8/74.9	49/75
s=2	48.3/71.3	49.9/70.5	49/71
Mean	49/73	50/73	**49/73**

Table 1: WER (%) of the model compared to the fullband PLP (613 sentences). Column: left and right microphone channels. Row: left (s=1) and right (s=2) sources. a/b: Model/Fullband. Mean WER of the model is 49%, whereas the fullband PLP WER is at 73%.

Finally, we establish cocktail-party recognition performances in WER (Tab. 1). The protocol is adapted to the simultaneous speech recognition task. The recognition is applied on each microphone channel, "left" and "right", to recognise "left" (s=1) and "right" (s=2) sources (i.e., sentences). This provides two recognition scores per microphone channel. After averaging, we show a great improvement in comparison with the fullband PLP WER (73%). The model is able to recognise about 50% of the words emitted by the two sources, so we can conclude it well recognises the dominant words. Note that 92% WER is observed for the model when we swap the labels "left" and "right", and that the WER of the "blind" full combination is at 76% (s=1, left; equal weights at 1/16).

6. CONCLUSION AND PERSPECTIVES

The localisation cue is effective to do robust cocktail-party recognition using an architecture, which is not the serial "segregate and recognise" scheme. But our alternative proposal is not complete, and more work has to be done in order to recognise well both members of the mixture. The ways to explore are: **(1)** the narrowing of the subbands in order to resolve the formants of each utterance **(2)** the use of a partial recognition model [4] **(3)** the use of a decomposition model guided by the recognition of the dominant word of each pair [8]. Another problem to solve is the adaptation of this method to variable conditions. For a non-static condition, the mapping-functions (Fig. 4) could be shifted according to a global TDOA estimate acquired independently, this for each source. An improvement allowed by the parallel architecture is to make use of multiple cues. A first solution is to make a product between multiple estimates. But fusion of multiple sources of information could be controlled by an "auditory organisation" module, which is not developed here. The current labelling principle, combined with the multistream technique, allows the control of the fusion of multiple labelling by adding constraints. In such a system, we can set a higher priority for a cue which is more specific, so more informative, such as the voicing [1] relatively to the localisation. The role of the control process is to optimise the fusion: **(1)** to override the less informative cues, **(2)** to achieve the co-operation of the cues (e.g. the localisation cue is useful to label unvoiced-consonant segments).

ACKNOWLEDGEMENTS

This work is a part of EEC projects TMR SPHEAR and LTR RESPITE.

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] Bourlard, H., Dupont, S., Hermansky, H., Morgan, N. (1996) Towards subband-based speech recognition, EUSIPCO, 1579-1582.

[4] Green, P.D., Cooke M.P., Crawford, M.D. (1995) Auditory scene analysis and HMM recognition of speech in noise, ICASSP, 401-404.

[5] Hagen, A., Morris, A., Bourlard, H. (1998) Subband-Based Speech Recognition in Noisy Conditions: The Full Combination Approach, Res. Report IDIAP, 15, Dec. 98.

[6] Hermansky, H., Morgan, M. (1994) RASTA processing of speech, IEEE Trans. on Speech and Audio Processing, 2:4:578-589.

[7] 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.

[8] Varin, L., Berthommier, F. (1997) A probabilistic model of double-vowel segregation, Proc. Eurospeech'97, 5:2791-2794, Rhodes.