INTERFACING OF CASA AND PARTIAL RECOGNITION BASED ON A MULTISTREAM TECHNIQUE

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

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

+IDIAP
P.O. Box 592
CH-1920, Martigny, SWIZERLAND
{glotin, bourlard}@idiap.ch

ABSTRACT

We propose a running demonstration of coupling between an intermediate processing step (named CASA), based on the harmonicity cue, and partial recognition, implemented with a HMM/ANN multistream technique [2]. The model is able to recognise words corrupted with narrow band noise, either stationary or having variable center frequency. The principle is to identify frame by frame the most noisy subband within four subbands by analysing a SNR-dependent representation. A static partial recogniser is fed with the remaining subbands. We establish on Numbers93 the noisy-band identification (NBI) performance as well as the word error rate (WER), and alter the correlation between these two indexes by changing the distribution of the noise.

1. INTRODUCTION

Speech recognition methods are sensitive to noise, because matching between input acoustic vectors and templates, even if intrinsically robust, does not support bias and variance introduced by interfering sources. Because interferents are generally non-stationary and their statistics unknown (i.e., no model of the interferent is available), one strategy is to optimise the use of the available features. This is referred as "speech enhancement". But enhancing the speech against a background, before recognition, requires *a priori* knowledge about the reliability, the specificity, and the redundancy of the features to be enhanced. The problem is: what kind of *a priori* knowledge ?

1.1 CASA methods

The goal of CASA (Computational Auditory Scene Analysis) is to model auditory integration of complex sounds presented in an auditory scene context, and to understand how percepts of these sounds are unified despite their apparent dispersion in the auditory representation, even when several sources interfere; i.e., how components belonging to each source are extracted and grouped. Perceptually, this results in a streaming effect [3]. Speech is a rich complex sound, expected to be processed and streamed by the auditory system in a similar manner to other complex sounds.

Since the task is to communicate, speech is decoded to achieve identification. This involves phonetic features, so the attributes of speech as a complex sound (i.e., the primitive attributes) are not necessarily useful at the recognition level, but these could participate in their extraction and/or in the streaming effect. The goal of coupling between CASA and speech recognition is to improve identification of the speech in the presence of interference. CASA could improve the extraction step to feed the recogniser with enhanced speech and/or could participate in the grouping of components belonging to different sources.

1.2 Structure and Robustness

Robustness is conferred by the structure of energy distribution observed in the time-frequency representation. Since the energy of one speech source embedded in noise is not uniformly distributed within this representation, salient regions of speech appear, those having a positive local SNR (Signal Noise Ratio). Spectrally, formants are robust phonetic features because the local SNR of peaks is likely to be better whatever the background. Temporally, bursts, amplitude and frequency modulations are other structures common to complex sounds. In the temporal domain, enhancement is allowed by combining a temporal derivative, preceded by spectral and temporal integration. This is a first example of a primitive extraction mechanism (common to complex sounds), in which energy is not the only factor of robustness, since it is coupled with another characteristic. This principle is the basis for the success of pre-processing methods like RASTA-PLP [8]. A fine observation of the acoustic structure of the speech provides other cues which could be efficiently combined with energetic salience to produce enhancement. The enhanced representation is named S(E, A) where E is energy and A a supplementary attribute. Here, we have S(E, dE/dt). We will show that harmonicity is another cue, providing S(E, H), but needing an intermediate representation to be processed.

1.3 Redundancy, SNR-dependent selection, and partial recognition

Now, the extracted information is not necessarily in the proper form to feed a normal recogniser. Because the speech signal is redundant, a truncated acoustic representation is sufficient to perform partial recognition, as demonstrated by Green, Cooke et al. [4,6]. Using a Gaussian Classifier, the most simple version is the "marginal" one, which ignores missing values. A second method reconstructs the data in order to evaluate the cepstral coefficients, to improve recognition. Investigations [4, 6] using these tools show that **(1)** deletions can be applied to the input time-frequency representation without great degradation of performance **(2)** for one mixture, a good selection criterion for time-frequency regions is produced by computing local SNR between the original signals, here clean signal and noise (the threshold is fixed around 0 dB). This shows that energetically salient regions carry a significant part of the information needed by the recognition process, and that E is the main factor of robustness. Selecting the regions where the local SNR is high and ignoring the rest is equivalent to a speech enhancement technique producing S(E, ø) without an additive cue. But these are "simulations", and without reference signals, the problem is to specify a SNR-dependent selection process with similar performance. Green, Cooke et al. [4, 6] suggest applying CASA methods to extract the same features; for example to track formants with interference; but this has never been shown. The purpose of this paper is **(1)** to define a model of S(E, H) carrying sufficient phonetic information to achieve robust recognition **(2)** to put forward an operational partial recognition method.

2. CASA LABELLING

2.1 How to build S(E, H) ?

Now, we combine the harmonicity cue with the energetic salience to derive S(E, H). Here, this representation is designed to be compatible with the partial recognition technique, and it is defined as a set of data selected from the time-frequency representation (i.e., "masked" data [6]), and not as a full "enhanced" representation. To understand the principle, remark first that the Fourier spectrum of a stationary voiced vowel is formed by two superimposed structures having different spectral scales: a comb (fine grain) carrying formant peaks (coarse grain). Here, harmonicity is the additive structural information represented by regularly spaced peaks. Harmonic structure can be used as a pointer to formants, but a sieve process is needed to decode it. This is the aim of the so-called "fo-guided" methods which extract a harmonic comb after characterising its fundamental frequency (fo). These remain SNR-dependent because these fine structures are masked when not sufficiently energetic. In auditory-scaled representations and in the high-frequency region, harmonicity appears in the temporal domain as beating, because these higher harmonics are unresolved. In order to work with auditory representations, we propose (as in [1]) to analyse harmonicity after demodulation, which is based on half-wave rectification (HWR) and bandpass filtering in the pitch domain. The main advantage of demodulation is to extract the envelope of beating harmonics to characterise it with other signal processing techniques than the sieve. In the modulation spectrum (i.e., a FFT applied after demodulation, Fig. 1a), a harmonic signal is represented by a peak at fo (with some harmonics, not figured), far easier to detect than the initial series of peaks.

The modulation spectrum allows an enhanced full spectral representation [1], but envelope analysis can be achieved in order to build a mask. The method proposed by Gaillard et al. [5] includes a variance estimate, the variance of zero-crossing intervals, to define a local decision model (Fig. 1b). This approach strongly differs from classical fo-guided methods based on the mean (eq. to 1/fo estimate). This allows a decision criterion to label time-frequency regions, in order to get a mask. Label 1 is assigned to regions where a harmonic comb is energetically dominant, according to a decision threshold. Otherwise, regions get the label 0. Finally, S(E, H) is the set of regions receiving the label 1. We show in Fig. 1 the principle of decision models based on other intermediate representations after demodulation: the modulation spectrum and the autocorrelation (Acorr).

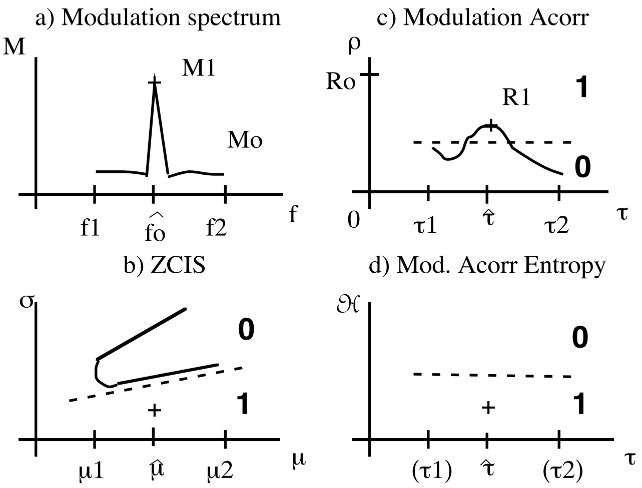

Figure 1 : Four SNR-dependent methods used to detect locally in the time-frequency plane a dominant harmonic comb (plotted +) competing with a background. Fundamental frequency estimate is an abscissa value. All methods are based on pre-processing by demodulation (HWR and Bandpass filtering), allowing dependence on harmonicity. Boundaries of the pitch domain are plotted. **a)** Modulation spectrum: After FFT (eq. to AMmap [1]). M1 is the peak module, and Mo is the sum within the pitch domain, so that M1/Mo estimates the modulation rate **b)** Zero Crossing Interval Statistics: Based on mean μ and standard deviation σ [5]. The threshold is defined to optimise harmonic comb detection, to label 1: "harmonic-dominant region" or 0: "other", noise-dominant or inharmonic, or interferent (hyperbolic boundary) **c)** Modulation Acorr: After autocorrelation (Acorr) R1 is the peak value taken within the pitch domain and Ro the 0 lag value. The modulation index R1/Ro of the noise sets a reference threshold **d)** Modulation Acorr entropy: the entropy ($\mathcal{H}$) of the same full autocorrelogram is a global index also depending on periodicity.

The characteristics and performances of these techniques differ (comparative results are not available yet). The modulation index R1/Ro, used in this paper, and the entropy of the full correlogram (tested in [7]) allow large time-frequency regions to be labelled.

2.2 Noisy Band Identification (NBI)

Now, when the mixture is speech with added localised noise, the representation S(E, H) adapted to partial recognition directly emerges from the selection of clean speech components, i.e. from the noise/speech segmentation. Since speech is composed of a majority of segments which are more or less voiced, there are two main segmentation strategies: **(1)** determination of boundaries of the noise region **(2)** differentiation between noise and harmonic signal owing to the fine spectro-temporal structure. The second strategy (which we use) is closer to an image texture analysis. The role of fine-grain structures is clear when compared with a low level detection strategy based on spectral peak-picking. This directly depends on SNR, and it is pertinent only if the task is to detect a sine-

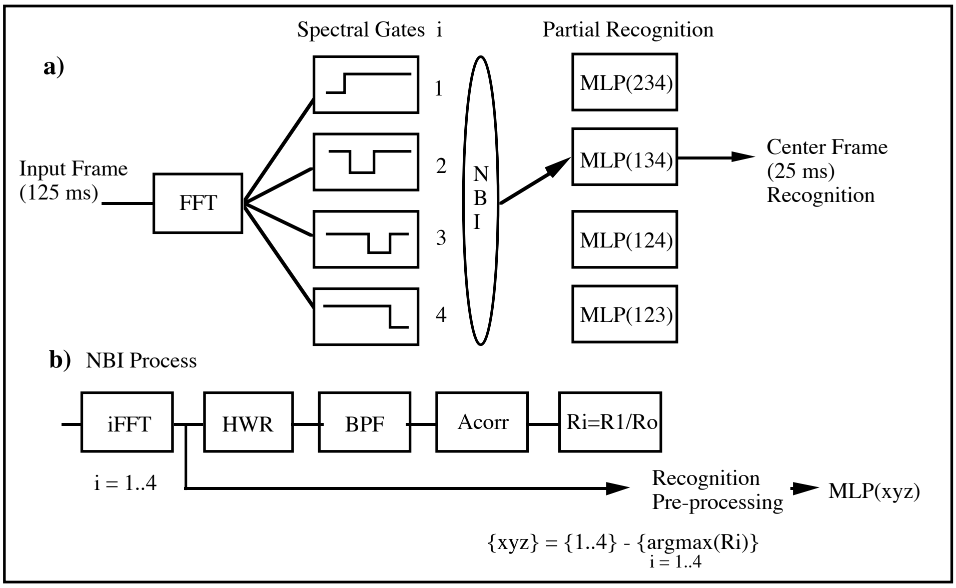

Figure 2 : Block-Diagram of the model. **a)** Four groups of subbands are built after FFT and spectral gating, indexed i=1..4. The NBI process (expanded in b)) selects one partial recogniser MLP(xyz) which performs recognition assigned to the center frame. **b)** For each group, after spectral gating, the demodulation process consists of Half Wave Rectification of the group-wave recovered by iFFT followed by Band-Pass Filtering in the pitch domain. The choice criterion is the modulation index R1/Ro: the more modulated group-wave is likely not to include the noisy band. The corresponding group-wave is addressed to the MLP(xyz).

wave, not a complex of peaks. The autocorrelogram of the demodulated signal is able to serve as a basis for differentiating between harmonic signal and noise with a time window shorter than the phoneme duration, but needs a large frequency bandwidth. A correlogram of a time-frequency region including a noisy band is less modulated. If the frequency domain is divided into four subbands, the following methods can detect (one or more) noisy subbands within the four: **(1)** use of a decision threshold to decide locally, subband by subband, if noisy or not **(2)** use of a threshold to decide group-wave (group of three subbands, see Fig. 2) by group-wave, if noisy or not **(3)** compare the subband-waves **(4)** compare the group-waves. When there is only one noisy subband, i.e. the task we have chosen, the choice of one subband is forced and **(4)** is the best one: this is robust because of the large frequency bandwidth and moreover, this does not require the tuning of a threshold. Consequently, we do not label the subbands independently, and the result is a noisy band identification (NBI). To have S(E, H), this subband, indirectly labelled 0, is removed from the input time-frequency representation during a time frame duration. Finally, we show that this algorithm is able to "pop-out" a band-limited noise corrupting harmonic segments of the speech, before recognition and frame by frame.

3. MODEL DESIGN

3.1 A static partial recogniser

We cut the frequency domain into four bands having limited overlap [0, 901]Hz, [797, 1661]Hz, [1493, 2547]Hz, [2298, 4000]Hz. Four partial recognisers MLP(xyz) are trained with the four combinations of three of these domains. After LPC pre-processing, input acoustic vectors are formed by merging energy, 1st and 2nd derivatives, and cepstral coefficients. Consequently, these four recognisers are not independent, but this is an advantage because covariance between subband data is taken into account. This differs significantly from the use of independent subband streams and we not intend to fuse their outputs. Secondly, multistream is based on a hybrid HMM/ANN recognition model [2] which is more robust than a Gaussian classifier. Consequently, good performance is obtained without reconstruction, even when large blocks are deleted from the acoustic representation. Frame by frame, the "best" MLP(xyz) is selected according to the evaluation of S(E, H). This is close to the simplest version, the marginal one, of partial recognition methods based on a Gaussian classifier. The main difference is we perform a static partition. Finally, interfacing between NBI and recognition is shown Fig. 2. It is dedicated to recognition of speech added with a narrow band of noise. The computational load is low, and our demonstrator is quasi-real time.

3.2 Implementation and testing

Recognition is implemented with the STRUT software package, allowing choice of different pre-processing as well as full-band and multistream recognition techniques. During the recognition stage, a MLP (full-band or MLP(xyz)) produces, frame by frame, a vector of 58 values. These are good estimates of posterior probabilities; i.e., probabilities of the current acoustic vector to be a member of each of the 58 phonetic classes. Training and test procedures are carried out using Numbers93. This is a set of 2167 sentences transmitted by telephone, only including numbers produced by 1132 speakers. A HMM is built for each word, also including probability of transitions between the phonetic states, to select the best word candidate within a limited dictionary and to correct it. Performance is expressed in WER (Word Error Rate).

Coupling of the two steps, CASA and Multistream recognition, is achieved with a forward model having compatible frames (Fig. 2). The frame duration is 125ms, sliding by steps of 12.5ms. NBI and recognition are established for the center frame of 25ms. Input signals are sampled at 8KHz. The same group-wave (spectrally gated signal) feeds both processes.

4. PERFORMANCE

The rectangular band of noise, 9dB global SNR and 400Hz bandwidth, is centred in each of the subbands previously defined. We establish statistics for NBI (Tab. 1) and WER (Tab. 2) by varying the noisy subband, on the same test database.

Nsb	**1**	**2**	**3**	**4**
a/b/c	15/65/78	36/81/89	28/77/88	21/82/87

Table 1 : Statistics of NBI-correct over all frames of the test database (from Numbers93) with stationary noise (9 dB, 400 Hz bandwidth). The noisy subband (Nsb) varies from 1-4. NBI method is based on modulation index R1/R0 (pitch within [90, 250]Hz). a/b/c rates (%) are respectively: a-rate of selection of this subband with clean signal, silence excluded (threshold at 40dB); b-NBI-correct all frames confused; c-NBI-correct silence excluded.

Nsb	**1**	**2**	**3**	**4**
FB a/b	38/47	40/41	45/40	73/24
c/d/e	19/20/27	15/16/18	12/13/17	12/14/18

Table 2 : WER statistics over all words of the test database. FB: full-band MLP in noise with 2 different pre-processing methods. a/b WER are respectively a-LPC (clean: 12%); b-LogRASTA-PLP (clean: 11%). c/d/e WER are respectively: c-MLP(xyz) with clean signal; d-Nsb given; e-model.

Table 2 shows a strong improvement relative to the full-band methods, even robust methods such as LogRASTA-PLP. NBI and WER are negatively well-correlated (cor=-0.98) for stationary noise. To decorrelate them, i.e. to get different WER with the same NBI (Tab. 3), the effect of noise distribution is analysed with two conditions having the same number of 125ms noisy frames in each subband: **(1)** random uniform; **(2)** regular, with a circular variation of the noisy band.

	Random	**Regular**
NBI-correct	62	62
Silence excl.	66	67

Table 3 : NBI statistics over all frames of the test database with non-stationarity of the noisy subband (9dB, 400 Hz bandwidth), random or regular. Rates (%) are respectively: NBI-correct all frames confused; NBI-correct silence excluded.

	Random	**Regular**
FB a/b	49/41	49/43
Nsb-given	30	26
Model	34	29

Table 4 : WER statistics over all words of the test database with Nsb variation. FB: full-band MLP. a/b WER are respectively: a-LPC; b-LogRASTA-PLP.

First, we have worse NBI and WER rates in the non stationary condition. Secondly, the better rates observed in the regular condition (Tab. 4) can be attributed to the higher degree of stationarity, but also to the time redundancy of the speech signal and to the larger time scale of word recognition, whereas the NBI process is memoryless.

5. MODEL IMPROVEMENT

The current model is adapted to a very limited range of interfering conditions (hence, it uses strong *a priori* knowledge of the interference characteristics). However, the underlying principles are promising and can be extended by: **(1)** counting the number of noisy subbands, by applying the decision model subband by subband; **(2)** use of an extended set of partial recognisers; **(3)** optimisation of the control. For example, if counting is zero, the frame is addressed to the full-band MLP, if > 1, integrate over, considering that 2 subbands are not sufficient to get reliable identification; **(4)** use of different frame duration's for CASA and recognition, because NBI frame duration can be shorten without great degradation; **(5)** building a probabilistic model of labelling (local estimate of P(label 0) and P(label 1)); **(6)** tracking of the noise, with some *a priori* knowledge; **(7)** use of some supplementary processing between the two steps; **(8)** use of other attributes: with a binaural input and when interfering signals are spatialised, Interaural Delay Difference is a cue to label the time-frequency representation and to get a S(E, ITD) closely compatible with the current design. Onset/Offset and amplitude modulation cues are other potential attributes to exploit.

Acknowledgements : This work is granted by the project COST249 and it is a part of EEC projects TMR SPHEAR and LTR RESPITE. Thanks to Ramsay Gordon for reading and comments on this paper.

6. REFERENCES

[1] Berthommier, F. & Meyer, G. (1995) Source Separation by a Functional Model of Amplitude Demodulation, Proc. Eurospeech'95, Madrid, pp. 135-138.
[2] Bourlard, H., Dupont, S. , Hermansky, H. & Morgan, N. (1996) Towards sub-band-based speech recognition, European Signal Proc. Conf., Trieste, pp. 1579-1582.
[3] Bregman, A.S. (1990) ASA, MIT Press.
[4] Cooke, M., Morris, A. & Green, P. (1996) Recognising occluded speech, Proc. of Workshop on the Auditory basis of speech perception, Keele, pp. 297-300.
[5] Gaillard, F., Berthommier, F., Feng, G. & Schwartz, J.L. (1997) A modified zero-crossing method for pitch detection in interfering conditions, Eurospeech'97, Rhodes, pp. 445-448.
[6] Green, P.D., Cooke M.P. & Crawford, M.D. (1995) Auditory scene analysis and HMM recognition of speech in noise, Proc. ICASSP, pp. 401-404.
[7] Glotin, H., Tessier, E., Bourlard, H. & Berthommier, F. (1998) Reconnaissance multi-bandes de la parole bruitée par couplage entre les niveaux primitifs et d'identification, XXIIèmes JEP, Martigny, pp. 375-378.
[8] Hermansky, H. & Morgan, M. (1994) RASTA processing of speech, IEEE Trans. on Speech and Audio Processing, 2:4:578-589.