A NEW SNR-FEATURE MAPPING FOR ROBUST MULTISTREAM SPEECH RECOGNITION

Frédéric BERTHOMMIER* and Hervé GLOTIN*+
*Institut de la Communication Parlée/INPG, Grenoble, France
+IDIAP, Martigny, Switzerland
bertho@icp.inpg.fr, glotin@idiap.ch

ABSTRACT

We describe a new model of CASA labelling which assigns to each time-frequency region a probability "clean" enough to feed a multistream recogniser only adapted to clean data. This labelling process is based on the harmonicity of the speech. The probability is evaluated according to a SNR-feature mapping and the choice of a SNR decision threshold. This allows an extension of a previous method [1] based on the binary detection of noisy time-frequency regions, followed by partial recognition of clean regions. The labelling process is adapted to a new multistream recognition approach [5], since the previous probabilities serve to weight the streams' posteriors.

1. INTRODUCTION

We propose to label the time-frequency representation after an analysis of primitive features of the speech, such as harmonicity or binaural cues. First, the CASA (Computational Auditory Scene Analysis, see [9]) approach is based on the definition of multiple representations of the signal allowing the extraction of this fine information. This is the *representational* aspect of CASA. But one main hypothesis grounding the development of CASA models is based on Gestalt principles (this is a particular interpretation of these principles): both segregation and grouping of characteristics belonging to different sources are supported by an auditory processing step preceding further stages such as phonetic identification. The natural robustness of human perception is entirely attributed to a low/intermediate level of processing able to perform the auditory scene analysis by itself. However, we assume that : **(1)** A crude application of Gestalt ruled-based principles has never shown results with real world signals; **(2)** the integrative role of Gestalt for the human perception of the speech is doubtful [8]. In order to find a compromise useful for performing robust speech recognition on the basis of the first (representational) CASA principle we want to preserve, we propose to weaken this function attributed to the auditory step:
The auditory level is able to extract information which is not well integrated in the input time-frequency representation, thanks to its representational properties, but we assume it is not able to produce the grouping of phonetic features by itself.

Therefore, our goal is not to model the whole auditory organisation process and the segregation of concurrent sources (speech+noise or multiple speech) without appealing to the phonetic levels, as most of the CASA models do. We limit the task to the robust identification of speech, at the word level in our simulations, embedded in non stationary noise.

Here, the role we assign to the auditory analysis step is to differentiate components of a target speech and interfering components as noise. TF is the reference representation and a labelling process is performed in parallel (Fig.1). This is based on a second representation of the same signal. The TF "acoustic information" is addressed to the recognition module together with the labelling information (Fig. 1, arrow 2). This is an improvement of a classical solution (Fig. 1, arrow 1) that consists in enhancing the features specific to the target source (i.e., to perform speech enhancement) or segregating interfering sources before recognition (i.e., as most of the CASA models of source segregation do). In Figure 1, the ASR module includes a statistical recognition model which works at the phonemic level and produces posteriors $P(q_k|X)$, q_k the set of phonemes, for each time-frame. The time-frames have a duration corresponding to the average phoneme duration, and they are parameterised in order to feed the recognition module with acoustic vectors.

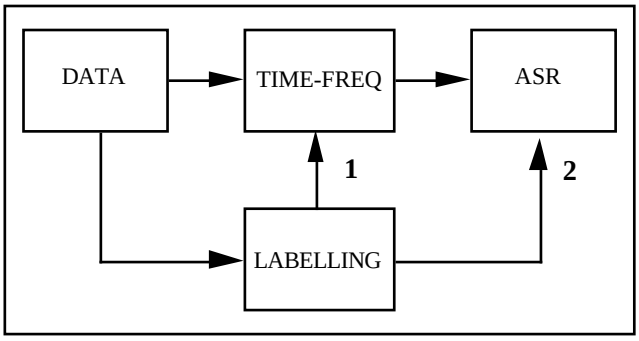

Figure 1 : Principle of CASA labelling. The input of the recognition process (ASR) is a time-frequency representation of the speech data. A labelling process compatible with this representation is performed in parallel. These estimates could be used in order to produce an enhancement of the speech directly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

First, the acoustic vectors feeding such a statistical recognition model are produced by an appropriate pre-processing of the time-frequency (TF) representation. For example, the RASTA-PLP pre-processing technique [6] integrates TF in time, as well as in frequency, in order to represent the formant trajectories well. Let us remark that, in current ASR models, this well-integrated TF representation is

the only source of information used to perform acoustic-phonetic decoding. The obvious drawback of this integration is the loss of fine spectro-temporal information which is a potential source of robustness of speech signals. Here, the CASA labelling step extracts information which is not currently used by the speech recognition model.

Secondly, the ASR module is adapted to integrate this labelling information. Three existing techniques, including a statistical recognition model, are compatible with this principle: **(1)** partial recognition [4], **(2)** multistream recognition [3] and **(3)** model decomposition. We will only discuss **(1)** and **(2)**, since they are adapted to be fed by the labelling information produced from our SNR-feature mapping. We will show that they are closely related to the nature of this labelling information.

To progress towards a realistic CASA modelling framework applied to robust speech recognition, we propose to follow the same route taken forty years ago for phonetic identification, when the stochastic characteristics of phonetic features were taken into account. Hence, the speech signal we hear and see in the real world is not deterministic, and the low level features are also variable. Therefore, the probabilistic CASA modelling style we defend will be compatible with phonetic recognition models, based on the same principles. So, we will develop this new approach in a common CASA/ASR framework in which the signal detection theory and the concept of data reliability have an important place. Low level characteristics which are not engaged in the speech recognition process can be analysed in order to extract information about the SNR, and then to label speech data to be recognised.

2. HOW TO ESTIMATE THE SNR ?

The goal of the labelling is to inform the recognition module about the data reliability *for* the recognition task. Since obtaining a fine SNR estimation is a difficult problem, a simplified task consists in localising noisy and clean time-frequency regions. The feasibility of such a detector was assumed by the missing data theory [4], which shows how to recognise the signals from partial information after detection. The ideal front-end of partial recognition evaluates Boolean masks: TF regions are labelled 0 for noise or 1 for clean data. The authors [4] performed a local measure of the SNR, having noise and signal available separately. Then, they labelled these data according to a threshold. Failures at the detection level are not taken into account. We propose to label the TF regions defined as subband time-frames automatically, and to take into account the characteristics of this labelling level.

2.1 A measure of harmonicity

This measure is based on a new representation of the same signal, developed in parallel with TF. After a bark-scaled filterbank analysis, a harmonic signal appears as a series of spectral peaks in the low frequency domain (resolved harmonics). The harmonics produce beats when they are not resolved, in the high frequency domain. The condition to get a beating envelope after band-pass filtering is to have at least two harmonics passing through the filter. If we cut the whole frequency domain in four equivalent subbands, this condition is easily fulfilled, and the evaluation of the degree of harmonicity of one signal can be based on the same temporal analysis. This allows a homogeneous processing within TF.

The "degree of harmonicity" is analysed after demodulation. This is based on half-wave rectification and trapezoidal bandpass filtering [0, 90, 350, 1000]Hz in the pitch domain. Then, we perform a cross-correlation of the envelope also bounded by the pitch domain. A similar algorithm is classically used in CASA models or in engineering applications to evaluate the pitch within several frequency channels. Following the Gestalt principle of coherence by similarity, the goal of most of the CASA models is then to group spectral components according to the coherence of their fundamental frequency. But we interpret the cross-correlogram differently: we do not extract the fundamental frequency, but a reliability measure. So, the parameter which is evaluated from the autocorrelogram is not the abscissa value of the first maximum taken within the pitch range [90, 350]Hz, but its normalised *amplitude* value. This index, the so-called R1/R0, depends on the harmonicity of the signal and decreases when the signal is corrupted by noise.

2.2 A SNR threshold

Therefore, a decision process about the reliability of the signal can be based on this observable R=R1/R0. A signal will be considered as reliable if it is "sufficiently" clean to be well pre-processed and recognised. In other words, the signal is reliable if the SNR exceeds a threshold. The reason is, above this threshold, the speech features are not masked by the presence of noise, and a pattern matching technique can be applied. The recognition module is trained with clean speech and its output is reliable only if the SNR is high. If the signal is a priori harmonic, the set-up of a R threshold value seems to be satisfactory for differentiating the clean signal from the corrupted signal, or the (non harmonic) noise. In fact, this simple method leads to bad results when it is applied locally in TF regions (hence, we choose a threshold value per subband in this case). This motivates the development of a more sophisticated technique in order to get information about the hidden SNR parameter from the observable R: the SNR-feature mapping (here, the "feature" is R). Such a mapping of the relation between observable speech features and the SNR has been proposed by [10] in order to recover an estimate of the SNR from noisy data. In our application, the mapping allows to substitute a SNR threshold for the previous R1/R0 decision threshold, and then to evaluate the probability of being above this SNR threshold from the observable R. We will see that the relationship between these parameters is non-linear. The SNR threshold Ti is fixed, for each subband i=1..4, according the degradation of the *recognition* performance.

2.3 Method of SNR-feature mapping

The goal is to acquire a statistical description of the relationship between the SNR and the observable parameter R. During the mapping step, the SNR is the controlled parameter, whereas it is the hidden variable during use. To build this statistical representation, white noise is added to a part of Numbers95 (100 sentences of the training set), silence excluded. This is repeated by varying the global SNR from -21 to 39dB. The same number of frames is included in this

statistic for each local level of SNRi. A 2D counting histogram is built for each subband separately. The two axes are: **(1)** the effective SNRi in each TF region **(2)** the Ri estimate. It counts the number of frames observed for each (SNRi,Ri) levels. A conditional probability distribution P(SNRi|Ri) is derived from this histogram, and then a cumulative density function, in which a new "threshold axis" replaces the SNRi axis (Fig.2). Finally, the function Mi which is extracted non-linearly relates the observable Ri and the probability Pi to be above a given local SNRi threshold Ti:

$$P_i=M_i(R_i)=P(SNR_i>T_i \mid R_i)$$

This function (Fig. 3) is a slice of the previous c.d.f.. The Ti values are deduced from simulations of subband recognition performance: this is the point where the degradation of performance reaches 10%. With the current subband recognition model, we found Ti = [12, 9, 9, 9]dB.

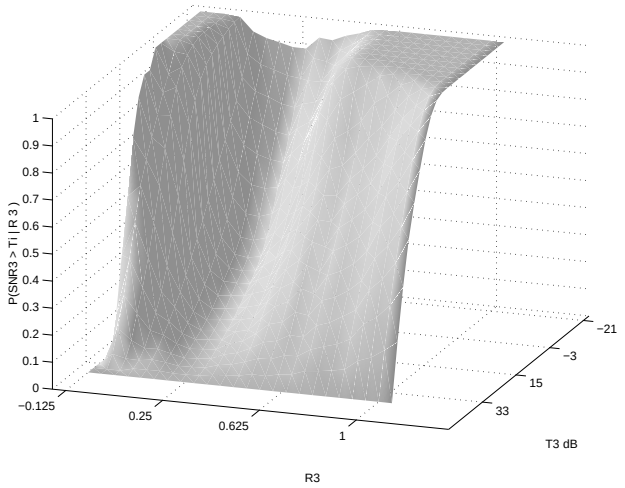

Figure 2 : The c.d.f. of subband 3.

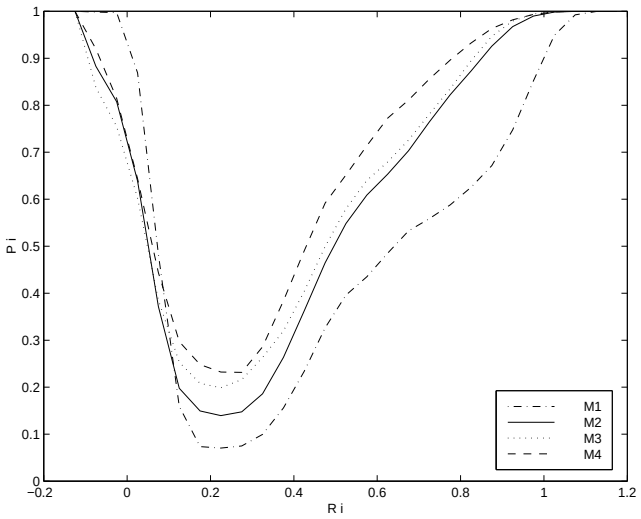

Figure 3 : The four Mi(Ri) functions.

3. A LINK BETWEEN PARTIAL AND MULTISTREAM RECOGNITION

The current work is an extension of a previous demonstration [1] which was based on the missing data approach. Strictly speaking, we performed a marginal partial recognition using the multistream technique, but this model was not realistic for applications: this was limited to the corruption of one subband, stationary or not. Each subband roughly carries a formant trajectory, so, when a subband is deleted, recognition remains feasible thanks to the spectral redundancy due to multiple trajectories. But we also remarked that this is limited by the noise detection performances. In this sense, one of our "improvement proposals" was to build a probabilistic model of labelling; i.e., a local estimate of P(label 0) and P(label 1). These estimates Pi are now available thanks to the SNR-feature mapping. These carry more information than a single Boolean mask, but the problem is to make use of this continuous information. The solution we adopt is based on the multistream technique.

The aim of the multistream technique is to perform multiple stream recognition in parallel, since this is not too combinatorial, and then to fuse the estimates. As proposed by [7], this allows a better estimate of the posteriors thanks to **(1)** the control of the merging process, **(2)** a possible integration of weights or of a heuristic choice of the better streams. 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. Let Cj be the event "j is the stream which carries all the clean speech data". Then the probability P(Cj|X) defines the weight assigned to each stream j=0..15. This stream level probability is derived from the previous subband probability Pi to carry clean data:

$$P(C_j \mid X)=\prod_{i\in S_j} P_i \prod_{i'\in \overline{S_j}} (1-P_{i'})$$

in which Sj is the set of subbands included in the stream j, with label($i\in$Sj)=1. The Boolean mask B(j) associated to a stream j is the binary value of the integer j. Finally, it calculates a weighted average of all streams' posteriors:

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

This was interpreted by the authors [5] as a posterior weighting favouring the cleanest streams relative to the others, i.e. the weighting remains a function of the data reliability only. Two remarks allow us to well relate this multistream method to the partial recognition technique:

(1) The approximation $P(q_k \mid X_j,C_j)\approx P(q_k \mid X,C_j)$ means that noisy subbands do not carry phonetic information. These two terms are equal only for one stream, when Cj is true. Otherwise, this is an approximation. There are two cases: **(1)** Cj is false, but the stream j does not include a noisy subband; this is the aim of marginal partial recognition in using clean data only [4]; **(2)** the noise degrades the evaluation of the posteriors; it is supposed to flatten the distribution of posteriors $P(q_k|X_j)$. Hence, the weights P(Cj|X) lower the influence of these terms.

(II) If we consider the local noise detector to be imperfect, and only able to deliver a probabilistic labelling, it turns out that the same weights represent the reliability of a Boolean labelling process. Therefore, an optimisation of the partial recognition method taking into account the reliability of masks B(j) will consist in averaging each possible marginal recognition of the full acoustic vector (here, the 16 combinations) by the probability of occurrence of the corresponding mask B(j), also given by P(Cj|X). Thus, when the partial recognition is performed with a multistream method, this exactly corresponds to the previous "full combination" version of multistream recognition.

4. MODEL DESIGN AND TESTING

4.1 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 with these subbands. Their input includes RASTA-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(Sj). So, the recognition estimates are initially evaluated locally in TF, as well as the reliability estimates, and then grouped to derive the stream estimates. Their TF windows are fully synchronous and compatible. We remark that the "covariance" information is lost in both levels.

4.2 Implementation and testing

During the recognition stage, each of the MLP(i) produces, frame by frame, a vector of 33 values. These are good estimates of posterior probabilities; i.e., probabilities of the current acoustic vector X_i being a member of each of the 33 phonetic classes. Training and test procedures are carried out using Numbers95. This 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 word, also including probability of transition 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). The coupling of the two steps, CASA labelling and Multistream recognition, is achieved with a forward architecture as shown Fig.1. The frame duration is 125ms, sliding by steps of 12.5ms. Labelling and recognition are established for the center frame of 25ms.

A rectangular band of noise is centred in each of the subbands previously defined. We establish statistics of the WER (Tab. 1) which show a strong improvement relatively to the fullband RASTA-PLP. The model WER is near the result for partial recognition of the 3 clean subbands (given condition), and this is better than the blind condition. The gain cannot then be attributed to the averaging itself.

Nsb	Clean	1	2	3	4
Full	11	73	81	75	71
Given	-	23	26	30	17
Blind	15	30	38	35	23
Model	16	25	33	32	18

Table 1 : % WER statistics over the test database (100 sentences from Numbers95) with stationary noise (9 dB, 300 Hz bandwidth, centred on non overlapped regions); Nsb: Noisy subband; Full: fullband RASTA-PLP; Given: partial recognition on 3 subbands with Nsb excluded; Blind: multistream model with constant weight 1/16; Model: set-up as described in the text.

5. CONCLUSION

These results show that the CASA labelling step differentiates clean and noisy regions of TF well on the basis of the harmonicity of the speech signal. Moreover, the coupling with a multistream model is a promising way to tackle wideband/non stationary noises, since **(1)** the full acoustic vector is taken into account at the recognition level, and **(2)** the labelling process is local. The principle of SNR-feature mapping, here based on harmonicity analysis, can be generalised to other primitive features. Similarly, the function relating the reliability of the signal and the Interaural-time-difference (ITD) has been built [2]. This allows two speakers (and then mixtures of harmonic sounds) to be differentiated according to their different azimuth location. These independent probability measures will be easily merged since SNR thresholds Ti are shared and the TF regions will be compatible. We expect a cumulative improvement of speech recognition from these independent CASA labelling processes.

ACKNOWLEDGEMENTS

This work is supported by the project COST249 and it is a part of EEC projects TMR SPHEAR (Task 3.3) and LTR RESPITE (Task 2.1).

REFERENCES

[1] Berthommier, F., Glotin, H., Tessier, E., Bourlard, H. (1998) Interfacing of CASA and partial recognition based on a multistream technique, Proc. ICSLP'98, Sydney.

[2] Berthommier, F., Tessier, E., Glotin, H. (1999) A CASA labelling method using the localisation cue for Cocktail-Party speech recognition, Proc. Eurospeech'99, Budapest (accepted).

[3] Bourlard, H., Dupont, S. , Hermansky, H. & Morgan, N. (1996) Towards sub-band-based speech recognition, European Signal Proc. Conf., Trieste, pp. 1579-1582.

[4] 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.

[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] Hermansky, H., Tibrewala, S. & Pavel, M. (1996) Towards ASR on partially corrupted speech, Proc. ICSLP, pp. 462-465.

[8] Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S. & Lang, J.M. (1994) On the perceptual organisation of speech, Psych. Rev., 101:1:129-156.

[9] Rosenthal, D.F., Okuno, H. G. (Eds.) (1998) Computational Auditory Scene Analysis, LEA Publisher, London.

[10] Teissier, P. , Schwartz, J.L. & Guérin-Dugué, A. (1997) Non-linear representations, sensor reliability estimation and context dependent fusion in audio-visual recognition of speech in noise, Proc. Eurospeech'97, pp. 1611-1614.