IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Test of several external posterior weighting functions for multiband Full Combination ASR

Hervé Glotin [a] [b] Frédéric Berthommier [b]

IDIAP–RR 00-27

August 2000

to appear in
ICSLP'00

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP Inst. of Perceptual Artificial Intelligence - Simplon 4 - 1920 Martigny - Switzerland
[b] ICP Inst.de la Communication Parlée - 46 Av. Viallet - 38000 Grenoble - France

IDIAP Research Report 00-27

Test of several external posterior weighting functions for multiband Full Combination ASR

Hervé Glotin Frédéric Berthommier

August 2000

to appear in
ICSLP'00

Abstract. Information about speech reliability can be extracted and then integrated in a recogniser by various means. The full combination (FC) approach allows the weighting of the posterior values estimated locally in the time frequency representation, according a speech reliability measure. Since most of the speech segments are voiced, we use a method exploiting the harmonicity of speech tos derive these weights. We test this method together with the direct integration of the a priori SNR. Then, we run speech recognition with different kind of weighting functions. The weights are continuous or binary values. This corresponds to a soft or to a hard decision function about the speech reliability, which is derived from an observable harmonicity index. Using a binary decision process, the effect is, for each time frame, to collapse the set of combinations of sub-bands into a single combination. On the other hand, we substitute empirical values to these terms, including functions of the a priori SNR, which are continuous or discrete, but not based on a probabilistic estimation. We establish the average scores in % WER for a panel of noises at different levels, stationary or not, narrow-band or wide-band. All these functions are found to be sub-optimal comparatively to the constant weighting, but a robustness of the FC for narrow-band noises is observed.

1 Introduction

Preliminary experiments leads us to expect that multistream speech recognition can be made more robust with the inclusion of estimates of the stream's reliability [5, 2]. In this field, reliability or weighting factors are empirically estimated, or calculated using a probabilistic method. Several techniques are available for generating subband speech reliability. However these often require a frame duration which is too long to provide accurate estimates in an environment where the noise changes rapidly. The SNR estimation proposed in [6] demonstrate this situation. In this paper we propose a short term reliability [1] measure based on a harmonicity index.

We will see that this harmonic index is well correlated with the SNR and provides solutions for estimation of weighting factors. We will compare this kind of weights to optimal estimation using the true SNR value. The multistream recognition is performed by the Full Combination model [7, 3], allowing an entry point for external posterior weigthing based on a speech reliability measure.

2 The Full Combination ASR

Multistream ASR aims to make an adaptive fusion of the different sources, according to the match between each stream and the set of trained data. In our case a stream J will be one of the 16 combinations of $d = 4$ subbands, including the empty stream of data x_0. We shown [7] that full-band posterior $P(q_k|X)$ for each phoneme can be written into a weighted sum of all combination of subband according to the reliability of corresponding subband posteriors. If we assume that the best estimates of posteriors is produced by the cleanest data, then the weighting factor for the stream x_j corresponds to the probability that the "data of x_j better matchs the data of the training set" (this event is called L_j). Using Bayes we estimate the posteriors of the full band X as:

$$P(q_k|X) = \sum_{j=0}^{2^d-1} P(q_k, L_j|X) \simeq \sum_{j=0}^{2^d-1} P(q_k|x_j).P(L_j|X). \qquad \text{(Eq0)}$$

In our approach each basic event is relative to the local time frequency cell subband : $P(SNR_i > T_i)$ where SNR_i is its SNR. This local probability is given by a detector which is more or less sensitive (good detection) and specific (noisy cells detected as noisy). Therefore a continuous weighting factor is more appropriate than a binary one as this will be confirmed by this study.

It has been shown that (1) FC is more efficient than other common subband models [7], and (2), avoiding the need to train $2^d - 1$ experts because good estimates of subband combinations can be obtained from a product of the d subband experts. We will only use this approximation approach AFC. Let $|J|$ be the number of subbands in the stream J. If we assume[1] that the subband data vectors x_i are independant given class q_k we then have :

$P(x_j|q_k) \simeq \prod_{i\in J} P(x_i|q_k)$, then[2] :

$$P(q_k|x_j)\frac{p(x_j)}{p(q_k)} \simeq \prod_{i\in J} P(q_k|x_i)\frac{p(x_i)}{p(q_k)} \qquad (1)$$

$$P(q_k|x_j) \simeq \frac{\prod_{i\in J} P(q_k|x_i)}{p^{|J|-1}(q_k)} \cdot \Theta \qquad (2)$$

Link with partial recognition

The estimation technique for weighting factors which is introduced by the AFC model can be empirical or, preferably, derived from a probabilistic approach as shown in next section. Then let C_i be : "Time

[1] weaker assumption than a complete independence

[2] Θ is canceled by normalisation.

frequency cell feature vector in subband i matches the clean speech vector" [3]. Assuming that all "C_i" are independent (which is almost always the case for non-adjacent subbands [8]), then weight for a stream J will be estimated as :

$$P(L_j|X) = \prod_{i \in J} P(C_i|x_i). \prod_{i \notin J}(1 - P(C_i|x_i)).$$

Partial recognition is run if the probability $P(C_i|x_i)$ is binary. In that case, the cells where speech is occluded by noise are ignored and FC is similar to a partial recognition process [4]. In [2] we performed a marginal partial recognition which remains feasible thanks to spectral redundancy of formants. However this model is limited by its noise detection performance. Probabilistic model carries more information than a binary mask indeed the local noise detector is not perfect and corresponding soft decision probabilities can be assigned to the terms $P(L|X)$ of AFC. We will now compare AFC to partial recognition, under various conditions (we use prefix B for binary functions, S for Soft).

3 A Probabilistic estimation of short time speech reliability from harmonicity

We develop here technique for estimating the probability that a cell is corrupted by noise. Most speech is composed of voiced segments. Therefore, the autocorrelogram of the demodulated signal can be used as a basis for differentiating between harmonic signal and noise. An interesting property is that this differentiation has been shown to be efficient with a time window in the same range than the average phoneme duration [2], and in a frequency domain divided in four subbands.

A correlogram of a noisy cell is less modulated than a clean one. We use that fact to estimate the reliability of a cell for which time and frequency definitions are compatible with the recognition process (125 ms of duration). Before the autocorrelation, we compute the demodulated signal after Half Wave Rectification followed by Band-Pass Filtering in the pitch domain ([90, 350] Hz). We calculate for each cell the ratio $R_i = R1/R0$, where $R1$ is the local maximum in time delay segment corresponding to the fundamental frequency and $R0$ the cell energy. This measure is comparable to the HNR index [9]. We construct the histogram of R_i relatively to its local SNR. Initial data population comprised 60 sentences of the training set with added gaussian white noise at ($SNR = [-21 - 18...39]dB$). The distribution of R_i relative to local SNR, Fig 1, show the strong correlation between SNR and R_i which has been demonstrated in [1]. We extract from these distributions the Probability Density Function at a given cell

$$P(SNR_i = T_i|R_i, x_i)$$

where SNR_i is the local SNR of the cell.

Based on the histograms we construct the Cumulative Density (Fig.2) of each subband for a given SNR threshold T_i : $P(C_i|x_i) = P(SNR_i > T_i|R_i, x_i)$ For various subbands we get functions (called Spro) with similar graphs that are shifted on the R_i axis depending on the subband definition. We set Ti at 0dB according to experimental results [3]. The left part of the Mi function is built upon very few samples, so the function is not well defined, but this is not a major issue because only a few test samples are concerned.

Partial recognition (Bpro) is easily derived :

if $P(C_i) > 1/2$ then $P(C_i) = 1$ else 0.

Actually this threshold value depends on the detector R_i and can also be estimated from the ROC curve.

[3] For simpler notation we do not note the time variable. In this paper equations are written for a given time frame called cell.

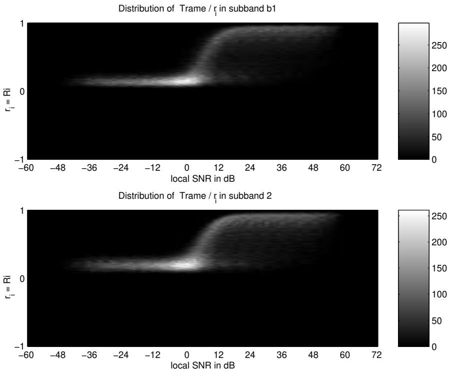

Figure 1: Histograms of the 2 first subband. Note the strong nonlinearly correlation between SNR and Ri

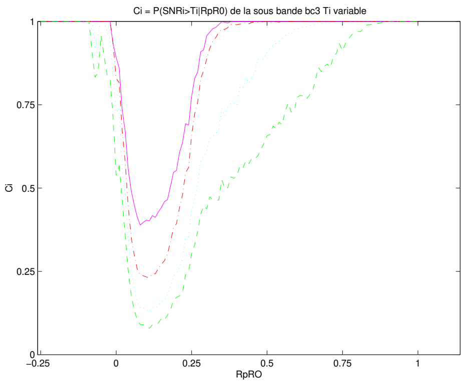

Figure 2: Reliability function in subband 3 for threshold T1 = -6,0,6,12 dB (from top to bottom). Similar functions are found for other subbands.

4 Empirical weighting factor

$P(C_i)$ can be estimated by an empirical weighting factor as long as it is correlated with the reliability of the different input streams. At first the empirical weighting factor is the value of R_i after rectification. We measured that its square root performs better (Method SqRS). The BsqR method is derived by thresholding SsqRS, we choose the threshold 0.5.

In order to refer to optimal values, we use the SNR calculated from clean and noisy samples on each time-frequency cell of 125 ms. The SNR weight is a priori chosen as

$$sqXNR = sqrt(1/(1 + 1/(10^{SNR/10}))).$$

The same decision threshold as the previous method is applied (BsqXNR) corresponding roughly to $SNR = 0dB$. We also studied another empirical weight (Srel) derived from a detector of speech pause [6]. In a similar way we estimate the relative SNR index

$$P(C_i) = \frac{(SNR_i - SNR_{iMIN})}{(SNR_{iMAX} - SNR_{iMIN})}$$

where SNR_{iMIN} and SNR_{iMAX} are the extreme values determined from statistics on our noisy subset. $SNR_{iMIN} = -64$ dB for all subbands i, but SNR_{iMAX} equals $[63, 54, 52, 43]$ dB for each respective subband.

The different methods are listed below with their characteristics. For all the binary functions $Td = 0.5$. fullband = no weight. Soft functions : blind = proba, given (equalised) / SsqXNR = empirical, given / Srel = empirical, given / SsqR = empirical, estimate / Spro = proba, estimate. Binary function : BsqXNR = empirical, given / BsqR = empirical, estimate / Bpro = proba, estimate.

5 Recognition evaluation

We use a hybrid ANN/HMM system and Numbers95, a multispeaker free format numbers telephone speech database. Our model is trained on 9 consecutive data frames. The posterior probabilities estimated by the ANN (about 1500 hidden units), divided by their priors, are passed as scaled likelihoods to a HMM for decoding using a 1 to 3 repeated-state model. No language model is used. For AFC only five ANN are trained: 1 for full band and 1 for each subband. We choose 1 subband for approximately each formant, and we carefully defined the subband with the minimal frequency overlap using the PLP filter bank. Frequency ranges are in Hz 115-629; 565-1370; 1262-2292; 2122-3769; 115-3769. The respective extracted coefficients are 5, 5, 3, 3, 11.

We tested two narrow band noises 300 Hz wide and centered in different subband. We also use a nonstationary noise composed of periodic sequences $[1, 2, 3, 4, 4, 3, 2, 1]$ of respective noisy subband number [2]. We used natural Factory noise from Noisex and a Daimler car noise.Tests were constructed by averaging scores obtained with 200 utterances repeated at 6 different SNRs from $-12to18$ dB, by step of 6db, silence included. All the features are processed by Jrasta (which is referenced in [5]).

	gwn	fact	car	narb1	narb3	n.st
fband	38.2	37.8	33.7	26.6	30.8	90.6
blind	46.9	45.6	44.2	24.5	21.7	49.9
Srel	47.2	45.0	43.9	21.4	20.1	51.9
SsqXNR	60.0	57.3	55.8	23.5	20.4	62.8
SsqR	47.7	45.6	44.8	28.9	20.4	49.6
Spro	47.3	45.0	45.0	27.1	19.3	59.8
BsqXNR	61.5	58.5	57.4	24.2	20.6	64.5
BsqR	60.9	57.6	53.6	45.9	30.6	67.1
Bpro	58.1	54.8	51.7	34.8	23.5	66.1

Table 1: Word Error Rate (WER) in % average on 200 sentences* 6 levels. Col: Gaussian White Noise, factory, car, narrow banb 1 and 3, nonstation. noise. Raw: fband : full band alone. Confidence interval = +-1 at WER=20%. Partial recognition of three subbands after exclusion of noisy sb1 or 3 in the case of narb1 or narb3 gives 22.7 or 19.0 WER%.

6 Discussion and conclusion

We tested various approaches to weight the posterior values estimated locally in the time frequency representation. We get some little significative improvement with Spro in narb3 noise. The definition of T_i of Spro method is an issue when optimizing the AFC's interface, but extensive experiments [3] show that there is no clear optimum value. A method's performance depended on the noise structure : wide band (gwn,factory, car noise) or narrow band noise (b1, b3) and whether it was stationary or not. AFC system will always outperform the full band when a significant part of each subband remains clean and the rest is stationary or not noise (90 % versus 50 % WER in the nonstationary case). The optimality of the constant weighting for narrow band noises in band 1 or 3, or in the case of nonstationary noise, demonstrates that AFC approach offers the potential rapid adaptation to changing and unpredictable narrow noise. This is due to unreliable posteriors which are minimized during normalisation (Eq2).

The main observation is whathever the weighting function integrated in Eq0. the result is worse than the use of a constant weighting value (blind). This effect does not depend on the nature (probabilist or empirical), neither on the support (estimated speech reliability or SNR given) of the variable which is introduced. Moreover, step functions, as binary decision functions, are worse than continuous functions, due to the largest difference with the constant function. So we conclude that the fusion of an external source of information cannot be realised well with the FC model in this condition, despite the apparent compatibility of the formalism with the introduction of such an information. This might be due to the Jrasta pre-processing which removes a lot of noise so SNR of the input signal and posteriors reliability might no longer be strongly correlated.

On the other hand, intrinsic weighting factors compatible with the FC formalism could improve the model [3]. This factors could be derived from outputs of the recognition system as well as a priori knowledge about the streams' reliability (e.g. the weighting of each subband or each stream by its reliability for speech recognition, knowing a priori that lower frequency subbands carry more phonetic information, or knowing that the larger streams are more reliable).

Acknowledgments We thank Hervé Bourlard (IDIAP) for his help.This work was supported by EC SPHEAR project and the EC RESPITE project.

References

[1] F. Berthommier and H. Glotin. A new snr-feature mapping for robust multistream speech recognition. In Berkeley University Of California, editor, *Proc. Int. Congress on Phonetic Sciences (ICPhS)*, volume 1 of *XIV*, pages 711–715, San Francisco, August 1999.

[2] F. Berthommier, H. Glotin, E. Tessier, and H. Bourlard. Interfacing of casa and partial recognition based on a multistream technique. In *Int. Conf. on Spoken Language Processing (ICSLP)*, volume 4, pages 1415–1419, Sydney, 1998. Australian Speech Science and Technology Association, Incorporated (ASSTA).

[3] H. Glotin. *Elaboration et étude comparative d'un système adaptatif de reconnaissance robuste de la parole en sous-bandes : incorporation d'indices primitifs F0 et ITD.* PhD thesis, Doctorat de l'Institut National Polytechnique de Grenoble, 2000.

[4] P. Green, M. Cooke, and M. Crawford. Auditory scene analysis and hidden markov model recognition of speech in noise. *TransSP*, pages 401–404, 1995.

[5] H. Hermansky, S. Tibrewala, and M. Pavel. Towards asr on partially corrupted speech. In *Int. Conf. on Spoken Language Processing (ICSLP)*, pages 462–465, 1996. origine multiband avec Osaka.

[6] H. Hirsch and C. Erlicher. Noise estimation techniques for robust speech recognition. In *Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 153–156, Detroit MI, may 1995.

[7] A. Morris, A. Hagen, H. Glotin, and H. Bourlard. Multi-stream adaptive evidence combination for noise robust ASR. *Speech Communication*, to appear, 2000.

[8] H. Steeneken and T. Houtgast. On the mutual dependency of octave-band-specific contribution to spech intelligibility. In *Proc. European Conf. on Speech Communication and Technology (EUROSPEECH)*, pages 1133–1136, 1991. Tres bon : montre la base de la reco sous bande, et le probleme des contributions des sous bqndes, bonnes references sous bandes.

[9] E. Yumoto, W.J. Gould, and T. Baer. Harmonic to noise ratio as an index of the degree of hoarseness. *JASA*, 1971:1544–1550, 1982.