Robust multi-stream speech recognition based on the combined reliabilities of speech signal (voicing) and phonemes estimates : the Posteriors Bias Prediction model

Hervé Glotin

ICP, INP Grenoble, France
IDIAP, EPF Lausanne, Switzerland
glotin@idiap.ch

Abstract

We discuss the fusion of speech and phoneme estimates reliabilities in a multi-stream Automatic Speech Recognizer (ASR) in order to improve recognition score in adverse condition. Recently the Full Combination approach (FC) [1] proposed a decomposition of the full-band posterior probability for each phoneme into a reliability weighted sum of corresponding combination posteriors. Actually we have shown in [2] that weighting factors in the FC should take into account not only the speech signal reliability, but also the intrinsic efficiency of sub-band experts. To tackle this problem we derive here a new model called ``Posteriors Bias Prediction" (PBP) in order to introduce reliability functions of each combination posteriors. We show that FC is a particular case of PBP. We show how PBP allows the integration of sub-stream reliability functions depending of the Signal to Noise Ratio (SNR) and the phoneme's class. Tests on telephonic free digits (Numbers95) under various noises demonstrate that PBP performs better than FC. We discuss the ameliorations that could be done to PBP implementation to increase its robustness to noise interference.

1. Introduction

Multi-band processing paradigm for noise robust ASR was originally motivated by the observation that human recognition appears to be based on independent processing of separate frequency sub-bands [3].

In the context of hybrid multi-band ASR [1], phoneme posterior probabilities can be estimated from each sub-band combination and increase ASR robustness to narrow-band noise [2]. Of the different multi-band models which have been proposed, only the ``Full Combination" approach (FC) [1,2] allows us to consistently overcome the difficult problem of combining sub-bands recognition, by integrating over all possible positions of noisy sub-bands.

We develop here some ameliorations of the FC model. The goal of any ASR system is to reliably detect the presence of a phoneme at a given time. The most informative event for recognition is tk. : ``the phoneme k occures at time t".

Let be qk : ``the phoneme k is detected (according to Maximum A Posteriori (MAP) criterion) at time t". Let J be the set of all possible data vector X, and fj="the recognizer observes only stream j at time t". Because fj are mutually exclusive and collectively exhaustive we have :

$$P(qk|X)=\sum_{j\in J}P(qk,fj|X) \qquad (1)$$

(1) is the decomposition used in FC model. Actually one can substitute qk by tk in (1). In FC model events tk and qk are collapsed. We propose to analyze in detail the relation between this two events, so that events tk and qk are not collapsed. Actually, we will demonstrate that tk and qk are different, introducing a bias in phoneme's estimations.

2. The Posteriors Bias Prediction Model

We have seen that equation (1) is equal to :

$$P(tk \mid X) = \sum_{j\in J} P(tk, fj \mid X) \qquad (2)$$

We always have :

$$P(tk \mid X) = \sum_{j\in J} [P(tk, qk, fj \mid X) + P(tk, \overline{qk}, fj \mid X)]$$

Using bayes rule and $Xj = X \cap fj$, we have :

$$P(tk \mid X) = \sum_{j\in J} P(fj \mid X).[(P(tk \mid qk, Xj) - P(tk \mid \overline{qk} \ , Xj)).P(qk \mid Xj) + P(tk \mid \overline{qk}, Xj)\,]$$

This equation includes :

$\varphi^{+}(k, Xj) = P(tk \mid qk, Xj)$ = the reliability of positive estimation, and

$\varphi^{-}(k, Xj) = 1 - P(tk \mid \overline{qk}, Xj)$ = the reliability of negative estimation.

These reliabilities are correlated with the rate of true positive and true negative phoneme's detections. They completely characterize the performance of any detector.

Let note $\phi(k, Xj) = (\varphi^{+} + \varphi^{-})(k, Xj) - 1$, then we see that we get a formula similar to the "Full

Combination" model, but includes a linear transformation of the posteriors :

$$P(tk \mid X) = \sum_{j \in J} P(fj \mid X).[\phi(k,Xj).P(qk \mid Xj) + 1 - \varphi^{-}(k,Xj)]$$

This is the PBP model. We show in [4] that FC is an ideal case of PBP : they are similar if the recogniser collapses tk and qk, this means that if we consider a MAP detector, its sensibility and sensibility are equal to one, so that this detector makes no mistakes, which is not realistic.

3. Link between reliability functions and voicing

We have shown in [4] that reliability functions $\varphi+(k,Xj)$ and $\varphi^{-}(k,Xj)$ are SNR dependant. Actually PBP model enables to optimally link intrinsic (MAP) and any signal reliabilities like SNR or the voicing index R which has been developed in [5]. We will focus here on the use of R, calculated from normalized autocorrelation pic in the pitch domain of time frequency cells (of around 128 ms * 700 Hz). R is calculated only in the 4 sub-bands 1,2,3,4, and then combined in order to approximate other R sub-streams. We have shown that R is well correlated with SNR [5]. We approximate R for any sub-stream with the sum of the pic of the sub-bands included in the sub-stream, divided by the sum of the energy of each of these sub-bands. This estimation is well correlated with the direct sub-stream measure (the correlation minimum equals to 98% over the different stream, with Gaussian white noise at different SNR over the whole test set of numbers95).

We then study the reliability functions $\varphi+(k,Xj)$ and $\varphi^{-}(k,Xj)$ mapped from confusion matrices based on the phoneme detected with the MAP criterion and the target phoneme, at different R interval and for different stream on noisy subsets of a development test set (200 utterances with Gaussian white noise at different SNR). For this first experiment we group phonemes' class in 3 subsets : Voiced (Voi.), Fricative non voiced (/s,f,th,hh/) (Fricat) and Plosive non voiced (Plos). (see figures 1 and 2).

We observe in figure 1 and 2 that reliabilities are phoneme- and stream-dependent. They are also depending on the automatically estimated voicing index R. Naturally reliability for Voi. is well correlated with R. For Silence, reliabilities are always strongly correlated with R in the case of the negative estimation (figure 2), but we observe the contrary in case of positive estimation. We see that these state (representing around 40% of the phonemes on Numbers95), are generating very often wrong recognition at low but also at high (see figure 1) SNR. PBP aims to correct their estimations using a priori functions of these bias. Non voiced fricative (Fricat.) are mostly silence when R is high for High Frequency streams (HF), then they share nearly the same reliability.

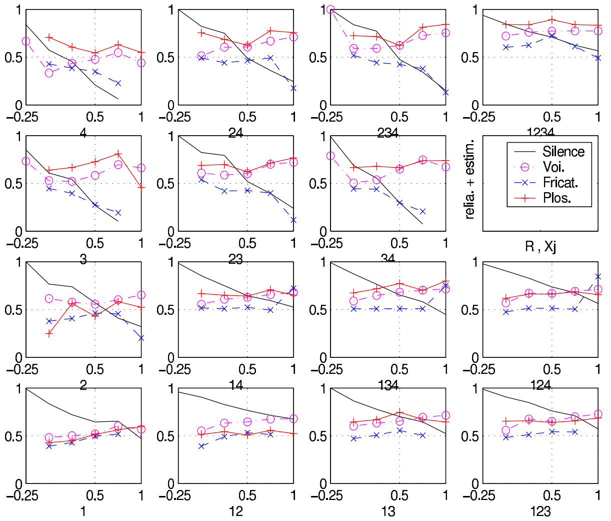

Figure 1 $\varphi+(k,Xj)$ for the 15 different sub-streams, the 3 phoneme subsets and the different R values. Sub-streams are noted by the sub-bands they contain (e.g. 134 is the stream containing sub-bands 1, 3 and 4. Full-band is the stream 1234). R is varying from –0.25 to 1 (abscissa). Reliability (ordinata) is varying from 0 to 1. We group phonemes in 3 subsets : Voiced (Voi.), Fricative non voiced (Fricat) and Plosive non voiced (Plos).

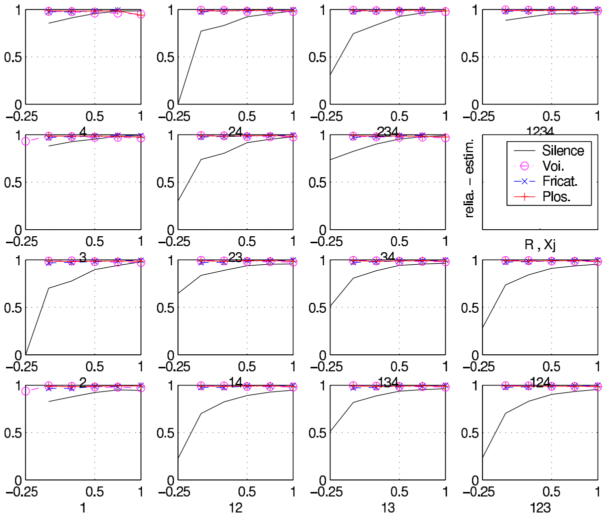

Figure 2 $\varphi-(k,Xj)$ for the 15 different sub-streams, legend is the same than in figure 1.

In the case of $\varphi^{-}(k,Xj)$, all the sub-stream functions are similar. Only silence, for low R differs. In the case of $\varphi+(k,Xj)$ and non silence phonemes, we see that we have similar behaviour for streams (1, 2, 12, 13, 14, 123, 124, 134) which are the Low Frequency (LF) streams, certainly because speech is mostly conveyed in LF. The HF streams generate another kind of behaviour. In particular, the Plos. have different function which is correlated with the fact that plosives release contains mostly silence in LF and most of the information is conveyed in HF.

4. Recognition tests

We can compute the reliability P(fj|X) of any stream j using R index and the SNR mapping technique as in [5], or we can consider that any stream is equally useful (blind model), and that the information about stream reliability will be given through $\varphi+(k,Xj)$ and $\varphi^{-}(k,Xj)$ functions. We will consider only this second method for this first implementation of the PBP model.
We choose 1 sub-band for approximately one formant (see table 1), and we carefully defined the sub-band with the minimal frequency overlap. All the features are the JRASTA [6].

sub-band	1	2	3	4
Hz	115 629	565 1370	1262 2292	2122 3769

Table 1:Set up of the 4 sub-bands and the full band in Hz (cut off 3dB) with their respective LPC order and number of coefficient extracted.

We are using the Numbers 95 database of multi-speaker US English free-format numbers telephone speech, with around 50 words, 27 phonemes. The training set is composed of 3500 utterances, the test set is composed of 200 utterances at various SNR dB : -12 -6 0 6 12 18 dB. Preliminary noise set is composed of 4 limited band noise named Bx, 300 Hz large and centred in each sud-band (x indexes the sub-band). We also use a periodic mixture of these Bx : each 125 ms, x is regularly picked up from the sequence [1,2,3,4,4,3,2,1].

Baseline tests were made with a full-band hybrid ASR in which neural networks with one hidden layer of 1700 units, using a context of 9 consecutive frames, generates the posterior probabilities P(qk|Xj) for each of 27 phonemes. Training used the full Numbers95 training set. During recognition, posterior probabilities divided by their priors, were passed as scaled likelihoods to a fixed parameter HMM for decoding. Sub-band posterior estimations are easily combined [1] to generate sub-stream posteriors.

	GWN	FACT	CAR	B1	B3	N.ST.	MEAN
Full Band	38.2	37.8	33.7	26.6	30.8	90.6	42.9
Spect. Sub.	33.7	41	35.6	29.2	38.4	56.3	39
FC blind	46.9	45.6	44.2	24.5	21.7	49.9	38.8
FC + R	47.3	45	45	27.1	19.3	59.8	40.6
PBP blind	46.1	43.7	42.4	25.3	21.1	44.3	37.1

Table 2 : Word Error Rate (WER) in % average on 200 sentences 6 levels (-12,-6,0,6,12,18 dB SNR). Col.: GWN:Gaussian White Noise, factory, car noises, B1(resp B3) narrow banb noise in band 1 (resp B3),NST: non-stationary noise. Fband= full band system with Jrasta processing. Spect Sub.=Spectral Substraction. FC blind= FC with uniform weights, compared to FC+R using R signal confidence (see [2]). PBP blind = PBP model using uniform weight, but the two functions described in figures 1 and 2 . Confidence interval for the means = +-0.5 at WER=20\%. Partial recognition of three sub-bands after exclusion of noisy sub-band 1 or 3 in the case of noise b1 or b3 gives 22.7 or 19.0 WER%.*

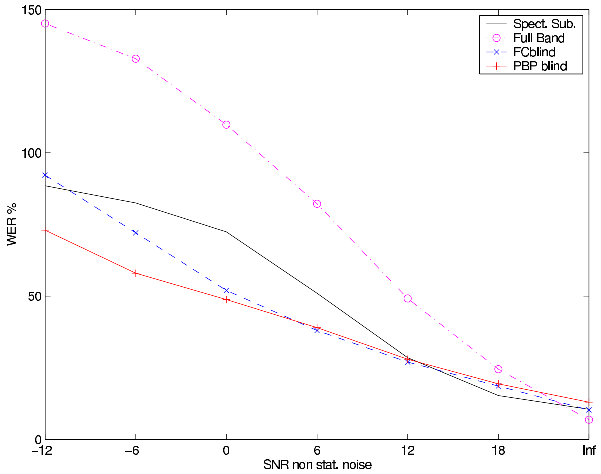

Figure 3 Details of WER% (ordinata) for the non stationary noise at different SNR in dB (abscissa) and the different models : Full band Jrasta, Spectral Substraction, Full Combination and PBP model using blind weights and the two functions described in figures 1 and 2.

The HMM for each phoneme used a 1 to 3 repeated-state model. No language model was used.
For the full-combination multi-band system a separate MLP (of identical design to the full-band MLP) was trained on clean data for each sub-band combination. Multiple MLP outputs were then merged [1] at the frame level (which here was also the state level) to give a single posterior probability for each class.

5. Discusion

We observed in the figure 1 and 2 the well known fact that a given frequency domain encodes more or less efficiently the information necessary for phoneme identification. WER % tables show that PBP performs in average better than FC model. PBP performance are detailed in the case of NST noise, showing that PBP is best at low SNR. We can assume that this is due to the fact that bias are stronger at low SNR. We can expect even more improvement if we use different reliability functions for each phonemes, and signal dependant weights P(fj|X) (but not equal or "blind" weights like in this paper).

The key result is that we have different PBP weighting functions for different phonemes and different streams and that PBP model integrates them. The architecture of PBP model is similar to a "correction system", described by Shannon in [7] (see fig. 4). During training, signal and noise, sent and received messages are observed, allowing us to build the reliability functions to pass on the "correction data" information to a correction device during testing. More comments can be found in [8]

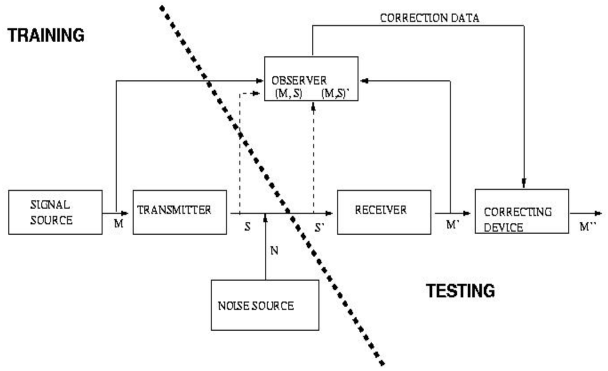

Figure 4 : Schematic diagram of a correction system inspired from [7]. During training, signal and noise, sent and received messages are observed, allowing us to build the reliability functions. During testing, only the corrupted signal s' and message M' are known, but the observer can use the reliability function to pass on the "correction data" stream information to a correction device (like our Posteriors Bias Prediction).

6. Conclusion

We analyse in detail the confusion matrices of each stream in order to characterize the predictive values of correct estimations (negative or positive) of each phoneme using the criterion of Maximum a Posteriori probabilities. We build these predictive value functions and we show that they are phoneme- and stream-dependant.

In order to use this information we develop a new model: the "Posterior Bias Prediction". We show that the previous "Full Combination Model" is a particular case of PBP. The PBP model allows us to optimally link intrinsic reliabilities (based on the MAP criterion) and extrinsic signal reliability SNR like (e.g. voicing [5]). We discuss the relation of PBP with the correction system proposed in [7]. Strong improvement in robustness of phoneme and word recognition is expected using more detailed functions, using a particular reliability function for each phoneme.

Acknowledgments

This work was supported by EC SPHEAR project and the EC RESPITE project.

References

1. Morris, A., Hagen, A., Glotin, H. and Bourlard, H., "Multi-stream adaptive evidence combination for noise robust ASR", *Speech Communication, Special Issue on Noise Robust* Vol.34 (1-2). 2001.

2. Glotin, H. and Berthommier, F., "Test of several external posterior weighting functions for multiband full combination ASR", *in Int. Conf. on Spoken Language Processing (ICSLP),* Beijing-China. 2000.

3. Allen, J., "How do humans process and recognize speech?", in *IEEE Trans. on Speech and Audio Processing*, 2(4) pp 567-577. 1994.

4. Glotin, H., "Optimal fusion of expert's confidence and speech reliability for robust multi-stream ASR : the bias prediction model ", in *IEEE International Workshop on Intelligent Signal Processing*. 2001.

5. Berthommier, F. and Glotin, H., "A new SNR-feature mapping for robust multistream speech recognition", in *B. University Of California, ed., Int. Congress on Phonetic Sciences (ICPhS),* Vol. 1 of XIV, San Francisco, pp.711-715. 1999.

6. Hermansky, H. and Morgan, N. , "RASTA processing of speech", *IEEE Transactions on Speech and Audio Processing*, 2(4), pp 578-589. 1994.

7. Shannon, C., "A mathematical theory of Communication", Technical Report 27, Bell System Technical Journal. 1948.

8. Glotin, H., "*Elaboration et étude comparative de systèmes adaptatifs multi-flux de reconnaissance robuste de la parole : incorporation d'indices de voisement et de localisation.*" PhD thesis, Institut National Polytechnique de Grenoble. June 2001.