Correcting Posteriors by using a feedback synthesis loop in robust ASR

Hervé Glotin

ERSS-CNRS
5 all. Machado; Toulouse Cedex 1 - France
glotin@univ-tlse2.fr

ABSTRACT

Current Automatic Speech Recognition (ASR) systems are not efficient in noisy speech conditions. We propose a new strategy to reinforce ASR robustness, based on a feedback loop from recognition of posteriors to signal synthesis. The key idea is to use phonemes' posteriors generated by recognition to calculate an acoustic image (AI) at each frame and to calculate its correlation with the input signal. AI is the weighted sum phonemes clean speech spectrum, where weights are directly taken as the corresponding phonemes' posteriors. Correlation between AI and the input spectrum gives a Recognition Index (RI). We then show how a simple correction function of posteriors' distribution using RI improves the Word Error Rate in a continuous speech recognition task compared to a state of the art ASR system (Jrasta).

1. INTRODUCTION

Current Automatic Speech Recognition (ASR) systems are not efficient in noisy speech conditions. We first present various strategies proposed to reinforce ASR robustness (Wiener filter, multistream approach, enhanced decoding - see figure). Then we propose a new strategy based on a feedback loop from recognition to signal synthesis.
The key idea is to use the recognition of phonemes to construct an acoustic image which is compared to the input signal. This derives from Gibson [1], who points out that information cannot be said to cause perception : « Perception … is never fully stimulated but instead can go into activity in the presence of stimulus information ». Some principles were outlined in [7], but here we present how it can be successfully used for robust ASR.

2. DIFFERENT ROBUST ASR STRATEGIES

For more than ten years, many researchers in ASR systems have focused on a weighting approach. Generally they enhance reliable features [2,3] or phonemes estimates [3,4,5,6]. We show in this section that they can be classified as "forward" systems, but that one can also elaborate "backward" systems as introduced in [7].
First, the Wiener approach can be represented by path 1 in figure 1, using Signal to Noise Ratio (SNR) information. But path 1 also represents the weighting approach used in general multi-stream ASR. Another technique consists of weighting the estimates after recognition based on the SNR: this is represented by path 2. In that case, each stream feeds one expert recognizer. Their estimates are combined through a fusion process and sent to the decoder. Depending on their SNR, some streams are best suited for the phoneme transmission. It is then interesting to overweight the best stream at a particular time during the fusion process. Results using audio and visual modalities [3,4,5] demonstrate the significant advantage of the weighted multi-stream approach in adverse conditions.
We see that path 1 and 2 feed the system from the SNR, and are thus classified as forward systems.

Another kind of system relies on the estimation of the quality of the phonemes estimates, which we call the Posterior to Noise Ratio (PNR) [3,10,11]. Therefore, after the recognition step, the PNR estimation feeds into the fusion of estimates [11] (path 3 in figure 1) or features extraction (path 4) process.
In this paper, we present a new method for estimating the PNR using a "Proactive" approach or feedback loop represented by path 4 in figure 1.

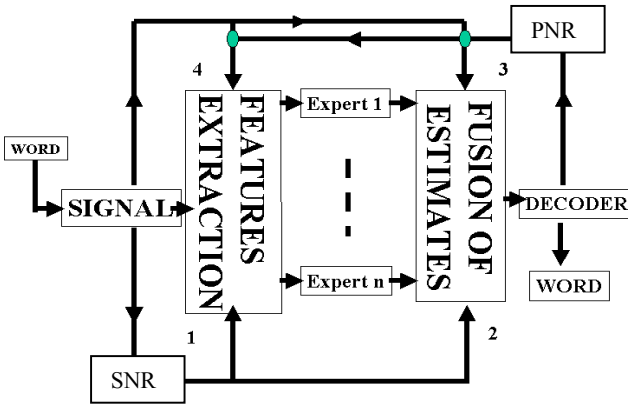

Figure 1: FOUR DIFFERENT ASR STRATEGIES. We represent elements of any ASR (here with n streams) : features extraction, expert recognition and fusion of their estimates, and finally the decoder and word generation. One can investigate four different correction or weighting strategies through the four paths described below.
Path 1: Wiener filtering using SNR estimation.
Path 2: SNR estimation can be used to control the fusion of different experts.
Path 3: The phonetic context can be used during expert fusion.
Path 4: feedback loop: the "Proactive ASR". We build an Acoustic Image (AI) in each stream. Correlation measure between PSD(AI) and the input frame's PSD is correlated with the quality of estimates. This information can be used during the extraction process or passed directly to the fusion process (3). One can also use path 4 in a mono-stream ASR.

3. RELIABILITY AND WEIGHT ESTIMATION

Usually weights are estimated for path 1 and 2 from the SNR step, derived from speech characteristics [3,4]. In path 3, the entropy of posteriors' distribution or mapping functions can be used [3].
In our approach, which follows path 4, we propose to derive these weights or Recognition Indices (RI) from a correlation measure between the input signal X and an Acoustic Image constructed from the estimates of phonemes P(qk|X), as introduced in [7], and we show how this information can be integrated into a simple mono-stream ASR.

The Feedback loop consists of building an Acoustic Image (AI) in each stream. Useful information can be extracted in path 4 if PSD(AI) and the input frame's PSD is correlated with the quality of estimates.
Therefore we use a Phoneme to Power Spectrum Density (PSD) mapping.
Let PSDm(k) = the mean PSD of phoneme k over the segmented training set.
Then we calculate the PSD(AI(t)) for frame at time t:

$$\mathbf{PSD(\ AI(t)\ ) = \sum_k P(qk \mid X(t))\ *\ PSDm(k)}$$

Then we directly calculate the Recognition Indice:

RI(t) = Correlation(PSD(X(t)) , PSD(AI(t)))

Results (see database reference in section 5) with 0dB Gaussian White noise show a good anti-correlation between RI and KullbackLeibler (dKL) distances of the phonemes' estimates with the target posteriors distribution (p(qk)=1 for the target phoneme, else 0). This correlation is -0.68 for the full spectrum. We confirm then that the larger the distance is, the lower the RI is. More correlation results are discussed later in section 6.

4. POSTERIORS ENHANCEMENT

We have seen in the previous section that we can calculate a quality factor RI of posterior distribution based on the KL distance. This quality factor is also correlated with the value of the Maximum a Posteriori (MAP) value. The higher (lower) the MAP value is, the better (worse) the recognition is.
As shown in figure 1, this information can then be passed directly to fusion process (path 3).
It is then intuitive to weight exponentially the posteriors frame value with their quality factor RI(t). Therefore we take :

P'(qk | X(t)) = P(qk | X(t)) ^ (RI(t) ^ N)

Then the P' posteriors are normalized and given to the decoder. N is an empirical factor which modulates the effect of the RI. Since the RI's confidence fluctuates, the best results are obtained using a nonlinear RI function. Therefore, we take RI^N. We tested N=1,2,3,4,5,6,7,8,9. As discussed

in the next section, the best results have been obtained using N = 5.

5. WORD RECOGNITION RESULTS

Tests were done on 200 utterances from the Numbers 95 database. NB95 is composed of multi-speaker, US English, free-format numbers telephone speech, with around 50 words and 27 phonemes. We used 128 ms frames in full spectrum. All PSD mapping in this study are only 16 bins long which requires fewer CPU resources.

Phoneme' estimates are produced by an ANN ASR [3,5]. All the features are the JRASTA [8]. The training set is composed of 3500 utterances.

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, generate the posterior probabilities P(qk|Xj) for each of 27 phonemes. During recognition, posterior probabilities divided by their priors, were passed as scaled likelihoods to a fixed parameter HMM for decoding. For each phoneme the HMM used a 1 to 3 repeated-state model. No language model was used.

5.1 Optimisation of factor n

Since the reliability of our Recognition index fluctuates, it is interesting to modulate its value. We show in figure 2 the WER for different N factors from 1 to 9, on 200 noisy utterances from the NB95 development test set, with 0dB Gaussian White noise.

The minimum WER is 31.4% for optimal N=5, (baseline system is 34.4% WER). Since (RI^5) is nearly null (<0.01) for RI < 0.5, this means that (RI^N) factor is optimal for the system when RI is higher than 0.5.

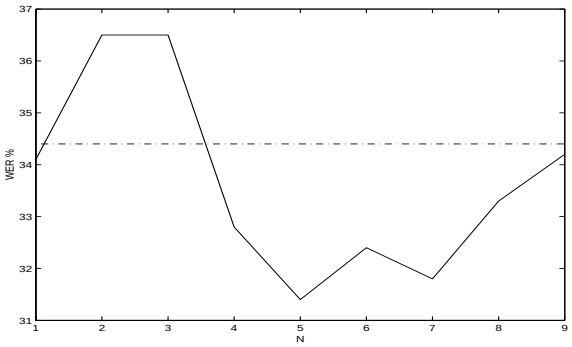

Figure 2 : optimization of factor N is based on the minimization of WER (plain line) for different N from 1 to 9, on 200 utterances from of the NB95 development test set, with 0dB Gaussian White noise. The baseline system (dashed line) = 34.4%. The minimum WER is 31.4 % for optimal N =5.

5.2 WER results on various noise

The test set is composed of 200 utterances at various SNR dB :-12,-6,0, 6,12,18 dB. The added noises are Gaussian White noise, factory noise from Noisex database and car noise from Daimler Benz [3]. Another noise set is composed of 4 limited noisy subbands named Bx, each 300 Hz large and centered in each subband (x is the subband number of Table 2). We also use a non-stationary noise as 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].

	GWN	Fact	Car	B1	B3	Nst	ME
Jrasta	38.2	37.8	33.7	26.6	30.8	90.6	42.9
PASR	35.0	34.8	30.5	24.2	28.8	86.3	**39.9**
Dwer	-8.4	-7.9	-9.5	-9	-6.5	-4.7	**-7**

Table 1 : *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 band noise in band 1 (resp B3), Nst : non-stationary noise, Mean WER. Jrasta=baseline fullband system with Jrasta processing. PASR = our "Pro Active" ASR system. Partial recognition of three subbands after exclusion of noisy subband 1 or 3 in the case of noise b1 or b3 gives 22.7 or 19.0 WER. Clean baseline = 11.2% WER. Confidence interval at 5% for the mean.= +- 0.71 at WER=40%. In the last line, we show the Delta WER in %, negative in case of improvement.*

6. DISCUSSION AND CONCLUSION

Future work will apply our PASR weighting strategy to multistream ASR. Indeed correlation with RI and SNR exist in sub-band analysis shown in table 2. Therefore backward weighting information can be applied to a system using multi-band or multi-stream weighting.

Subband	1	2	3	4	FULL
Hz	115 629	565 1370	1262 2292	2122 3769	115 3769
Corr(RI,dKL)	-0.57	-0.36	-0.48	-0.33	-0.68

Table 2 : Set up of the 4 subbands and the full spectrum in Hz (cut off 3dB) and their correlation between our Recognition Index (RI) and the KullbackLeibler (dKL) distances of the phonemes' estimates with the target posteriors distribution (p(qk)=1 for the target phoneme, else 0). Results from the 200 utterances of the Numbers95 test set.

This Proactive architecture (PASR) can be seen as a simple illustration of a sensory map system [9] and it gives new information for use in robust ASR. This PASR can be extended to multi-stream ASR. Another interesting perspective is to study the iterative property of the PASR system. Indeed, after the first loop, one can construct a new AI from the new estimates and reiterate the process. Another application of RI is to use it during the features extraction process. Such a system will be studied in future work.

We have seen that ASR systems generate a lot of word errors in noisy conditions, and a large amount of the research in robust ASR is focused on the forward approach. Therefore we present a new kind of correcting/weighting strategy based on a backward system. Then we have shown how to extract relevant information and how to integrate it in a simple ASR system to significantly reinforce its robustness.
More sophisticated systems can be based on this Proactive architecture, like recurrent correction, and will be investigate in future work.

7. ACKNOWLEDGMENTS

This work takes place in a French COGNISUD project. We thank IDIAP-EPFL for some material support.

8. REFERENCES

[1] Gibson, *The ecological approach to visual perception*. Hillsdale, 1986

[2] S.Ris and S.Dupont , "Assessing local noise level estimation methods : application to noise robust ASR", *Speech Communication, Sp. Issue on Noise Robust* Vol.34 (1-2). 2001.

[3] H.Glotin, Phd Thesis, *Elaboration of robust multistream automatic speech recognition using voicing and localisation cues*, Grenoble, June 2001.

[4] H.Glotin, D.Vergyri, C.Neti, G.Potamanios and J.Luettin, "Weighting schemes for audio-visual fusion in speech recognition", In ICASSP 2001.

[5] A.Morris, A.Hagen, H.Glotin and H.Bourlard, « Multi-stream adaptive evidence combination for noise robust ASR », *SpeechCommunication*, Vol 34 (1- 2), April 2001.

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

[7] H.Glotin. **"**Recognition using speech synthesis : a reactive dynamic for robust ASR". Accepted in *ISCA International Workshop*. To appear in Cambridge Press. Aix-en-Provence France. April 2002.

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

[9] B.Lindblom and J.Lubker, "Phonetics Linguistics", *The speech Homonculus and a problem of Phonetics Linguistics*, Academic Press, Orlando, 1985

[10] H.Glotin, "Optimal fusion of expert's confidence and speech reliability for robust multistream ASR : the PBP model", *In IEEE international Workshop on Intelligent Signal Processing*, 2001.

[11] H.Glotin, "Enhanced Posteriors Bias Prediction for robust multi-stream ASR combining voicing and estimates reliabilities". In *IEEE International Conf. ASSP*. Orlando-USA, May 2002.