ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

THE FULL COMBINATION SUB-BANDS APPROACH TO NOISE ROBUST HMM/ANN BASED ASR

Andrew Morris, Astrid Hagen, Hervé Bourlard[1]

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
P.O. Box 592, 4 Rue du Simplon, CH-1920, Martigny, Switzerland

morris,hagen,bourlard@idiap.ch
http://www.idiap.ch/

ABSTRACT

The performance of most ASR systems degrades rapidly with data mismatch relative to the data used in training. Under many realistic noise conditions a significant proportion of the spectral representation of a speech signal, which is highly redundant, remains uncorrupted. In the "missing feature" approach to this problem mismatching data is simply ignored, but the need to base recognition on unorthogonalised spectral features results in reduced performance in clean speech. In multiband ASR the results from independent recognition on a number of within-band orthogonalised sub-bands are combined. This approach more accurately reflects the uncertainty in mismatch detection, but loss of joint information due to independent sub-band processing can also result in reduced performance with clean speech. In this article the "full combination" approach to noise robust ASR is presented in which multiple data streams are associated not with individual sub-bands but with sub-band combinations. In this way no assumption of sub-band independence is required. Initial tests show some improved robustness to noise with no significant loss of performance with clean speech.

Keywords: missing feature theory, noise robust ASR, multiband ASR, HMM/ANN hybrid

1. INTRODUCTION

One of the main factors limiting ASR applications is the rapid degradation of performance which occurs due to mismatch between the data encountered during recognition and the data used in training. Under many realistic noise conditions a significant proportion of the highly redundant spectral representation of speech remains uncorrupted. The "missing feature" approach [12,14] to the mismatch problem is based on the following:

> *In noisy speech, automatic recognition can often be improved by simply ignoring the parts of the spectral signal most affected by noise.* *(rule 1)*

While this approach is attractive because it requires no assumptions about noise type or level, it has two serious drawbacks. One is that data mismatch is not easy to detect, and even when correct, data ignored in this way may contain useful speech information. Another is that the need to avoid mixing clean with noisy spectral coefficients precludes the possibility of spectral data orthogonalisation prior to recognition, which results in unacceptably low performance in clean speech.

An alternative way to exploit spectral redundancy has been motivated by the *product of errors rule* [6,1] for human speech perception whereby (under certain conditions):

> *In human perception the full-band error rate is equal to the product of sub-band error rates.* *(rule 2)*

This is formally equivalent to saying that sub-band errors are independent and that correct recognition occurs whenever there is correct recognition in any sub-band. This shows that we have a strong ability to select the right answer from multiple independent "guesses". This has inspired a lot of recent work with multiband ASR [3,11] in which recognition is performed in a number of spectral subbands and later combined. In this approach the difficult problem of mismatch detection is replaced by the problem of expert combination. However, the loss of joint spectral information which results from independent processing of sub-bands can also give reduced performance with clean speech.

In Section 2 we present the "full combination" approach to noise robust ASR in which multiple data streams are associated not with individual sub-bands but with all possible sub-band combinations. In this way the number of experts for a given number d of sub-bands increases to 2^d and no assumption of independence between sub-bands is required.

In Section 3 we show how results comparable to those obtained by training all 2^d combination experts can sometimes be obtained from a suitable approximation in terms of the function of the d sub-band experts alone.

2. POSTERIORS DECOMPOSITION

The degree of mismatch (or reliability) of any data component or sub-band is by definition relative to the training data pdf. Data reliability is therefore a stochastic quantity even before the method used for its estimation has been taken into account. Every possible selection of sub-bands therefore has an a-priori non zero probability of giving best recognition performance.

Let the 2^d different combinations of 0..d sub-bands from a set of d sub-bands be denoted c_i, i=1..2^d.

1. Also Professor at the Swiss Federal Institute of Technology, Lausanne, CH

Let the event that c_i is the combination which gives best recognition performance be denoted $best_i$.

Let the components of data vector x selected by combination i be denoted x_{c_i}.

As the events $best_i$ are exhaustive and mutually exclusive, we can now decompose the full-band phoneme posterior probability for each class q into a weighted sum of sub-band combination posteriors as follows:

$$P(q|x) = \sum_i P(q, best_i|x) = \sum_i P(best_i|x)P(q|best_i, x)$$

By the definition of $best_i$, to obtain best recognition we should select $\hat{P}(q|best_i, x) = P(q|x_{c_i})$ and

$$\hat{P}(q|x) = \sum_i P(best_i|x)P(q|x_{c_i}) \quad (1)$$

Sub-band posterior probabilities $P(q_k|x_{c_i})$ for each phoneme q_k are obtained by training an MLP "expert" for each sub-band combination c_i (see Fig.2a).

Various methods for estimating the expert weighting factors $P(best_i|x)$ are discussed in Sections 2.1 and 7.

2.1 Expert weighting estimation

By rule 1 there must exist a level of reliability r_0 in each sub-band below which recognition will improve when this sub-band is ignored. If we can measure r_0 and also estimate the reliability $r(x_j)$ of each sub-band x_j, then we can estimate the probability that any given subband combination will give the best recognition results.

If we assume that r_0 is independent of which other sub-bands are selected, and has the same value for each sub-band, and define $reliable_j \equiv r(x_j) > r_0$, then

$$best_i \Leftrightarrow (reliable_j \forall j \in c_i) \wedge (\neg reliable_j \forall j \notin c_i) \text{ and}$$

$$P(best_i|x) = \prod_{j \in c_i} P(reliable_j) \prod_{j \notin c_i} P(\neg reliable_j) \quad (2)$$

In our initial experiments, instead of estimating r_0 and $r(x_j)$, sub-band reliability was modelled as a piecewise linear function of the estimated local sub-band SNR. An SNR range $[SNR_{min}, SNR_{max}] = [0, 30]/dB$ was decided below which $P(reliable) = 0$ and above which $P(reliable) = 1$. $P(reliable_j)$ was then obtained as:

$$\hat{SNR}_j' = min(max(\hat{SNR}_j, SNR_{min}), SNR_{max})$$

$$\hat{P}(reliable_j) = \frac{\hat{SNR}_j' - SNR_{min}}{SNR_{max} - SNR_{min}} \quad (3)$$

The simple form of local SNR estimate used [3, p.9] is based on sub-band energy histograms spanning several hundred ms and is therefore not able to track highly non stationary noise.

3. FULL COMBINATION APPROXIMATION

As the number of sub-bands increases it soon becomes impractical to train a separate combination expert for every sub-band combination. It has been found that the assumption of full independence between sub-bands can lead to an unacceptable decrease in performance for clean speech. However, combination posteriors can be approximated from sub-band posteriors without making an assumption of *full* independence as follows. Let y_j denote the j^{th} sub-band in combination x_{c_i} and $b_i = |c_i|$.

$$P_{ki} = P(q_k|x_{c_i}) = P(q_k|y) = P(y|q_k)P(q_k)/P(y)$$

If we assume independence between y_j *when conditioned on* q_k then:

$$P(y|q_k) = \prod_{j=1\ldots b_i} P(y_j|y_1 \ldots y_{j-1}, q_k) \cong \prod_j P(y_j|q_k)$$

and $P_{ki} \cong \Theta \bar{P}_{ki}$ where Θ is independent of k, and

$$\bar{P}_{ki} = P^{1-|c_i|}(q_k) \prod_{j \in c_i} P(q_k|x_{c_i})$$

But $\sum_k P_{ki} = 1$ so $\Theta = 1/\sum_k \bar{P}_{ki}$ and $P_{ki} \cong \bar{P}_{ki}/\sum_m \bar{P}_{mi}$ (4)

4. EXPERIMENTS

4.1 Data preparation

Speech was taken from the Numbers95 database of US English connected digits telephone speech [5]. Car noise from the Noisex92 database [15] was added at varying SNR levels relative to the average signal energy in each utterance (excluding non-speech periods).

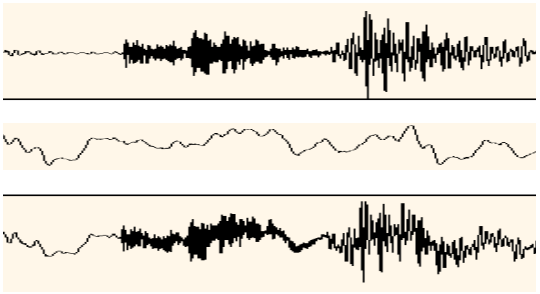

Figure 1. Waveform for section of a clean speech signal, for car noise and for speech-noise mixture at -10 dB SNR

For the purpose of local sub-band SNR estimation, data was preprocessed using bark scaled cube-root spectral coefficients, with a 25 ms analysis window and 12.5 ms between window centres.

For recognition we ran experiments using both PLP [9] and J-Rasta-PLP coefficients [10], with the same analysis window size and shift. In order that noise was not mixed between separate data substreams, the DCT orthogonalisation transform, which is used in both cases, was applied only to the data within each sub-band combination.

4.2 Recognition system

The full-combination method assumes access to a posteriori phoneme probabilities and therefore cannot be applied to HMM based ASR which is likelihood based. All tests were therefore made with an HMM/ANN hybrid [4] in which the ANN used was a one hidden layer MLP with 1000 hidden units, trained to input 9 consecutive data frames and output the posterior probabilities $P(q_k|x)$ for each of 33 phonemes q_k and each data frame, x. In recognition, scaled posterior probabilities

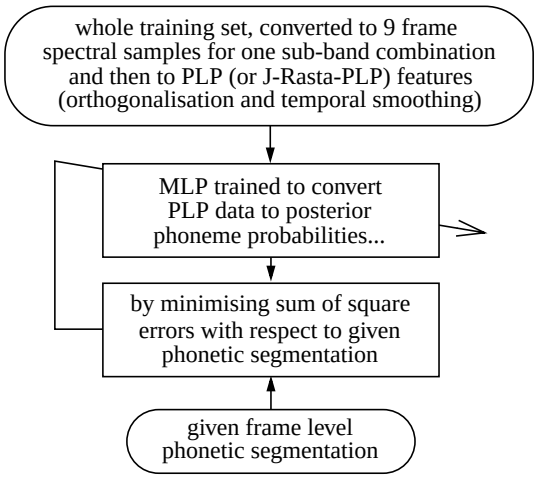

Figure 2a. Full combination hybrid system training. An MLP is trained for each sub-band combination to convert 9 frames of clean data to phoneme probabilities

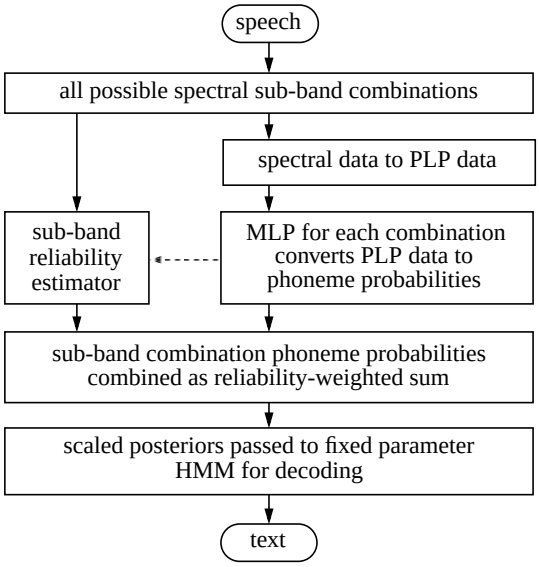

Figure 2b. FC hybrid system used in recognition[2]

from the MLP are passed to the HMM for decoding. The HMM has fixed parameters. All state transition probabilities are 0.5. Each phoneme uses a one state model for which emission likelihoods are supplied as scaled posteriors from the MLP. Phoneme specific minimum durations are modelled by forcing from 1 to 3 repetitions of the same state for each phoneme. No language model was used.

For the full-combination system a separate MLP is trained for each sub-band combination. Multiple MLP outputs are then merged at the frame level (which here is also the state level), using Eqns. 1, 2, 3 (& Eq. 4 for FC approximation), to give a single posterior probability for each class, before passing these scaled posteriors to the same HMM as used by the full-band system.

2. The input of phoneme posterior probabilities to the sub-band reliability estimator from the MLP is not yet tested.

4.3 Recognition tests

Recognition tests were made to compare the following systems at SNR levels from clean down to -10 dB SNR:

1. normal (full-band) hybrid ASR system
2. FC hybrid, with 4 sub-bands (16 combinations)
3. FC approximation (AFC)
4. one expert per sub-band, with equal weights

Tests used the first 100 examples in the Numbers95 test set. Results are summarised in Figures 3a and 3b below.

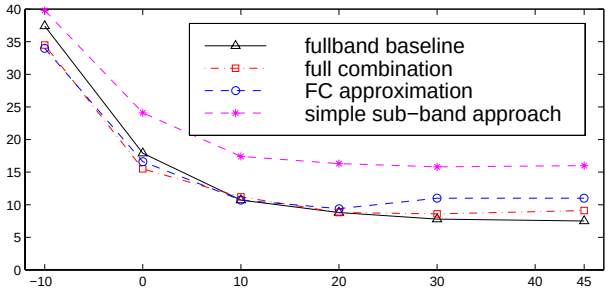

Figure 3a. WER (vertical axis) against SNR for full-band, FC, FC approximation and early sub-band ASR, with PLP data

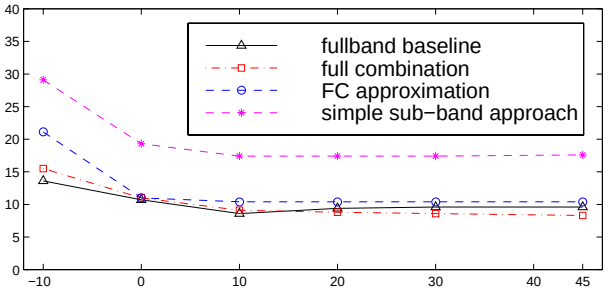

Figure 3b. WER (vertical axis) against SNR for full-band, FC, and FC approximation, with J-Rasta-PLP data

5. DISCUSSION

Our initial experiments were with car noise because in-car recognition is one of the areas we are particularly interested in. The SNR in a car interior does not usually fall much below 10 dB, and our results show that none of the systems tested here is very disturbed by this kind of noise. However, SNRs in a speeding car can fall to -10 dB, and here we see a significant drop in performance for the full-band baseline system using PLP data features.

For both PLP and J-Rasta-PLP features the simple sub-band approach used here shows reduced performance at all SNRs. With PLP features both the FC and AFC systems give near baseline performance down to 10 dB and begin to show some advantage for SNRs below this.

With J-Rasta-PLP features both FC and AFC systems give near baseline performance. J-Rasta-PLP features are like PLP features except that a filtering is applied to remove near stationary noise. This would explain why J-Rasta-PLP features here give increased robustness to noise (for all systems tested) without affecting performance with clean speech. This would also

explain why the FC approaches cannot further improve noise robustness in this case. We found that most of the advantage for the FC approach was also removed when a 300 Hz highpass filter was used. This is because a large part of the car noise lies below 300 Hz.

6. CONCLUSION

Our test results have shown that the full-combination method can overcome some of the problems associated with independent sub-band processing. However, with the simple procedures used here for expert weighting, the FC method does not yet show any advantage over the full-band baseline, except for SNRs less than 10 dB. In experiments reported elsewhere [8] we found that FC performance does not change significantly when all combination weights are simply equal. This suggests that the potential for reliability sensitive expert weighting has not yet been effectively exploited.

7. FUTURE DIRECTIONS

The full-combination approach is critically dependent on the method used for expert weight estimation.

A first step in improving weight estimation would be to improve sub-band reliability estimation. For this we are considering a number of approaches, including improved local SNR estimation, the detection of speech-specific characteristics such as harmonicity [2], direct use of data likelihood [16] and measures such as entropy derived from phoneme posteriors.

Another step would be to improve the method used to obtain the weight estimation in terms of these sub-band reliabilities. When fixed combination weights were trained on clean labelled data, with both LMSE and ML objectives, it was noted that the resulting combination weights tend to favour larger sub-band combinations. Better results were also obtained for FC with clean data by training separate weights for each phoneme.

The adaptive expert weight estimation used in Section 2.1 was oversimplified and could easily be improved in a number of ways. As with the fixed weighting, we could estimate separate weights for each phoneme. Then, instead of estimating $P(reliable_j)$ directly as a function of $\hat{SNR}_j$, we could measure r_0 and use $\hat{SNR}_j$ to measure $r(x_j)$. $P(r(x_j) > r_0)$ could then be estimated by modelling SNR_j as Gaussian with mean $\hat{SNR}_j$ and some given variance representing confidence in the SNR estimate.

Modelling of $P(reliable_j)$ could be further improved by measuring a separate r_0 for each sub-band, and also for each different number of other sub-bands which are reliable. This would more closely model rule 1.

ACKNOWLEDGEMENTS

This work was carried out in the framework of both the EC/OFES SPHEAR (Speech, Hearing and Recognition) project and the Fonds National Suisse MULTICHAN project (Non-stationary multichannel signal processing).

REFERENCES

[1] Allen, J. B. (1994) “How do humans process and recognise speech?”, IEEE Trans. on Speech and Signal Processing, Vol.2, No.4, pp.567-576.

[2] Berthommier, F. & Glotin, H. (1999) “SNR-feature mapping for robust multistream speech recognition”, Proc. ICPhS’99, (in press).

[3] Bourlard, H., Dupont, S. & Ris, C. (1996) “Multi-stream speech recognition”, Research Report IDIAP-RR-96-07.

[4] Bourlard, H. & Morgan, N. (1997) “Hybrid HMM/ ANN Systems for Speech Recognition: Overview and New Research Directions”, Proc. “International School on Neural Nets: Adaptive Processing of Temporal Information”

[5] Cole, R.A., Noel, T., Lander, L. & Durham, T. (1995) “New telephone speech corpora at CSLU”, Proc. European Conf. on Sp. Comm. and Tech., 1, pp. 821-824.

[6] Fletcher, H. (1922) “The nature of speech and its interpretation”, J. Franklin Inst., 193(6), pp.729-747.

[7] Hagen, A., Morris, A.C. & Bourlard, H. (1998) “Sub-band based speech recognition in noisy conditions: The Full-Combination approach”, Research Report IDIAP-RR 98-15.

[8] Hagen, A., Morris, A.C. & Bourlard, H. (1999) “Different weighting schemes in the full combination sub-bands approach for noise robust ASR”, Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions.

[9] Hermansky, H. (1990) “Perceptual linear predictive (PLP) analysis of speech”, J. Acoust. Soc. Am., 87(4), pp.1738-1752.

[10] Hermansky, H. & Morgan, N. (1994) “RASTA processing of speech”, IEEE Trans. on Speech and Audio Processing, 2(4), pp.578-589.

[11] Hermansky, H., Tibrewela, S. & Pavel, M. (1996) “Towards ASR on partially corrupted speech”, Proc ICSLP’96, pp. 462-465.

[12] Lippmann, R. P. & Carlson, B. A. (1997) “Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise”, Proc. Eurospeech’97, pp. 37-40

[13] Morgan, N., Bourlard, H. & Hermansky, H. (1998) “Automatic speech recognition: an auditory perspective”, Research Report IDIAP-RR 98-17.

[14] Morris, A. C., Cooke, M. & Green, P. (1998) “Some solutions to the missing feature problem in data classification, with application to noise robust ASR”, Proc. ICASSP'98, pp.737-740.

[15] Varga, A., Steeneken, H.J.M., Tomlinson, M. & Jones, D. (1992) “The Noisex-92 study on the effect of additive noise on automatic speech recognition”, Tech. Rep. DRA Speech Research Unit.

[16] de Veth, J., de Wet, F., Cranen, B. & Boves, L. (1999) “Missing feature theory in ASR: make sure you missing the right type of features”, Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions.