IDIAP RESEARCH REPORT

CONFUSION MATRIX BASED POSTERIOR PROBABILITIES CORRECTION

Andrew C. Morris & Hemant Misra

IDIAP-RR 02-53

December 2002

SUBMITTED FOR PUBLICATION

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
email secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP-RR 02-53

CONFUSION MATRIX BASED POSTERIOR PROBABILITIES CORRECTION

Andrew C. Morris, Hemant Misra

December 2002

Abstract

An MLP classifier outputs a posterior probability for each class. With noisy data classification becomes less certain and the entropy of the posteriors distribution tends to increase, therefore providing a measure of classification confidence. However, at high noise levels entropy can give a misleading indication of classification certainty because very noisy data vectors may be classified systematically into whichever classes happen to be most noise-like. When this happens the resulting confusion matrix shows a dense column for each noise-like class. In this article we show how this pattern of misclassification in the confusion matrix can be used to derive a linear correction to the MLP posteriors estimate. We test the ability of this correction to reduce the problem of misleading confidence estimates and to increase the performance of individual MLP classifiers. Word and frame level classification results are compared with baseline results for the Numbers95 database of free format telephone numbers, in different levels of added noise.

Keywords: MLP classifier, classifier correction, confidence measures, confusion matrix

Acknowledgements: This work was supported by the Swiss Federal Office for Education and Science (OFES) in the framework of both the EC/OFES HOARSE (Hearing Organisation And Recognition of Speech in Europe) project and the Swiss National Centre of Competence in Research (NCCR) on Interactive Multimodel Information Management (IM2).

1. Introduction

In the context of multi-stream HMM/ANN speech recognition [1,9,10], the logic behind using the entropy in the output vector from each MLP classifier to weight the evidence supplied by that MLP [3,4,5,7,8,11,13] is that uncertain outputs with a flat distribution will have high entropy, while a confident peaked distribution will have low entropy. However, it has been observed that while MLP output entropy does usually increase with noise level, at high noise levels an MLP may also output a high probability that the noisy data comes from just the one or two "attractor" classes which happen to be closest to the noisy input in feature space. In this case a low entropy high confidence measure can result for an MLP which is performing a highly inaccurate classification.

In this article we develop and test a correction which is designed to exploit the pattern of errors revealed by a validation set confusion matrix to correct posterior probability estimates such that false low entropy posterior distributions can be avoided.

In Section 2 we illustrate the way in which this problem with misleading entropy estimates typically arises during recognition as the noise level increases. In Section 3 we show how the posteriors correction for each MLP was derived from the cross-validation set confusion matrix for this MLP. Section 4 shows recognition test results for the Numbers95 database of free format telephone numbers, in different levels of added noise. Here it is shown that the proposed posteriors correction leads to improved entropy based confidence scores and significantly reduced frame- (though not word) error rates under various noise conditions. Section 6 follows with a discussion and conclusion.

2. Problem with posteriors entropy as a confidence measure

While classifier entropy generally increases with noise level, as the noise level increases beyond a certain point many frames start to be classified erroneously as the "attractor" phoneme which happens to be most noise-like (see Fig.1). As this classifier saturation occurs the entropy may stop increasing and start decreasing, so leading to a misleading indication of classification confidence (see Fig. 2). The resulting confusion matrix usually shows a dense column for each attractor phoneme. The idea underlying the posteriors correction proposed in this article is to use the pattern of misclassification observed in a noisy cross-validation set confusion matrix to derive a linear correction to the MLP posteriors which should redistribute attractor phoneme confusion throughout the confusion matrix, thereby increasing the utility of the posteriors entropy as a confidence measure.

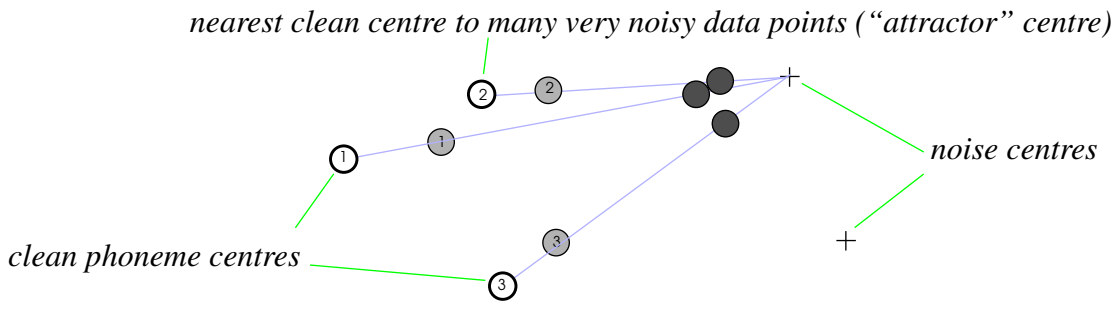

Figure 1. White points denote clean class centres. When a small amount of noise is added (light grey points) each centre shifts a small distance towards the noise centre. This results in some points being "randomly" misclassified. When a large amount of noise is added, many centres move close to the noise centre (+).. Most of these are classified not as noise, but as the original class which happens to be nearest to the noise centre.

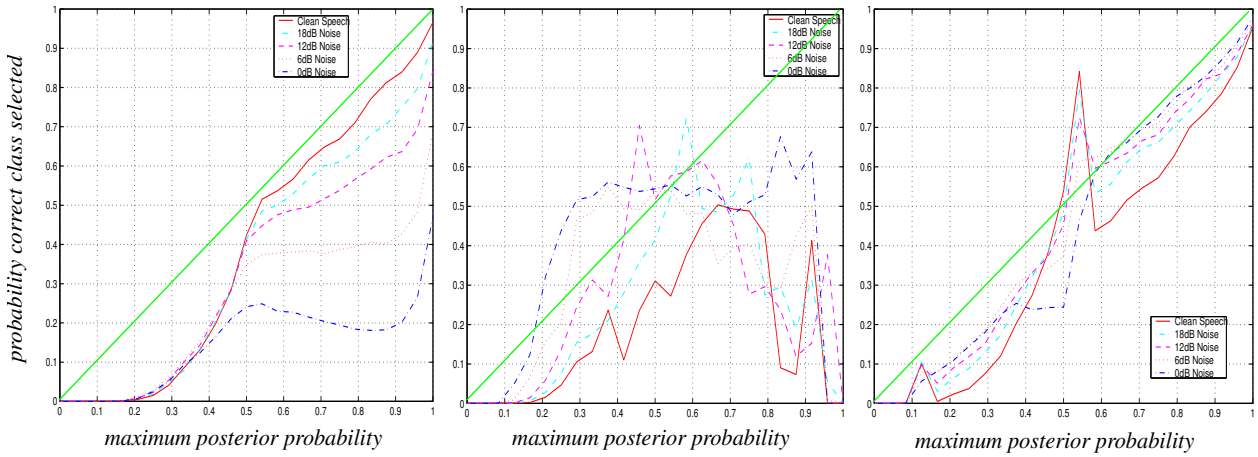

Figure 2. Maximum posterior size (horizontal) vs probability largest probability selects correct class (vertical), for all-streams MLP, for data conditions clean (solid), SNR 18 (dot-dashed), SNR 12 (dashed), SNR 6 (dotted) and 0 dB SNR (darker dot-dashed), before correction (left), after correction using Model-1 (centre) and after correction using Model-2 (right). Noise is Noisex factory.

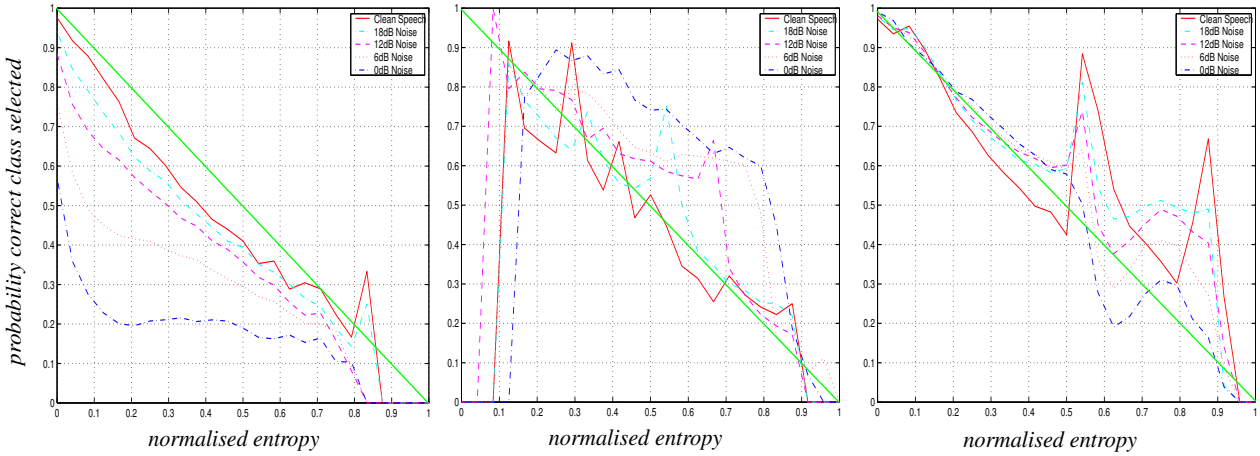

Figure 3. Normalised entropy size (horizontal) vs probability largest probability selects correct class (vertical), for all-streams MLP (line types as in Fig.2), before correction (left), after correction using Model-1 (centre) and after correction using Model-2 (right).

The original idea of this correction was just to correct the posteriors used in the entropy calculation for the purpose of classifier confidence estimation. It was not intended to apply this correction to the posteriors which are passed as scaled likelihoods to the decoder. However, in Section 4 we test both of these together. In future they should also be tested separately.

3. Confusion matrix based posteriors correction

For each trained MLP (see Section 4 for HMM/ANN system and test database used) we first obtain a confusion matrix $C(i, j)$ of $(true, guessed)$ co-occurrence counts on a multiple noise-condition validation set. This gives the number of names a frame labelled with "true" class (i) is given a maximum posterior probability of being from

"guessed" class (j). This is then converted into a matrix of conditional probabilities $P(T = i|G = j)$ by dividing each column by its sum.

$$P(T = i|G = j) = P(T = i \wedge G = j)/P(G = j)$$

$$P(G = j) = \sum_i P(T = i \wedge G = j)$$

3.1 Model 1

We can obtain corrected estimates for these true-class probabilities from the guessed-class MLP output probabilities by using this matrix of conditional probabilities as follows (where $(T = true)$ and $(G = guessed)$ class).

$$P(T = k|x) = \sum_j P(T = k \wedge G = j|x) \tag{1}$$

$$= \sum_j P(G = j|x)P(T = k|G = j \wedge x) \tag{2}$$

$$\cong \sum_j P(G = j|x)P(T = k|G = j) \tag{3}$$

A vector of corrected posterior probabilities $P_k{}' = P(T = k|x)$ can therefore be obtained from the vector of initial posterior probability estimates $P_j = P(G = j|x)$, using (3), as follows.

$$P' = C \cdot P \tag{4}$$

In the extreme case where all data is identified as belonging to the same noise-like class, the confusion matrix would be empty except for one column, which would be full. This correction would convert all such zero entropy posteriors vectors into the vector of class priors, with correspondingly high entropy.

Direct application of (4) leads, however, to a catastrophic failure to recognise the large proportion of frames which are originally labelled, correctly or otherwise, as a "space". This increases clean WER using only the single full-band MLP from 10.9% to 41.6%. Model-2 was therefore introduced to get over this problem.

3.2 Model 2

The main attractor class in our tests was the important "space" phoneme. The effect of the posteriors correction in (4) is to strongly flatten every posteriors vector for which the highest probability is for an attractor class. Flattening all "space" frame posteriors leads to a high rate of word deletion. We therefore developed another correction procedure in which different corrections are applied to frames estimated as speech or non-speech (Spe = "x is speech").

$$P(T = k|x) = P(T = k \wedge Spe|x) + P(T = k \wedge \neg Spe|x) \tag{5}$$

$$= P(Spe|x)P(T = k|Spe \wedge x) \tag{6}$$

$$+ P(\neg Spe|x)P(T = k|\neg Spe \wedge x) \tag{7}$$

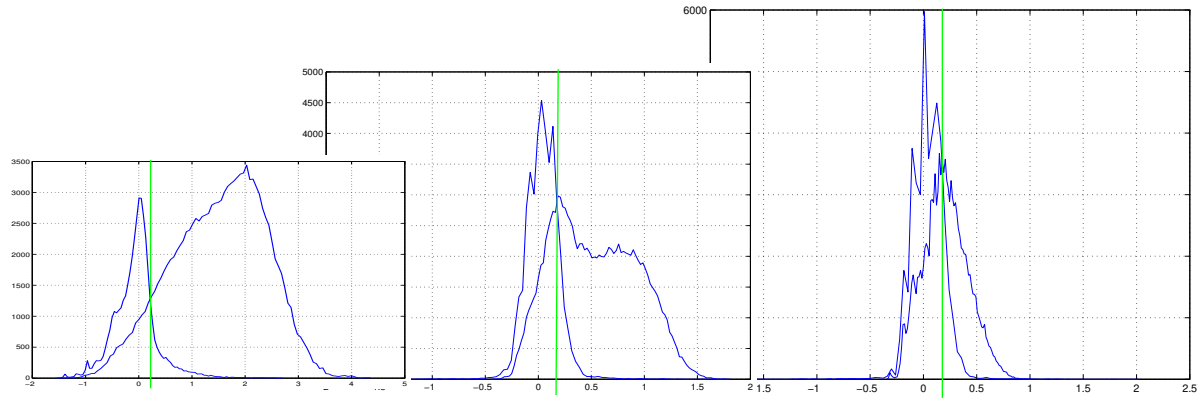

Figure 4. Plots show (unnormalised) histograms (c0-c0n) = log(E/nE) for speech frames superimposed on same for non-speech frames, for noise conditions clean (left), SNR 12 dB (centre) and SNR 0 dB (right), for the Numbers95 development set with Noisex factory nose. Crossing points are (respectively) at 0.213, 0.176 and 0.175 (giving E/nE = 1.237, 1.192, 1.191).

$P(T = k|Spe \wedge x)$ can now be expanded as $P(T = k|x)$ was expanded before. With $P(Spe|x) = \alpha$, this gives

$$P' = (\alpha C_{Spe} + (1 - \alpha) C_{\neg Spe}) P \tag{8}$$

where C_{Spe} is the validation confusion probabilities matrix estimated using only frames detected as speech, and $C_{\neg Spe}$ is estimated from all other frames. We used here a hard speech/non-speech decision based on the frame energy E level and estimated noise energy nE alone (one could use a more sophisticated speech/non-speech detector) as follows,

$$(Spe \Leftrightarrow E > \theta \cdot nE) \tag{9}$$

$$P(Spe|x) = 1 \text{ if } (E > \theta \cdot nE) \text{ else } P(Spe|x) = 0 \tag{10}$$

where E in each frame is estimated as e^{c0} ($c0$ is the unused PLP $c0$ coefficient), and nE is estimated as the average e^{c0} value over the first 10 frames in each utterance (see Fig. 4).

4. Recognition tests

Tests were made with the Numbers95 corpus of multi-speaker free format telephone numbers [2] (e.g. “one hundred fourty two”) recorded at 8 kHz, with artificially added noise from Noisex [14]. Speech features used were cepstral domain PLP [6], excluding the energy coefficient “c0”. Recognition was performed using the “all combinations multi-stream hybrid” model [9,10]. In this system a separate one hidden layer MLP is trained for each of the seven non-empty combinations of the three cepstral feature streams (PLP, delta PLP, delta delta PLP) (where delta PLP coefficients were over 9 windows, and delta delta over 9 windows). Each MLP was trained to map a 9-frame window of these data vectors onto a probability for each speech unit [12]. Speech units were the standard 27 monophones which come with this database, in which the before/after silence and inter-word space phonemes have been merged into one class. Each MLP had one hidden layer with 600 hidden units per input stream.

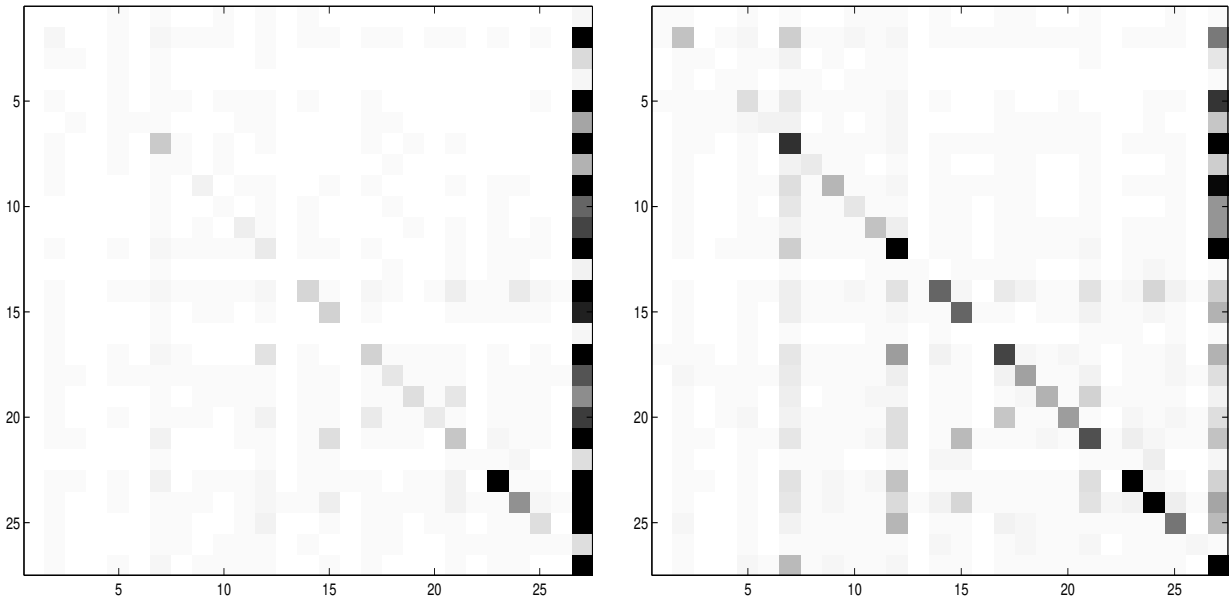

Figure 5. Confusion matrix for all-streams MLP, SNR 0 dB, baseline (top) & Model-2 corrected (bottom). Space class last. Correction gives improved frame-level recog. performance.

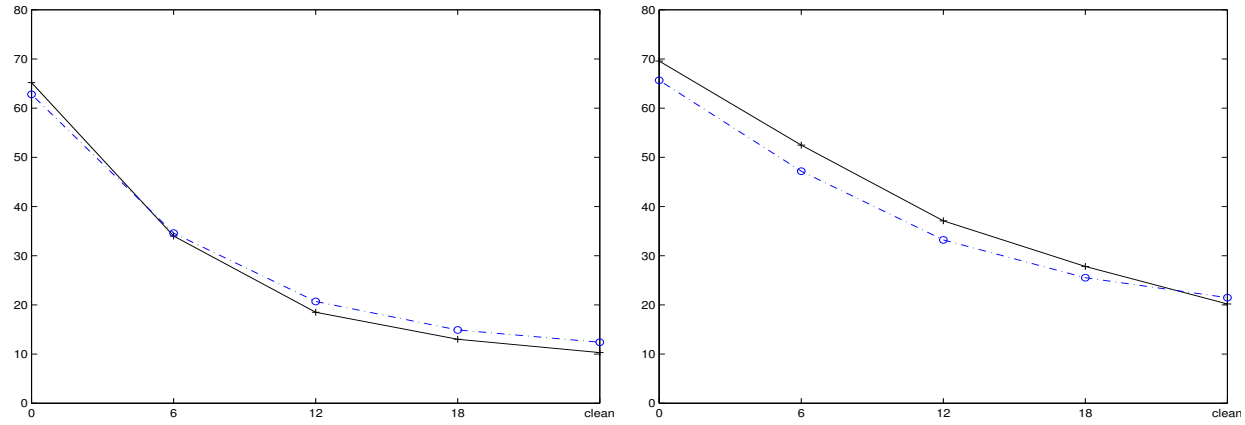

Figure 6. WER (left) and frame error rate (right) for noise levels clean to SNR 0 dB, factory noise, for entropy weighted multistream baseline (+, solid) and corrected (o, dot-dashed).

4.1 Entropy based posteriors combination rule

During recognition the output from all seven MLPs was combined in a linear confidence-weighted sum for each data frame, where the confidence weight w_n^i for the MLP for each stream combination $i = 1 \ldots S$ at frame (n) was given by the following IEWAT (inverse entropy with average entropy threshold) function of the posteriors entropy h_n^i from that MLP [8].

$$P(q_k|x_n, \Theta_k) = \sum_{i=1}^{S} w_n^i P(q_k|x_n^i, \theta_{i,k}) \tag{11}$$

$$\bar{h}_n = [\sum_{i=1}^{S} h_n^i]/S \tag{12}$$

$$\text{If } (h_n^i > \bar{h}_n)\ \overline{1/h_n^i} = 0\text{, else } \overline{1/h_n^i} = 1/h_n^i \tag{13}$$

$$w_n^i = \overline{1/h_n^i}/[\sum_{j=1}^{S} \overline{1/h_n^j}] \tag{14}$$

4.2 Tests made

Model-2 corrected posteriors (8, 10) were used for estimating both the entropies used in MLP weighting and in calculating the scaled posteriors used in decoding. The effect of this correction on posteriors classification power and on the relation between true confidence and posteriors entropy is shown in Figs. 2 and 3. Recognition tests were made with artificially added Noisex factory noise, under conditions (clean, SNR 18, SNR 12, SNR 6 and SNR 0 dB). Optimal values for θ in (10), obtained for the development set at noise levels (clean, SNR 12, SNR 0 dB), were (1.24, 1.19, 1.19). Tests at all noise levels used $\theta = 1.19$. WER and FER (frame error rate) results are shown in Fig.6.

5. Discussion and conclusion

Figures 1, 2 and 5 illustrate that one of the root causes for misclassification at high noise levels in HMM/ANN ASR is the tendency for ANNs to confidently classify everything as one or other non-speech "attractor" phoneme, such as "space" or "silence". In Section 3 it was shown how the speech and non-speech confusion matrices for a moderately noisy cross validation data set can be used, together with speech/non-speech detection, to provide a linear correction to the posteriors vector output from an MLP classifier which takes into account the pattern of cross validation errors. Tests showed that, in multiple noise conditions, Model-2 posteriors correction consistently improved the accuracy of both posteriors entropy estimation and frame level phoneme recognition. WER, however, was not improved. The experiments reported in this aricle represent only an initial test of concept. Several potential improvements to this system can easily be identified.

Separate entropy and probabilities correction. As the "space" class is automatically identified as unreliable, every posteriors vector with a strong space probability becomes heavily flattenned, with the result that many spaces get deleted. This leads to a high word deletion rate, with the result that, while using corrected posteriors for *both entropy estimation and scaled likelihoods estimation*, the word error rate (WER) is increased. It would be instructive to test the separate effects of Model-2 correction for (i) corrected entropy only, and (ii) corrected scaled likelihoods only.

Improved speech/non-speech separation. The speech/non-speech detection method used here was very crude. Model 2 posteriors correction could therefore probably be improved by focusing on ways of combining more noise robust speech/non-speech detection techniques with ANN classification. Furthermore, ideal speech units should be more equidistant from non-speech, thereby avoiding attractor classes.

Noise-level based correction. The pattern of confusion changes significantly with noise level. As noise level is relatively easy to estimate (inital data frames are often representative of noise throughout each utterance) the correction could possibly be improved by training and selecting a different correction for each noise level.

Multicondition training. Multicondition training invariably leads to a large increase in noise robustness. One advantage of this confusion-correction procedure is that, being based on classifier confusion characteristics rather than SNR based weighting, it can be applied to experts trained with any kind of data.

References

[1] Bourlard, H. & Dupont, S. (1996) “A new ASR approach based on independent processing and recombination of partial frequency bands”, Proc. ICSLP’96, Philadelphia, pp. 422-425.

[2] 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.

[3] Deroo, D., “Modèles dépendents du contexte et méthodes de fusion de données appliqués a la reconnaissance de la parole par modèles hybrides HMM/MLP”, PhD thesis (Faculté Polytechnique de Mons), 1998.

[4] Dupont, S. & Luettin, J. (1998) “Using the multi-stream approach for continuous audio-visual speech recognition: experiments on the M2VTS database”, Proc. ICSLP’98, pp. 1283-1286.

[5] Heckmann, M., Berthommier, F. & Kroschel, K. (2002) “Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition”, to appear in Applied Signal Processing, Vol. 2, No. 11, special issue on Audio-Visual Processing.

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

[7] Kirchhoff, K. & Bilmes, G. (1999) “Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values”, Proc. ICASSP 1999.

[8] Misra, H., Bourlard, H. & Tyagi, V. (2002) “Entropy-Based Multi-Stream Combination”, IDIAP-RR 02-31, 2002.

[9] Morris, A. C., Hagen, A. & Bourlard, H. (1999) “The full-combination sub-bands approach to noise robust HMM/ANN based ASR”, Proc. Eurospeech’99, pp. 599-602.

[10] Morris, A.C., Hagen, A., Glotin, H. & Bourlard, H. (2001) “Multi-stream adaptive evidence combination for noise robust ASR”, Speech Communication, Vol.34, pp.25-40.

[11] Okawa, S., Bocchieri, E. & Potamiamos, A. (1999) “Multi-band speech recognition in noisy environments”, Proc. Eurospeech’99.

[12] Richard, M. D. & Lippmann, R. P. (1991) “Neural network classifiers estimate Bayesian a-posteriori probabilities”, J. Neural Computation, 3(4) MIT Press, pp.461-483.

[13] Tomlinson, J., Russel, M. J. & Brooke, N. M. (1996) “Integrating audio and visual information to provide highly robust speech recognition”, Proc. ICASSP’96, pp. 821-824.

[14] 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.