Confusion Matrix Based Entropy Correction in Multi-stream Combination

Hemant Misra, Andrew Morris

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
Martigny, Switzerland
{misra,morris}@idiap.ch

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 providing a measure of classification confidence. However, at high noise levels, entropy can give a misleading indication of classification certainty. Very noisy data vectors may be classified systematically into classes which happen to be most noise-like and 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 enhance the performance of entropy based full-combination multi-stream approach. Better word-error-rates are achieved for Numbers95 database at different levels of added noise. The correction performs significantly better at high SNRs.

1. Introduction

In the context of multi-stream Hidden Markov Model (HMM)/ Artificial Neural Network (ANN) speech recognition [1, 2, 3], the logic behind using the entropy in the output vector from each multi-layered perceptron (MLP) classifier to weight the evidence supplied by that MLP [4, 5, 6, 7, 8, 9, 10] is the following: 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 cross-validation (CV) set confusion matrix. The correction is applied to modify posterior probability estimates such that the 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, for multi-stream combination approach, we show how the posteriors correction for each MLP was derived from the CV set confusion matrix for that MLP. Section 4 shows recognition test results for *Numbers95* [11] database of free format telephone numbers, with noise added at different signal-to-noise-ratios (SNRs). Here it is shown that the proposed posteriors correction leads to improved entropy based confidence scores and reduced word error rates (WERs) under various noise conditions. In addition, the performance is shown to be significantly better at very high SNRs, the typical condition for which the approach is proposed. Section 5 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 getting classified erroneously as the "attractor" phoneme which happens to be most noise like (Fig 1). As this classifier

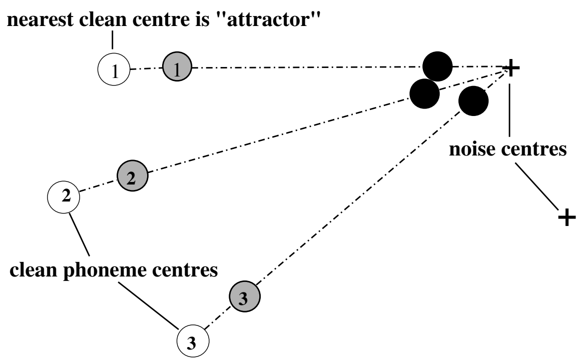

Figure 1: *As a small amount of noise is added, each class centre shifts towards the noise centre (white points move to light grey points) and some points may be randomly misclassified. However, when a large amount of noise is added, many centres move very close to the noise centre (+) and are classified (by forced choice) according to the phoneme class centre which is nearest.*

saturation occurs, the entropy may stop increasing and instead start decreasing, giving a misleading indication of classification confidence (Figs 2 and 3). The resulting confusion matrix usually shows a dense column for each attractor phoneme (Fig 4). The idea underlying the posteriors correction proposed in this article is to use the pattern misclassification observed in a noisy CV set confusion matrix to derive a linear correction to the MLP posteriors. The correction should redistribute attractor phoneme confusion throughout the confusion matrix, thereby increasing the utility of the posteriors entropy as a confidence measure.

The idea of this correction is only to correct the posteriors used in the entropy calculation for the purpose of classifier confidence estimate. It is not intended to apply this correction to the posteriors which are passed as scaled likelihoods to the decoder (the reason for not doing this is clear from Fig 2 where we see that the posteriors correction doesn't give a linear relationship between *maximum posterior probability* and *probability correct class selected*). If we analyse this result closely, we realize that

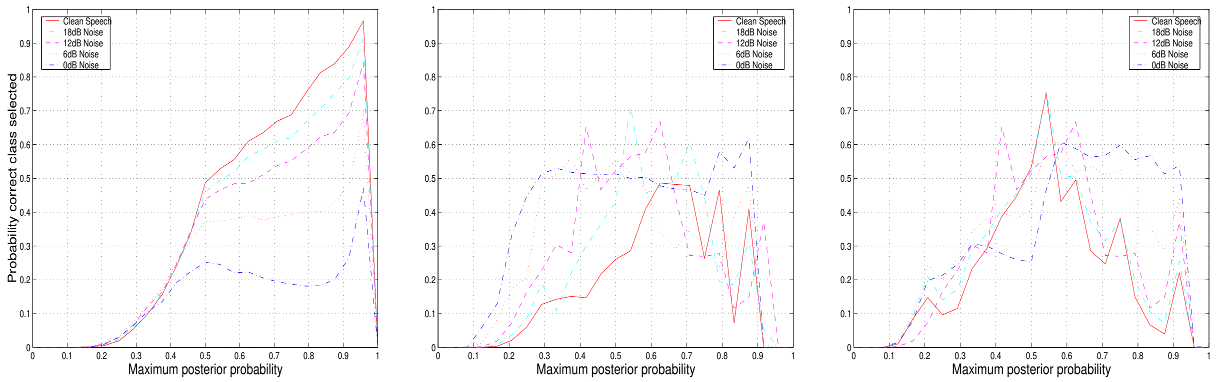

Figure 2: *Maximum posterior (horizontal) Vs probability that the largest probability selects the correct class (vertical). The plot is for all-stream MLP for the following data conditions: Clean (–), SNR18 (-.), SNR12 (- -), SNR6 (..), SNR0 (dark -.). Before correction (left), Model-1 correction (midle), and Model-2 correction (right). Noise is factory noise from Noisex database.*

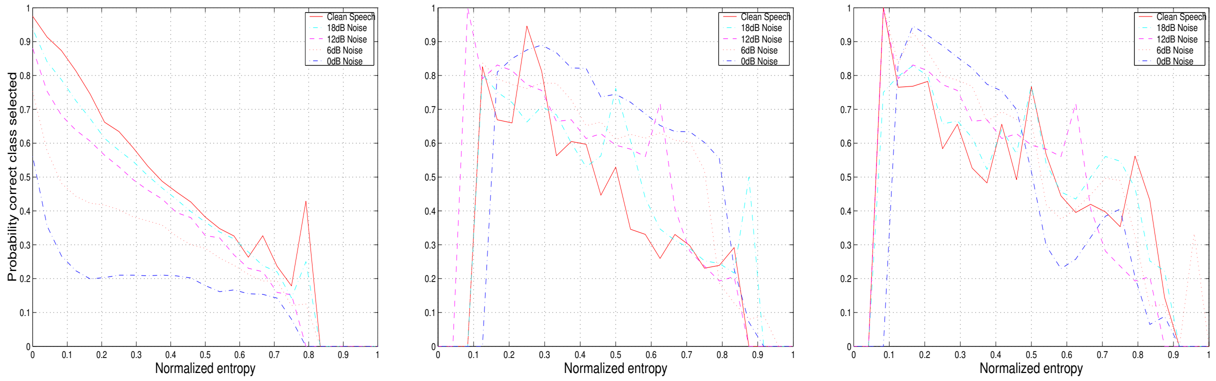

Figure 3: *Normalised entropy (horizontal) Vs probability that the largest probability selects the correct class (vertical). The plot is for all-stream MLP. Line types as in Fig 2. Before correction (left), Model-1 correction (middle) and, Model-2 correction (right). Noise is factory noise from Noisex database.*

indeed it is not possible to correct the posteriors otherwise we could have solved the problem of ASR by applying this correction repeatedly till we obtained the desired accuracy.

3. Confusion matrix based posteriors correction

For each trained MLP (see Section 4 for entropy weighted multi-stream 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 CV set. This gives the number of times 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.

$$C(i,j) = P(T=i|G=j) = \frac{P(T=i \wedge G=j)}{P(G=j)} \qquad (1)$$

$$P(G=j) = \sum_i P(T=i \wedge G=j) \qquad (2)$$

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).

$$P(T=k|x) = \sum_j P(T=k \wedge G=j|x) \qquad (3)$$

$$= \sum_j P(G=j|x)P(T=k|G=j \wedge x) \qquad (4)$$

$$\cong \sum_j P(G=j|x)P(T=k|G=j) \qquad (5)$$

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 (5), as follows,

$$P^{'} = C \cdot P \qquad (6)$$

In the extreme case where all data is identified as belonging to the same noise-like class, the confusion matrix would be empty

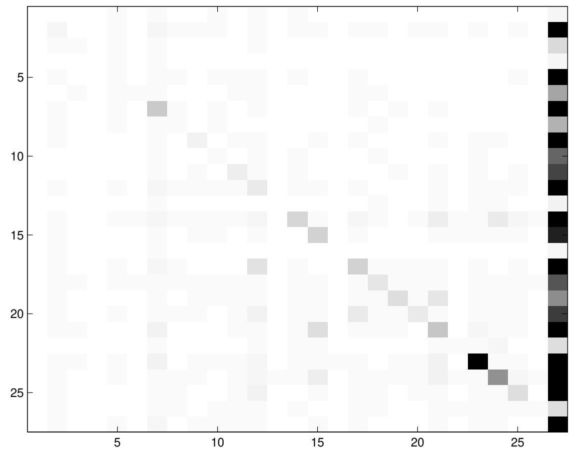

Figure 4: *Confusion matrix for all-streams MLP classifier for 0dB SNR (factory noise from Noisex database added to Numbers95 database). Space (silence) class is the last column.*

except for one column belonging to that noise-like phoneme. The column of this noise-like phoneme would be full (Fig 4). The proposed correction (6) would convert all such zero entropy posteriors vectors into the vector of class priors, with correspondingly high entropy.

3.2. Model 2

The main attractor class in our tests was the important "space" phoneme. The effect of the posteriors correction in (6) is to strongly flatten every posteriors vectors for which the highest probability is for an attractor class. Flattening all "space" frame posterior vectors leads to a high word deletion count. We therefore developed an amended correction procedure in which different corrections are applied to frames estimated as speech or non-speech (*Spe* = "x is speech").

$$\begin{aligned} P(T=k|x) =& P(T=k \wedge Spe|x) \\ &+ P(T=k \wedge \neg Spe|x) \quad (7) \\ =& P(Spe|x)P(T=k|Spe \wedge x) \\ &+ P(\neg Spe|x)P(T=k|\neg Spe \wedge x) \quad (8) \end{aligned}$$

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

$$P' = (\alpha C_{Spe} + (1-\alpha)C_{\neg Spe})P \quad (9)$$

where C_{Spe} is the CV 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 level (E) and estimated noise energy (E_{noise}) as follows,

$$(Spe \Leftrightarrow E > \theta_{threshold} + E_{noise}) \quad (10)$$

$$P(Spe|x) = \begin{cases} 1 & : \quad E > \theta_{threshold} + E_{noise} \\ 0 & : \quad E \leq \theta_{threshold} + E_{noise} \end{cases} \quad (11)$$

where E in each frame is estimated as e^{c0} (c0 is the unused PLP c0 coefficient), and E_{noise} is estimated as the average e^{c0} value over the first 8 frames in each utterance. One could use a more sophisticated speech/non-speech detector for better estimates.

4. Recognition tests

Tests were made with the *Numbers95* [11] corpus of multi-speaker free format telephone number (e.g. "one hundred forty two") recorded at 8 KHz. Noise was added artificially from Noisex [12]. Speech features used were cepstral domain PLP [13], excluding the energy coefficient "c0". Recognition was performed using the "full combinations multi-stream hybrid" [2, 3] model. 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 and delta-delta PLP). Both delta and delta-delta PLP coefficients were over 9 windows. Each MLP was trained to map a 9-frame context window of these data vectors onto a probability for each speech unit [14]. Speech units were the standard 27 monophones which come with this database. The before/after silence and inter-word space phonemes have been merged into one class. Each MLP had one hidden layer with number of hidden units proportional to the dimension of the input vectors fed to that MLP.

4.1. Entropy based posteriors combination rule

During recognition, the outputs from all the seven MLPs were combined in a linear confidence-weighted sum for each data frame. The confidence weight w_n^i for the MLP for each stream combination $i = 1...S$ 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].

$$P(q_k|x_n,\Theta) = \sum_{i=1}^{S} w_n^i P(q_k|x_n^i,\theta_i) \quad (12)$$

where θ_i are the parameters of the i^{th} MLP classifier and q_k is the k^{th} output class. Also, $h_n^i = -\sum_{k=1}^{K} P(q_k|x_n^i.\theta_i) \cdot \log_2 P(q_k|x_n^i.\theta_i)$ and K = 27 (number of monophones).

$$\bar{h}_n = \frac{\sum_{i=1}^{S} h_n^i}{S} \quad (13)$$

$$\overline{h_n^i} = \begin{cases} 10000 & : \quad h_n^i > \bar{h}_n \\ h_n^i & : \quad h_n^i \leq \bar{h}_n \end{cases} \quad (14)$$

$$w_n^i = \frac{1/\overline{h_n^i}}{\sum_{i=1}^{S} 1/\overline{h_n^i}} \quad (15)$$

4.2. Tests made

Model-1 and Model-2 corrected posteriors were used for estimating the entropies used in MLP weighting. The effect of these corrections on posteriors classification power and on the relation between true confidence and entropy is shown in Figs 2 and 3, respectively. Recognition tests were made with artificially added Noisex factory noise under different SNRs (clean, 18, 12, 6 and 0 dB). Optimal value for $\theta_{threshold}$ in (10) obtained for the development set at noise levels clean, 18 and 12 dB were 0.2106, 0.1472 and 0.1386, respectively. We used $\theta_{threshold}$ = 0.1758 for tests at all noise level. WER results are shown in Fig. 5

5. Discussion and conclusion

Figs. 1, 2 and 4 illustrate that one of the root causes for misclassification at high noise levels in HMM/ANN ASR is the tendency of the ANNs to confidently classify everything as one or

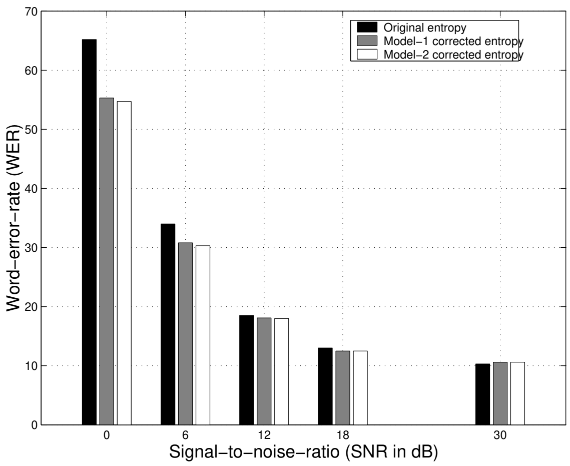

Figure 5: *WER for noise levels 'Clean (30dB)' to SNR 0dB for factory noise; Results are for multi-stream combination using IEWAT; Posteriors of the MLPs are used as scaled likelihood for decoding; Original, Model-1 corrected or Model-2 corrected entropy is used for weighting different streams.*

other non-speech "attractor" phoneme, such as "space" or "silence". In Section 3 we introduced two models for correcting this error. Model-1 is relatively simple and in Model-2 it was shown how the speech and non-speech confusion matrices for a moderately noisy CV data set can be used for correction. Together with speech/non-speech detection, Model-2 provided a linear correction to the posteriors vector output from an MLP classifier which takes into account the pattern of CV errors. Tests showed that in multiple noise conditions, Model-1 and Model-2 posteriors correction consistently improved the accuracy of both posteriors entropy estimation and WER.

Model-2 gives very slight improvement over Model-1 and the reason could be as follows: While speech/non-speech detection was crucial to the advancement from Model-1 to Model-2 posteriors correction, the speech/non-speech detection method used here was very crude. Problems with Model-2 posteriors correction could therefore be possibly addressed in future by focusing on ways of combining more noise robust speech/non-speech detection techniques with ANN classification. It would also be of interest to design a set of speech units which are more equidistant from non-speech classes and thereby avoid attractor classes.

Acknowledgements The authors want to thank the Swiss National Science Foundation for supporting this work through the National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)", as well as DARPA for supporting through the EARS (Effective, Affordable, Reusable Speech-to-Text) project.

6. References

[1] Hervé Bourlard and Stéphane Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proceedings of International Conference on Spoken Language Processing*, Philadelphia, PA, USA, Oct. 1996, pp. 426–429.

[2] Andrew C. Morris, Astrid Hagen, and Hervé Bourlard, "The full combination sub-bands approach to noise robust HMM/ANN based ASR," in *Proceedings of European Conference on Speech Communication and Technology*, Budapest, Hungary, Sept. 1999, vol. 2, pp. 599–602.

[3] Andrew C. Morris, Astrid Hagen, Hervé Glotin, and Hervé Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Comm.*, vol. 34, pp. 25–40, 2001.

[4] D. Deroo, *Modeles dependents du contexte et methodes de fusion de donnees appliques a la reconnaissance de la parole par modeles hybrides HMM/MLP*, PhD thesis, Faculte Polytechnique de Mons, Belgium, 1998.

[5] Stéphane Dupont and Juergen Luettin, "Using the multi-stream approach for continuous audio-visual speech recognition: Experiments on the M2VTS database," in *Proceedings of International Conference on Spoken Language Processing*, 1998, pp. 1283–1286.

[6] Martin Heckmann, Frédéric Berthommier, and Kristian Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *To be published in Journal on Applied Signal Processing (special issue on Audio-Visual Processing, 2002)*, vol. 2, no. 11, 2002.

[7] Katrin Kirchhoff and Jeff A. Bilmes, "Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, Phoneix, Arizona, Mar. 1999, pp. 2395–2398.

[8] Hemant Misra, Hervé Bourlard, and Vivek Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, Hong Kong, Apr. 2003.

[9] Shigeki Okawa, Enrico Bocchieri, and Alexandros Potamianos, "Multi-band speech recognition in noisy environments," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, Seattle, Washington, May 1998, pp. 641–644.

[10] J. Tomlinson, M. J. Russel, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 1996, pp. 821–824.

[11] Richard Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at cslu," in *Proceedings of European Conference on Speech Communication and Technology*, 1995, vol. 1, pp. 821–824.

[12] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the affect of additive noise on automatic speech recognition," Technical report, DRA Speech Research Unit, Malvern, England, 1992.

[13] Hynek Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[14] M. D. Richard and Richard P. Lippmann, "Neural network classifiers estimate bayesian a-posteriori probabilities," *Neural Computation*, vol. 4, no. 3, pp. 461–483, 1991.