NEW ENTROPY BASED COMBINATION RULES IN HMM/ANN MULTI-STREAM ASR

Hemant Misra, Hervé Bourlard, Vivek Tyagi*

IDIAP, Martigny, Switzerland
{misra, bourlard, vivek}@idiap.ch

ABSTRACT

Classifier performance is often enhanced through combining multiple streams of information. In the context of multi-stream HMM/ANN systems in ASR, a confidence measure widely used in classifier combination is the entropy of the posteriors distribution output from each ANN, which generally increases as classification becomes less reliable. The rule most commonly used is to select the ANN with the minimum entropy. However, this is not necessarily the best way to use entropy in classifier combination. In this article, we test three new entropy based combination rules in a full-combination multi-stream HMM/ANN system for noise robust speech recognition. Best results were obtained by combining all the classifiers having entropy below average using a weighting proportional to their inverse entropy.

1. INTRODUCTION

Many variations of the multi-stream Hidden Markov Model (HMM)/Artificial Neural Network (ANN) based hybrid ASR system [1] have been proposed, whereby complementary data streams are combined to improve recognition performance. Multiple data streams may be from different sensory modalities, e.g. video and audio [2], or from different representations of the same input stream, such as analysis on different time scales [3], or static and time difference features as used in this paper. We are working with the full-combination multi-stream (FCMS) HMM/ANN approach for noise robust ASR, whose superiority was shown in [3]. A central issue in multi-stream combination is expert weighting. A widely used measure of classifier confidence is the entropy [4] of the output posteriors distribution. The combination rule most commonly used is to select the ANN with the minimum entropy. In this article, we compare the performance of this rule with several new entropy based combination rules. In the next section, we introduce the FCMS HMM/ANN model used in our experiments. In Section 3, we present the three new entropy based weighting rules tested in this paper. Sections 4 and 5 present the experimental details and discuss the results obtained. This is followed by a conclusion in Section 6.

* Also with EPFL, Lausanne, Switzerland

2. FULL-COMBINATION MULTI-STREAM

In an HMM/ANN based hybrid ASR system, the output of the ANN are estimates of posterior probabilities, $P(q_k|x_n, \theta)$, where q_k is the k^{th} output class, x_n is the acoustic feature vector for the n^{th} frame, and θ is the set of parameters of the ANN model.

In FCMS, one ANN expert is trained for each stream combination. In Fig. 1, we have 3 feature representations

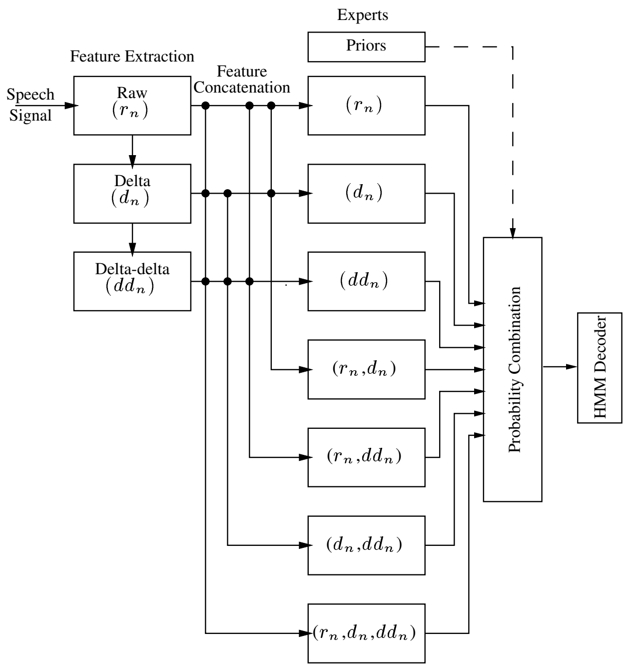

Fig. 1. *Multi-stream full-combination approach using Raw features (r_n), Delta features (d_n) and Delta-delta features (dd_n), and all their possible combinations, as separate streams in the frame work of an HMM/ANN hybrid system.*

giving 2^3 = 8 possible stream combinations. However, the 8^{th} combination is empty and is the a-priori probabilities in case none of the 7 experts were reliable [3].

The combined output posterior probability for the k^{th}

class and n^{th} frame is then computed according to:

$$\hat{P}(q_k|X_n,\Theta) = \sum_{i=1}^{I} w_n^i P(q_k|x_n^i,\theta_i) \qquad (1)$$

where I is the number of experts or streams (7 in the present case), $X_n = \{x_n^1, \cdots, x_n^I\}$, the set of all possible stream combinations built up from x_n, $\Theta = \{\theta_1, \cdots, \theta_I\}$, the set of parameters for each expert trained for each possible stream combination, and w_n^i is the weight assigned to the output of the i^{th} expert.

3. ENTROPY BASED COMBINATION

The entropy of the i^{th} expert for n^{th} frame, h_n^i, is computed by the equation,

$$h_n^i = -\sum_{k=1}^{K} P(q_k|x_n^i,\theta_i) \cdot log_2 P(q_k|x_n^i,\theta_i) \qquad (2)$$

where K is the number of output classes or phonemes (27 phonemes in our case), x_n^i is the input acoustic feature vector for the i^{th} expert for the n^{th} frame, and θ_i is the parameter set of the i^{th} ANN expert.

In our study, we observed that if an ANN has been trained on clean speech, the average entropy (averaged over all the frames) at the output of the ANN increases in case of noisy speech (Tables 1 and 2). The tables show that average entropy is high for low signal-to-noise-ratio (SNR) speech signals. In other words, for noisy speech, the discriminatory power of the ANN decreases and the posterior probabilities tend to become more uniform. This mismatch between the training and testing conditions is reflected through the entropy at the output of the ANN. We have used this information in our FCMS approach for weighting the outputs of different experts.

At the time of testing, the experts associated with the streams that are more corrupted by noise will face more mismatched conditions. Consequently, their output entropy will increase indicating the fact that the posterior probabilities are approaching towards *equal probabilities for all the classes*. The experts having high entropy have less discrimination, therefore output of such experts should be weighted less. Similarly, the experts having low entropy will have higher discrimination among classes and their output should be weighted more.

To achieve the above, the idea of inverse entropy weighting is investigated in this paper. The weight, w_n^i (1), assigned to the output of the i^{th} expert is given by,

$$w_n^i = \frac{1/h_n^i}{\sum_{i=1}^{I} 1/h_n^i} \qquad (3)$$

The scaled likelihoods are obtained by dividing the combined posterior probabilities (1) by a-priori probabilities of their respective phones, and sent through an HMM decoder to get the decoded output [1].

In the following, we discuss some variations of this inverse entropy method. The results of these methods are also presented in this paper.

3.1. Inverse entropy weighting with static threshold

In this variation, a fixed maximum threshold is chosen for the entropy (empirically optimized for clean speech and is 1.0 in our studies). If the entropy of a particular expert for a frame is more than the threshold, the output of that expert is penalized by a *static weight* of $\frac{1}{10000}$ (other values of static weight gave similar performance). For the same frame, the output of the experts with entropy lower than the threshold are still weighted inversely proportional to their respective entropies. The modified equations for *Inverse entropy weighting with static threshold* (IEWST) are:

$$\tilde{h}_n^i = \begin{cases} 10000 & : \quad h_n^i > 1.0 \\ h_n^i & : \quad h_n^i \leq 1.0 \end{cases} \qquad (4)$$

$$w_n^i = \frac{1/\tilde{h}_n^i}{\sum_{i=1}^{I} 1/\tilde{h}_n^i} \qquad (5)$$

3.2. Inverse entropy weighting with average entropy at each frame level as threshold

In this weighting scheme, the average entropy of all the streams for a frame is calculated by the equation,

$$\bar{h}_n = \frac{\sum_{i=1}^{I} h_n^i}{I} \qquad (6)$$

This average entropy is used as a dynamic threshold for the frame and output of all the experts having entropy greater than the threshold are weighted very less ($\frac{1}{10000}$), whereas output of the experts having entropy lower than the threshold are weighted inversely proportional to their respective entropies. The equations in case of *Inverse entropy weighting with average threshold* (IEWAT) are:

$$\hat{h}_n^i = \begin{cases} 10000 & : \quad h_n^i > \bar{h}_n \\ h_n^i & : \quad h_n^i \leq \bar{h}_n \end{cases} \qquad (7)$$

$$w_n^i = \frac{1/\hat{h}_n^i}{\sum_{i=1}^{I} 1/\hat{h}_n^i} \qquad (8)$$

3.3. Minimum Entropy Criterion

In this approach, for every frame the output from the expert that has the minimum entropy is chosen and used for decoding while the output of rest of the experts are ignored. The modified equations in this case are:

$$\hat{P}(q_k|X_n, \Theta) = P(q_k|x_n^j, \theta j) \tag{9}$$

with

$$j = \underset{i}{\operatorname{argmin}} \quad \{h_n^i\} \tag{10}$$

4. EXPERIMENTAL SETUP

In the experiments reported in this paper, Numbers95 database of US English connected digits telephone speech [5] is used. There are 30 words in the database represented by 27 phonemes. Training is performed on clean speech utterances and testing data, which is different from the training data, is corrupted by different kinds of noises. To simulate noisy conditions Noisex92 database [6] is used and the car and factory noises are added at different SNRs to Numbers95 database. We ran the experiments using Rasta-PLP [7] features.

The ANNs used were a single layer multi-layer perceptron (MLP) and the number of units in the hidden layer of an ANN expert were proportional to the dimension of the input feature vector stream fed to that ANN. The feature vectors used in our FCMS system (Fig. 1) were: 12 dimensional raw cepstral coefficients (0^{th} coefficient is not used) represented by r_n, 13 dimensional delta cepstral coefficients (d_n), and 13 dimensional delta-delta cepstral coefficients (dd_n). The input layer was fed by 9 consecutive data frames.

The HMM used for decoding had fixed state transition probabilities of 0.5. Each phoneme had a 1 state model for which emission likelihoods were supplied as scaled posteriors [1]. The minimum duration for each phoneme is modeled by forcing 1 to 3 repetitions of the same state for each phoneme. *Phone deletion penalty* parameter was empirically optimized for clean speech test database and then it was kept constant for all the experiments.

5. RESULTS AND DISCUSSION

WERs and average entropy values of the above experimental setup are presented in Tables 1 and 2 for car and factory noises, respectively. The performance of the proposed schemes is either better or comparable to standard full-band system under different noise conditions. In general, the performance in the presence of *factory noise* is poor as compared to car noise and in most of the cases inverse entropy weighting with average threshold (IEWAT) performs the best. The relative average improvement in performance by different methods over the baseline full-band system are: *1.7% by Inverse entropy weighting, 8.5% by IEWST, 10.5% by IEWAT and 6.7% by Minimum entropy criterion*. Similar results, though not reported in this paper, are obtained for PLP [8] features where relative average improvement in performance is more significant as compared to Rasta-PLP features, but the absolute performance is relatively poorer.

Apart from WERs, the average entropy values also reveal a few important things. Average entropy at the output of each MLP expert (results not shown in the table), as well as for each combination, is high for low SNR input speech and low for high SNR input.

5.1. Relation between WER and Entropy

In the general framework of developing new speech recognition approaches targeting at consistently minimizing conditional entropy while introducing new knowledge sources, some interesting relationship is observed between WER performance of different combination methods and their respective entropies.

Entropy for any linear combination is always higher than the lowest entropy among all the combined experts and the same is observed from the entropy results of inverse entropy weighting. In this weighting, the entropies for the combination are high and at the same time improvement in WER performance is less significant. As expected, the minimum entropy criterion gives the least average entropy values. This is a situation where only the stream having the lowest entropy is chosen at every frame level and the other streams don't contribute to the decision. But the results indicate that even this highly constrained situation gives an improvement in the WER performance as well as a decrease in average entropy. Out of the other two non-linear combinations, average entropies and WERs for IEWST are always higher as compared to IEWAT. WER performances of IEWAT is best in most of the cases and also the entropy of the combination is always the lowest.

6. CONCLUSION

As with any multi-stream combination technique, the entropy based weighting schemes tested here with noise robust RASTA-PLP features give a much less dramatic performance improvement than they do with PLP features. However, the IEWAT (inverse entropy with average entropy threshold) weighting scheme, in which all experts with below average entropy are dynamically selected at each frame, outperforms all of the other schemes under almost all noise conditions. IEWAT gives a relative WER improvement, averaged over both the noise cases and all their SNRs, of 10.5% compared to the full-band baseline and 4.3% compared to minimum entropy selection. We observe that although the WER tends to decrease as the combined posteriors entropy decreases, selecting only the MLP with the minimum entropy does not usually give the best performance. The value of combining the posteriors from several experts

Stream	Car Noise (in db)				Clean
	0	6	12	18	Speech
r-d-dd (Baseline)	13.3 (0.94)	11.4 (0.87)	10.8 (0.82)	10.4 (0.78)	10.2 (0.74)
Equal Weights	13.9 (1.50)	12.1 (1.14)	12.2 (1.36)	12.4 (1.31)	12.0 (1.24)
Minimum Entropy	11.7 (0.60)	10.8 (0.55)	9.5 (0.52)	9.1 (0.49)	9.1 (0.47)
Inverse Entropy	12.0 (1.22)	10.8 (1.14)	11.0 (1.07)	11.0 (1.03)	10.6 (0.97)
Inv. Entr. Static Threshold	11.2 (0.94)	10.0 (0.86)	9.1 (0.82)	9.1 (0.78)	10.1 (0.75)
Inv. Entr. Avg. Threshold	11.0 (0.89)	9.2 (0.83)	9.2 (0.78)	9.1 (0.75)	10.2 (0.71)

Table 1. *Word-Error-Rates (and Average Entropy Values) for Car Noise. The baseline full-band system is r-d-dd. r - cepstral, d - delta cepstral and dd - delta-delta cepstral features.*

Stream	Factory Noise (in db)			
	0	6	12	18
r-d-dd (Baseline)	56.2 (1.42)	33.1 (1.33)	18.9 (1.09)	12.7 (0.90)
Equal Weight	55.8 (1.88)	32.6 (1.82)	18.7 (1.59)	14.2 (1.41)
Minimum Entropy	55.7 (0.96)	32.1 (0.89)	17.7 (0.72)	12.5 (0.60)
Inverse Entropy	54.5 (1.67)	31.9 (1.60)	18.5 (1.35)	13.0 (1.16)
Inv. Entr. Static Threshold	55.2 (1.45)	31.8 (1.36)	18.1 (1.11)	12.5 (0.92)
Inv. Entr. Avg. Threshold	54.7 (1.30)	31.5 (1.23)	17.2 (1.02)	12.2 (0.86)

Table 2. *Word-Error-Rates (and Average Entropy Values) for Factory Noise.*

appears to outweigh the advantage of pure entropy minimization. It remains to be seen whether IEWAT weighting will also improve the performance of audio-visual and other multi-stream combination applications which have up to now used minimum entropy selection.

7. ACKNOWLEDGMENTS

We wish to thank Andrew C. Morris for his useful suggestions. The authors want to thank the Swiss National Science Foundation for supporting this work through the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)", as well as DARPA through the EARS (Effective, Affordable, Reusable Speech-to-Text) project.

8. REFERENCES

[1] Nelson Morgan and Hervé Bourlard, "An introduction to the hybrid HMM/connectionist approach," *IEEE Signal Processing Magazine*, pp. 25–42, May 1995.

[2] 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.

[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] C. E. Shannon, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, 1949.

[5] 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.

[6] 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.

[7] Hynek Hermansky and Nelson Morgan, "Rasta processing of speech," *IEEE Trans. Speech, Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

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