IDIAP RESEARCH REPORT

Entropy-based multi-stream combination

Hemant Misra [a] Hervé Bourlard [a b]

Vivek Tyagi [a]

IDIAP–RR 02-31

PUBLISHED IN

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), April 2004, Hong Kong

[a] IDIAP - Dalle Molle Institute of Perceptual Artificial Intelligence, P.O. Box 592, CH-1920 Martigny, Switzerland

[b] EPFL, Lausanne, Switzerland

IDIAP Research Institute **www.idiap.ch**

Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 02-31

Entropy-based multi-stream combination

Hemant Misra   Hervé Bourlard   Vivek Tyagi

PUBLISHED IN

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), April 2004, Hong Kong

Abstract. Full-combination multi-band approach has been proposed in the literature and performs well for band-limited noise. But the approach fails to deliver in case of wide-band noise. To overcome this, multi-stream approaches are proposed in literature with varying degree of success. Based on our observation that for a classifier trained on clean speech, the entropy at the output of the classifier increases in presence of noise at its input, we used entropy as a measure of confidence to give weightage to a classifier output. In this paper, we propose a new entropy based combination strategy for full-combination multi-stream approach. In this entropy based approach, a particular stream is weighted inversely proportional to the output entropy of its specific classifier. A few variations of this basic approach are also suggested. It is observed that the word-error-rate (WER) achieved by the proposed combination methods is better for different types of noises and for their different signal-to-noise-ratios (SNRs). Some interesting relationship is observed between the WER performances of different combination methods and their respective entropies.

1 Introduction

Multi-band approach has been discussed in the literature for its superior performance for band-limited noise [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. Recently, many variations of multi-band approach have been proposed. The most promising ones include full-combination approach [6, 7, 8]. The superiority of full combination approach for different kinds of band-limited noise was shown in [10, 11]. In [11], it was reported that in case of wide-band noise the performance could not be improved further as compared to a full-band system.

To overcome this problem of wide-band noise, multi-stream approaches are proposed in the literature with varying degree of success. Full-combination multi-stream is one such approach [12, 11]. In general, the streams in a multi-stream system can be from different sources, let us say, video and audio. In some other cases, the streams could be different representations obtained from the same common source. In the present work, the streams are the different feature vectors obtained from the same common source, that is, full-band speech, and all the possible combinations of these feature vectors. The three feature vector representation we have considered in our experiments are raw cepstral coefficients, delta cepstral coefficients obtained from raw cepstral coefficients and delta-delta cepstral coefficients obtained from delta-cepstral coefficients. Though delta and delta-delta features are obtained from raw features, it is assumed that they include dynamic information incorporated in them and this information is complementary information. As in the full-combination approach, all the possible combinations of these three feature vectors are considered in our full-combination multi-stream approach. The novelty of our approach is the method used to combine these streams to make the system more robust to different types of noises.

In multi-stream paradigm, evidences from a number of representations of the full-band speech signal are combined to achieve robustness. It is assumed that different representations of the signal carry complementary information and thus can compensate for errors that are not common among the different representations. In this paper we propose a new entropy based method for combining different streams. We have investigated the method for full-combination multi-stream approach. In the next section we introduce our entropy based weighting approach in the frame work of full-combination multi-stream approach. In Section 3 the database used and the experimental setup has been explained. The results and conclusions are presented in Section 4 and 5, respectively.

2 Entropy based full-combination multi-stream

Entropy plays a central role in information theory as a measure of information, choice and uncertainty [13]. In case of several classes, the uncertainty as well as entropy is maximum when all the classes have equal probabilities. In case of Hidden Markov Model (HMM)/Artificial Neural Network (ANN) based hybrid automatic speech recognition (ASR) system [14, 15], 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 (each class corresponds to a particular phoneme or an HMM state) of the ANN, x_n is the acoustic feature vector for the n^{th} frame and θ is the set of parameters of the ANN model. Instantaneous entropy, h_n, at the output of such an ANN is computed by the equation,

$$h_n = -\sum_{k=1}^{K} P(q_k|x_n, \theta) \cdot log_2 P(q_k|x_n, \theta) \tag{1}$$

where, K is the total number of output classes (or phonemes in our case).

In one previous study [16] it has been observed that if ANN has been trained for speech signal, it gives low entropy for speech signals but gives high entropy for music signals. This result has been successfully used in speech/music discrimination in [16]. In our study, we observed that if the ANN has been trained on clean speech, the average entropy (averaged over all the frames of a particular stream) at the output of the ANN increases in case of noisy speech (Tables 3 and 5). From the tables it can be concluded that average entropy is high for speech signals having low SNRs, and lower the SNR of the speech signal is, higher is the entropy at the output of the ANN experts. In other words, for noisy speech, the posterior probabilities tend to become more uniform and the discriminatory power of the ANN decreases. It can be stated that if there is mismatch between the training and testing conditions, this will be reflected through the entropy at the output of the ANN. We have used this information in our full-combination multi-stream approach for weighting different streams.

In our full-combination multi-stream approach (Fig. 1), three different feature representations were used. As mentioned before, the three feature vector representation were raw cepstral coefficients, delta cepstral coefficients and delta-delta cepstral coefficients. These three representations as well as all possible combinations of these three representations were treated as individual streams. One ANN expert was trained for every stream. The total number of experts trained in our case were 7 (the 8^{th} combination being the prior probabilities in case none of the 7 experts are reliable).

As mentioned earlier, in HMM/ANN hybrid systems, output of the ANN expert trained on the i^{th} stream is the estimate of the posterior probabilities $P(q_k|x_n^i, \theta_i)$ for each of the 27 phonemes q_k and n^{th} data frame for the i^{th} stream x_n^i. θ_i is set of ANN parameters for i^{th} expert classifying x_n^i in terms of q_k classes, $k = 1, \cdots, K$. It is assumed that the streams that have higher entropy (posterior probabilities of different phonetic classes being similar) have less discriminatory information than those streams which have less entropy. The other interpretation is, the output of the ANN expert has higher entropy when there is more mismatch between training and testing conditions. Therefore at the time of testing the streams that are more corrupted by noise, their expert will face more mismatched conditions. Consequently, the entropy at the output of such experts will be high. High entropy implies that the posterior probabilities are approaching towards *equal probabilities for all the classes.* The experts having

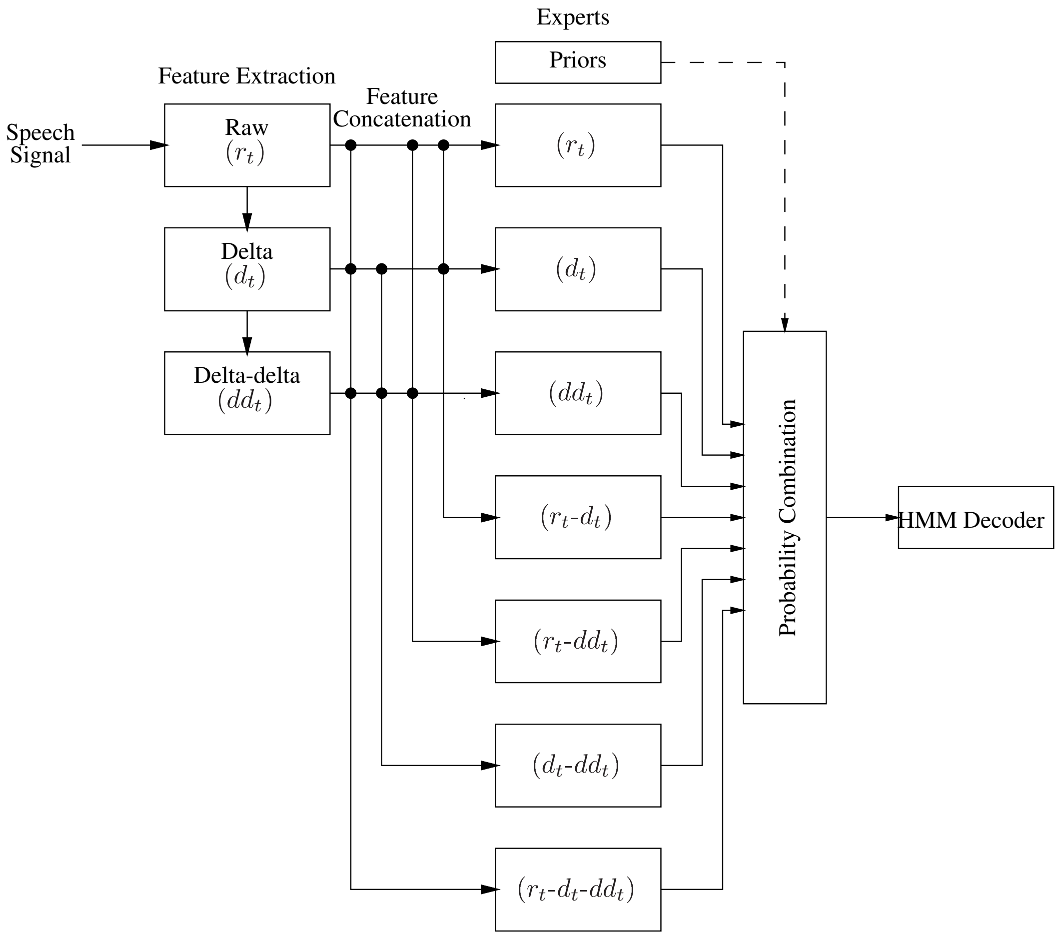

Figure 1: Multi-stream full-combination approach using Raw features ($R = r_1, \cdots, r_t, \cdots, r_T$), Delta features ($D = d_1, \cdots, d_t, \cdots, d_T$) and Delta-delta features ($Dd = dd_1, \cdots, dd_t, \cdots, dd_T$) as individual streams as well as all possible combinations of the three features as separate streams in the frame work of a HMM/ANN hybrid system.

high entropy have less discrimination, therefore such experts should be weighted less. Similarly, the experts having low entropy will have higher discrimination among classes and should be weighted more.

To achieve the above, the idea of inverse entropy weighting is investigated in this paper. Entropy of i^{th} expert for n^{th} frame is computed by the equation

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

where K is the number of output classes or phonemes (27 in our case), x_n^i is the input acoustic feature vector for the i^{th} stream for the n^{th} frame and I is the number of experts or streams (7 in the present case).

The combined output posterior probability for k^{th} class and n^{th} frame is then computed according to:

$$\hat{P}(q_k|X_n, \Theta) = \sum_{i=1}^{I} w_n^i P(q_k|x_n^i, \theta_i) \qquad \text{for } k = 1, \cdots, K \tag{3}$$

where $X_n = \{x_n^1, \cdots, x_n^i, \cdots, x_n^I\}$, $\Theta = \{\theta_1, \cdots, \theta_i, \cdots, \theta_I\}$ and

$$w_n^i = \frac{1/h_n^i}{\sum_{i=1}^{I} 1/h_n^i} \tag{4}$$

In the above equations, I is number of streams. The term in the denominator of Eq 4 is for normalizing so that the sum of combined probabilities is equal to one ($\sum_{k=1}^{K} \hat{P}(q_k|X_n, \Theta) = 1$).

The obtained combined posterior probabilities are send through a decoder after being divided by a-prior probabilities of their respective phones to get the decoded output.

Some simple variations of this inverse entropy method were also tried, including

Inverse entropy weighting with static threshold: In this approach, a static threshold is chosen for the entropy (1.0 in our studies). If the entropy of a particular stream for a frame is more than 1.0, the weight assigned to that stream is heavily penalized by a *static weight* of $\frac{1}{10000}$[1]. The streams with low entropy (less than the threshold of 1.0) for the same frame are still weighted inversely proportional to their respective entropies. We called this approach as *Inverse entropy weighting with static threshold* (IEWST). In this case, the equations are modified as:

$$\tilde{h}_n^i = \begin{cases} 10000 & : \; h_n^i > 1.0 \\ h_n^i & : \; h_n^i \leq 1.0 \end{cases} \tag{5}$$

$$w_n^i = \frac{1/\tilde{h}_n^i}{\sum_{i=1}^{I} 1/\tilde{h}_n^i} \tag{6}$$

Minimum Entropy Criterion: In this variation, for a frame, the stream that has the minimum entropy is chosen and rest of the streams are ignored. This is done for all the frames in the test utterance. Therefore at each frame level the stream that has the minimum entropy is chosen and is used for decoding. The corresponding equations in this case are:

$$\hat{P}(q_k|X_n, \Theta) = P(q_k|x_n^j, \theta j) \tag{7}$$

such that

$$j = \underset{i}{\text{argmin}} \quad h_n^i \tag{8}$$

Inverse entropy weighting with average entropy at each frame level as threshold: In this weighting scheme, entropy of each stream at each frame level is computed. Then the average entropy of all the streams for that frame is calculated by the equation,

$$\bar{h}_n = \frac{\sum_{i=1}^{I} h_n^i}{I} \tag{9}$$

[1] Some other values of static weight were also tried and all of them gave similar performance

This average entropy is used as a threshold for that frame. For that frame, all the streams having entropy greater than the threshold are weighted very less ($\frac{1}{10000}$) and the streams having entropy lower than the threshold are weighted inversely proportional to their respective entropies. The equations in this case are:

$$\tilde{h}_n^i = \begin{cases} 10000 & : \quad h_n^i > \bar{h} \\ h_n^i & : \quad h_n^i \leq \bar{h} \end{cases} \tag{10}$$

$$w_n^i = \frac{1/\tilde{h}_n^i}{\sum_{i=1}^{I} 1/\tilde{h}_n^i} \tag{11}$$

The approach is referred to as *Inverse entropy weighting with average threshold* (IEWAT).

Ideally we would like the discrimination by the combination to be superior than any of the individual streams. Therefore, it will be interesting to know about the average entropy of the combined posterior probabilities ($\hat{P}(q_k|X_n, \Theta)$). It can be verified that any linear combination of the streams will always yield an entropy value between the highest and lowest entropy values among all the streams. Therefore, the simple inverse entropy weighting combination will always give an entropy value between these two limits. But in case of non-linear combinations investigated in this paper, some interesting results regarding the average entropy of the combination were obtained. These results are reported along with the WERs in Section 4. Some interesting parallels are drawn between WER and average entropy of the different combination methods.

3 Experimental setup

In the experiments reported in this paper, Numbers95 database of US English connected digits telephone speech [17] is used. There are 30 words in the database and there are 27 phonemes to represent the phonetic transcription of these 30 words. To simulate noisy conditions Noisex92 database [18] is used and the car, factory and lynx noises are added at different SNRs to Numbers95 database. Also, we used an in-house recorded car noise provided by our project partner Daimler Chrysler (reported as Noise50 in our paper). We ran the experiments using two types of features, PLP [19] and J-Rasta PLP [20]. The window size and window shift were 25.0 ms and 12.5 ms, respectively.

In our full-combination method it is the a-posterior phoneme probabilities that are required. An ANN/HMM hybrid system was used for all the experiments because it gives an estimate of a-posterior phoneme probabilities at the output of the ANN. The ANN used was a single layer Multi-layer Perceptron (MLP) with variable number of units in the hidden layer. The number of hidden units in an ANN expert were proportional to the dimension of the input feature vector stream fed to that expert (Table 1). The feature vectors used in our study were: raw cepstral coefficients (13 dimension but the 0^{th} coefficient is

Feature Stream	Dimension of feature stream	Number of hidden units in ANN expert
Raw (R)	12 * 9 = 108	600
Delta (D)	13 * 9 = 117	600
Ddelta (Dd)	13 * 9 = 117	600
R-D	25 * 9 = 225	1200
R-Dd	25 * 9 = 225	1200
D-Dd	26 * 9 = 234	1200
R-D-Dd (Baseline)	38 * 9 = 342	1800

Table 1: Number of hidden units in every ANN expert. In the experiments the 0^{th} cepstral coefficient has not been used. Raw is static cepstral coefficients, Delta is delta cepstral coefficients and Ddelta is delta-delta cepstral coefficients. **The baseline system is R-D-Dd**.

not used), delta cepstral coefficients (13 dimension) and delta-delta cepstral coefficients (13 dimension). The input layer was fed by 9 consecutive data frames. The output of the i^{th} ANN expert was the posterior probabilities $P(q_k|x_n^i, \theta_i)$ for each of the 27 phonemes q_k and n^{th} acoustic feature vector frame x_n^i.

At the time of recognition, combined scaled posterior probabilities (Eq 3) as well as scaled posterior probabilities from the MLP are passed to the HMM for decoding. The HMM used for decoding had fixed parameters with fixed state transition probabilities of 0.5. Each phoneme had a 1 state model for which emission likelihoods were supplied as scaled posterior from the MLP. The minimum duration for each phoneme is modeled by forcing 1 to 3 repetitions of the same state for each phoneme. A language model was used such that all word sequences were equal probable. It is similar to not having a language model for decoding. *Phone deletion penalty* parameter was empirically optimized for clean speech test database and then it was kept constant for all the experiments reported in this paper.

4 Results and discussion

WER results of the above experimental setup are presented in Tables 2 and 4 for PLP and J-Rasta PLP features, respectively. From the tables it can be seen that PLP features are more degraded by noise as compared to J-Rasta PLP features. Robustness of Rasta features towards noise is well established and the same observation is reported in [8].

Word-Error-Rates (WERs) for PLP Features																	
Stream	Car Noise (in db)				Factory Noise (in db)				Lynx Noise (in db)				Noise50 (in db)				Clean Speech
	0	6	12	18	0	6	12	18	0	6	12	18	0	6	12	18	
R	28.8	20.9	15.4	12.9	71.1	44.3	25.1	16.7	63.1	38.9	23.9	17.3	63.1	35.1	20.9	15.4	12.2
D	16.7	15.2	14.2	13.3	58.0	35.6	21.2	16.2	44.4	27.2	19.4	15.0	48.0	29.4	19.6	14.5	12.9
Dd	16.4	16.6	16.3	15.5	60.4	36.3	22.2	16.0	45.2	28.8	22.1	17.2	49.9	30.1	20.4	15.2	13.9
R-D	16.0	12.2	11.2	10.4	69.1	36.8	21.0	13.3	49.4	27.6	18.1	13.4	56.0	27.5	15.5	11.6	9.2
R-Dd	18.4	12.9	11.0	10.5	67.5	37.7	18.5	13.3	53.4	30.7	18.9	13.4	54.9	27.6	15.9	11.3	8.8
D-Dd	12.4	12.2	11.0	10.9	55.8	31.3	17.7	13.1	39.6	23.7	14.3	12.1	46.0	26.7	16.2	11.4	9.9
R-D-Dd (Baseline)	17.7	12.5	10.9	10.3	67.7	33.9	19.2	13.9	50.5	27.2	17.7	13.3	51.9	25.4	16.8	11.4	8.1
Inverse Entropy	11.2	10.6	10.5	10.1	63.3	31.1	19.1	11.6	43.6	24.2	16.3	11.3	47.8	25.1	15.1	11.2	9.2
IEWST	13.5	11.3	9.9	9.6	67.7	34.0	18.9	11.7	42.7	24.2	16.2	10.8	47.4	25.0	15.5	10.4	8.7
Minimum Entropy	13.8	11.0	10.5	10.1	69.3	34.8	18.3	12.0	43.9	25.5	15.8	11.3	49.4	25.5	15.4	10.5	8.9
IEWAT	13.7	10.5	9.7	9.5	67.9	32.5	17.2	11.8	41.5	23.4	15.4	10.4	46.3	24.2	14.7	10.6	8.9

Table 2: Word-Error-Rates for PLP features. Variable number of hidden units for each network (Table 1). The baseline full-band system is *R-D-Dd*. R - Cepstral coefficients, D - Delta cepstral coefficients, Dd - Delta-delta cepstral coefficients, IEWST - Inverse entropy weighting with static threshold, IEWAT - Inverse entropy weighting with average threshold.

Average Entropy values for PLP Features																	
Stream	Car Noise (in db)				Factory Noise (in db)				Lynx Noise (in db)				Noise50 (in db)				Clean
	0	6	12	18	0	6	12	18	0	6	12	18	0	6	12	18	Speech
R	1.27	1.18	1.12	1.09	1.24	1.32	1.23	1.14	1.96	1.68	1.43	1.27	2.07	1.81	1.54	1.35	1.13
D	1.33	1.28	1.23	1.18	1.82	1.74	1.51	1.30	1.76	1.55	1.36	1.23	1.79	1.64	1.43	1.27	1.07
Dd	1.52	1.46	1.40	1.35	1.85	1.82	1.64	1.46	1.99	1.79	1.58	1.44	1.91	1.79	1.60	1.45	1.24
R-D	0.94	0.85	0.80	0.76	0.90	1.07	0.99	0.86	1.46	1.24	1.01	0.87	1.47	1.31	1.07	0.91	0.76
R-Dd	1.01	0.89	0.82	0.78	0.98	1.11	1.00	0.87	1.66	1.36	1.10	0.93	1.70	1.45	1.17	0.98	0.80
D-Dd	1.14	1.09	1.04	1.00	1.61	1.55	1.31	1.11	1.65	1.41	1.20	1.07	1.56	1.44	1.23	1.07	0.90
R-D-Dd (Baseline)	0.88	0.79	0.74	0.71	0.93	1.07	0.95	0.82	1.47	1.21	0.97	0.82	1.48	1.29	1.04	0.87	0.72
Inverse Entropy	1.14	1.03	0.96	0.92	1.22	1.34	1.19	1.03	1.78	1.49	1.23	1.05	1.76	1.57	1.29	1.10	0.91
IEWST	0.83	0.75	0.71	0.69	0.92	1.06	0.93	0.79	1.48	1.20	0.95	0.80	1.50	1.30	1.03	0.86	0.70
Minimum Entropy	0.51	0.46	0.44	0.43	0.59	0.68	0.60	0.51	0.96	0.78	0.62	0.51	1.00	0.86	0.68	0.56	0.44
IEWAT	0.81	0.73	0.69	0.66	0.87	0.99	0.88	0.75	1.35	1.12	0.90	0.77	1.36	1.20	0.97	0.81	0.66

Table 3: Entropy values for PLP features. Variable number of hidden units for each network (Table 1). The baseline full-band system is *R-D-Dd*. R - Cepstral coefficients, D - Delta cepstral coefficients, Dd - Delta-delta cepstral coefficients, IEWST - Inverse entropy weighting with static threshold, IEWAT - Inverse entropy weighting with average threshold. The corresponding WER Table 2

Word-Error-Rates (WERs) for Rasta-PLP Features																	
Stream	Car Noise (in db)				Factory Noise (in db)				Lynx Noise (in db)				Noise50 (in db)				Clean Speech
	0	6	12	18	0	6	12	18	0	6	12	18	0	6	12	18	
R	20.4	18.5	15.8	14.7	62.2	36.0	23.3	17.9	45.6	31.3	19.4	16.4	54.4	36.8	24.0	16.6	13.8
D	16.8	15.5	14.3	12.9	57.4	35.1	22.9	16.3	45.3	28.5	18.7	15.0	53.0	32.5	21.3	15.2	13.0
Dd	18.1	17.5	16.4	16.3	60.7	37.5	25.8	18.7	49.9	34.4	24.0	18.7	54.4	34.4	24.2	18.4	14.8
R-D	15.4	13.0	12.5	11.6	57.6	35.9	20.4	13.9	43.1	25.2	17.0	12.5	50.7	31.8	19.6	15.0	11.2
R-Dd	15.5	12.5	11.4	11.2	57.4	31.3	18.1	12.6	40.6	23.9	16.2	11.8	46.4	27.5	17.3	12.9	11.7
D-Dd	12.9	11.6	11.7	10.8	55.6	34.2	20.6	15.2	42.4	25.1	16.6	12.6	49.3	28.3	18.7	13.8	10.5
R-D-Dd (Baseline)	13.3	11.4	10.8	10.4	56.2	33.1	18.9	12.7	38.4	22.2	14.6	11.0	48.2	26.0	17.0	13.3	10.2
Inverse Entropy	12.0	10.8	11.0	11.0	54.5	31.9	18.5	13.0	39.9	23.1	15.6	11.3	48.2	27.7	17.7	12.5	10.6
IEWST	11.2	10.0	9.1	9.1	55.2	31.8	18.1	12.5	40.1	22.5	15.1	10.5	47.4	27.6	17.0	12.2	10.1
Minimum Entropy	11.7	10.8	9.5	9.1	55.7	32.1	17.7	12.5	39.3	22.6	14.8	10.2	48.2	24.8	15.8	12.0	9.1
IEWAT	11.0	9.2	9.2	9.1	54.7	31.5	17.2	12.2	39.7	22.1	14.3	10.0	47.6	26.4	16.0	11.7	10.2

Table 4: Word-Error-Rates for Rasta-PLP features. Variable number of hidden units for each network (Table 1). The baseline full-band system is *R-D-Dd*. R - Cepstral coefficients, D - Delta cepstral coefficients, Dd - Delta-delta cepstral coefficients, IEWST - Inverse entropy weighting with static threshold, IEWAT - Inverse entropy weighting with average threshold.

Average Entropy values for Rasta-PLP Features																	
Stream	Car Noise (in db)				Factory Noise (in db)				Lynx Noise (in db)				Noise50 (in db)				Clean
	0	6	12	18	0	6	12	18	0	6	12	18	0	6	12	18	Speech
R	1.46	1.38	1.32	1.27	1.94	1.76	1.54	1.38	1.88	1.66	1.48	1.36	1.82	1.66	1.48	1.37	1.25
D	1.35	1.30	1.25	1.21	1.70	1.65	1.46	1.29	1.61	1.46	1.32	1.22	1.61	1.52	1.37	1.25	1.14
Dd	1.55	1.49	1.44	1.39	1.89	1.86	1.68	1.52	2.00	1.81	1.63	1.49	1.84	1.76	1.60	1.48	1.30
R-D	1.10	1.03	0.98	0.96	1.53	1.48	1.25	1.07	1.48	1.30	1.12	1.01	1.44	1.34	1.16	1.03	0.89
R-Dd	1.06	0.98	0.91	0.87	1.53	1.42	1.18	1.00	1.52	1.28	1.08	0.96	1.46	1.31	1.10	0.97	0.82
D-Dd	1.12	1.07	1.03	0.99	1.50	1.46	1.26	1.08	1.46	1.30	1.13	1.02	1.41	1.33	1.16	1.03	0.91
R-D-Dd (Baseline)	0.94	0.87	0.82	0.78	1.42	1.33	1.09	0.90	1.36	1.16	0.97	0.85	1.29	1.18	0.98	0.86	0.74
Inverse Entropy	1.22	1.14	1.07	1.03	1.67	1.60	1.35	1.16	1.64	1.43	1.23	1.10	1.56	1.46	1.25	1.11	0.97
IEWST	0.94	0.86	0.82	0.78	1.45	1.36	1.11	0.92	1.41	1.19	0.98	0.86	1.34	1.22	1.02	0.88	0.75
Minimum Entropy	0.60	0.55	0.52	0.49	0.96	0.89	0.72	0.60	0.93	0.78	0.64	0.55	0.91	0.81	0.67	0.57	0.47
IEWAT	0.89	0.83	0.78	0.75	1.30	1.23	1.02	0.86	1.27	1.09	0.92	0.81	1.21	1.12	0.95	0.83	0.71

Table 5: Entropy values for PLP features. Variable number of hidden units for each network (Table 1). The baseline full-band system is *R-D-Dd*. R - Cepstral coefficients, D - Delta cepstral coefficients, Dd - Delta-delta cepstral coefficients, IEWST - Inverse entropy weighting with static threshold, IEWAT - Inverse entropy weighting with average threshold. The corresponding WER Table 4

Moreover, for PLP features, the performance of the proposed entropy weighting schemes is either better or comparable to standard full-band system under different noise conditions. Of all the noises considered in this paper, the performance in the presence of *factory noise* is generally the worst. In most of the cases it is the Inverse entropy weighting with average threshold (IEWAT) which performs the best (except in few cases where Inverse entropy weighting performs slightly better). In case of factory and noise50 noises, usually the improvement in the performance is not significant but for the other two noises (car and lynx noise) the improvement in performance is as high as 22.6% (for car noise at 0 db by IEWAT) and 36.7% (for car noise at 0 db by Inverse entropy weighting). The average improvement in performance by different methods are as follows: **9.9% by Inverse entropy weighting, 9.2% by IEWST, 7.4% by Minimum entropy criterion and *11.8% by IEWAT*.**

Similarly, for J-Rasta PLP features, the performance of the proposed methods is either better or comparable to standard full-band system under different noise conditions, but the improvement in performance is less significant as compared to the PLP features. Again the IEWAT performs the best in most of the cases. The Inverse entropy weighting fails most of the times and in few cases it is the Minimum entropy criterion which performs slightly better than IEWAT. The average improvement in performances by different methods are: **-0.5% by Inverse entropy weighting, 4.2% by IEWST, 4.8% by Minimum entropy criterion and *6.9% by IEWAT*.**

Apart from WERs, the entropy tables (Table 3 and 5) also reveal few important things. Average entropy at the output of each expert is high for low SNR input speech signal. As mentioned earlier, the experts were trained on clean speech. When they were tested on noisy speech (mismatch conditions during training and testing), it is reflected in their output entropies. More the mismatch is, higher is the entropy. There is a slight exception to this rule in case of PLP features in presence of factory noise. In this particular case, the 0 dB SNR speech signal has less entropy as compared to 6 dB SNR speech signal. It could be because of the statistical conditions of noise in the particular case being too biased. Except this, in all the cases, we observe the trend that a low SNR speech signal at the input of an expert leads to a high entropy at the output of that expert.

There is something more to look at in terms of entropy of the combination for different weighting strategies. As stated earlier, entropy for any linear combination always lies between the highest and the lowest entropies among all the streams. Same can be observed from the entropy results of inverse entropy weighting. In this weighting the entropies for the combination are high and at the same time WER performance is also not significantly better (specially for J-Rasta PLP features where performance drops for this kind of weighting).

As expected, the minimum entropy criterion gives the least average entropy values. But this is a highly constrained situation where only the stream having the least entropy is chosen at every frame level. In this situation the other streams don't contribute in the decision. But from the results it can be seen that even this highly constrained situation gives an improvement in the WER performance as well as a decrease in entropy. Out of the two other non-linear combinations, average entropy 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 least (except the highly constrained case of Minimum entropy mentioned above).

5 Conclusion

Though it is a preliminary conclusion, in general, we observe that the WER for a combination strategy is lower when entropy is low. Though decreasing the entropy does not always give the best result (Minimum entropy criterion's performance is inferior to IEWAT), but it can be stated that if constraints are put properly, the WER and entropy both decrease. And it indicates a strong correlation between the two. It would be interesting to do more studies in this direction and see if the concept can be strengthened.

From the experimental results we can conclude that the entropy based weighting scheme for full-band multi-stream approach and its variations proposed in this paper help in improving the performance. The streams considered in our paper are the basic streams and it is felt that entropy based weighting can further improve the performance if streams carry information which is more complementary. It is seen that the performance is best for IEWAT. This is a case where threshold for entropy is dynamic and dynamically the streams having low entropies are the one which are considered for combination at every frame level. It opens up a new research direction where some new methods can be thought to reject the streams which are less reliable and this reliability is expressed in their respective entropies.

References

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

[2] H. Bourlard, S. Dupont, and C. Ris, "Multi-stream speech recognition," IDIAP-RR 7, IDIAP, 1996.

[3] H. Bourlard, S. Dupont, H. Hermansky, and N. Morgan, "Towards sub-band based speech recognition," in *Proceedings of European Signal Processing Conference*, (Italy), pp. 1579–1582, 1996.

[4] H. Bourlard and S. Dupont, "Sub-band based speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 1251–1254, 1997.

[5] H. Bourlard, "Non-stationary multi-channel (multi-stream) processing towards robust and adaptive ASR," in *Proceedings of the ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pp. 1–9, 1999.

[6] A. Hagen, A. Morris, and H. Bourlard, "Sub-band based speech recognition in noisy conditions: The full-combination approach," IDIAP-RR 15, IDIAP, 1998.

[7] A. Hagen and A. Morris, "Comparison of HMM experts with MLP experts in the full combination multi-band approach to robust ASR," in *Proceedings of International Conference on Spoken Language Processing*, (Beijing, China), 2000.

[8] A. Hagen, *Robust Speech Recognition based on Multi-stream Processing*. PhD dissertation, École Polytechnique Fédérale de Lausanne, Départment d'Informatique, EPFL, Lausanne, Switzerland, Dec. 2001.

[9] A. Morris, A. Hagen, and H. Bourlard, "The full combination sub-bands approach to noise robust HMM/ANN based ASR," in *Proceedings of Eurpean Conference on Speech Communication and Technology*, vol. 2, pp. 599–602, 1999.

[10] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Comm.*, vol. 34, pp. 25–40, 2001.

[11] A. Hagen, A. Morris, and H. Bourlard, "From multi-band full combination approach to multi-stream full combination processing in robust ASR," in *ISCA Tutorial and Research Workshop ASR2000*, (Paris, France), pp. 175–180, 2000.

[12] A. Hagen and H. Bourlard, "Using multiple time scales in the framework of multi-stream speech recognition," in *Proceedings of International Conference on Spoken Language Processing*, vol. 1, (Beijing, China), pp. 349–352, 2000.

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

[14] H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*. 101 Philip Drive, Assinippi Park, Norwell, Massachusetts 02061 USA: Kluwer Academic Press, 1994.

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

[16] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," *To be published in Speech Communication*, 2002.

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

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

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

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