IDIAP RESEARCH REPORT

Spectral Entropy Feature in Multi-Stream for Robust ASR

Hemant Misra [a] [b] Hervé Bourlard [a] [b]

IDIAP–RR 05-45

SUBMITTED FOR PUBLICATION

[a] IDIAP Research Institute, Martigny, Switzerland
[b] EPFL - Swiss Federal Institute of Technology, 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 05-45

Spectral Entropy Feature in Multi-Stream for Robust ASR

Hemant Misra Hervé Bourlard

SUBMITTED FOR PUBLICATION

Abstract. In recent papers, entropy computed from sub-bands of the spectrum was used as a feature for automatic speech recognition. In the present paper, we further study the sub-band spectral entropy features which can give the flatness/peakiness of the sub-band spectrum and in turn the position of the formants in the spectrum. The sub-band spectral entropy features are used in hybrid hidden Markov model/artificial neural network systems and are found to be noise robust. The spectral entropy features are investigated along with PLP features in multi-stream combination. Separate multi-layer perceptrons (MLPs) are trained for PLP features, spectral entropy features and both the features concatenated. The output posteriors of the three MLPs are combined after weighting such that the weight to a particular MLP's outputs are inversely proportional to the entropy of the output posterior distributions of that MLP. In Tandem framework, the combined output, after decorrelation, is fed to standard hidden Markov model/Gaussian mixture model system. Significant improvement in performance is reported when spectral entropy features are used along with PLP features in multi-stream combination.

1 Introduction

Feature extraction is an integral part of any automatic speech recognition (ASR) system. Standard features used in present ASR systems include mel-frequency cepstral coefficients (MFCC) [1], perceptual linear prediction (PLP) [2] and RASTA [3] based cepstral coefficients.

Robustness is an important issue in ASR systems, that is, an ASR system should be able to perform well under different conditions. There are several methods to improve the robustness of an ASR system, for example, a) doing cepstral mean subtraction (CMS) [4] and variance normalization at feature level reduces the mismatch across different channels, b) in multi-condition training, the data collected from different environments is used to train the models thus reducing the sensitivity to unknown (noise) conditions encountered at the time of testing [5]. Yet another technique to improve the noise robustness that has emerged in the recent past is multi-stream combination [6]. In multi-stream combination, different feature representations are obtained from the speech signal and modelled either jointly or separately. Assuming that different feature representations have different performance (error) characteristics, if they are combined properly we can achieve performance which could be better than the performance of the individual feature representations. Generally, in multi-stream combination, training is done on clean speech. At the time of testing, the weights assigned to different streams are adapted.

In this paper, we study the robustness issue at feature and posterior levels in multi-stream combination systems. In an earlier paper [7], we proposed spectral entropy computed from the short-time-Fourier transform (STFT) of the speech signal as a feature for ASR. In this paper, we explore the idea further and study the performance of the proposed spectral entropy features along with PLP features in multi-stream combination.

In the next section, we discuss the motivation for studying spectral entropy features and we present the idea of multi-band spectral entropy features. In Section 3, we explain the full-combination multi-stream (FCMS) and the inverse entropy weighting approaches. In Section 4, we present the Tandem system and extend the idea of inverse entropy weighting to combine Tandem systems. The database used to carry out the studies as well as the ASR system implementation details are described in Section 5. In the next section, we present the results followed by conclusions.

2 Spectral Entropy Feature

Entropy is a measure commonly used in communication theory to find the information content of a message. The entropy measure can also be employed to measure the "peakiness" of a probability density function (PDF) or probability mass function (PMF). A flat PMF does not carry any information and has the highest entropy possible while a PMF with a peak for only one class gives information about the high probability of that particular class and has low entropy. Therefore entropy can be used to measure the peakiness of a distribution.

STFT spectrum of a speech signal is characterized by peaks and valleys. Peaks usually correspond to the location of the formants and have relatively fixed position for a particular sound. At the same time, formants are less sensitive to noise as compared to valleys, for example, voiced sounds have strong formants and are less affected by noise while unvoiced sounds having weak formants get easily affected by noise. In [8], the author used the position of the formants as additional features for ASR. Similar idea of formant location as features were recently tried in spectro-temporal activity pattern (STAP) features for noise robust ASR [9]. In the same spirit, in [7], spectral entropy was used as features to capture the position of the formants and use them in ASR.

STFT spectrum not being a PMF we cannot compute entropy of it. Nevertheless, we can normalize the spectrum and convert it into a PMF like function.

$$s_i = S_i / \sum_{i=1}^{N} S_i \qquad \text{for } i = 1 \text{ to } N \tag{1}$$

where S_i is the energy of i^{th} frequency component of the spectrum, $\mathbf{s} = (s_1, \cdots, s_N)$ is the PMF of the spectrum and N is the number of points in the spectrum (order of STFT). Entropy for each frame is then defined as:

$$H = -\sum_{i=1}^{N} s_i \log_2 s_i \tag{2}$$

Similar method to compute entropy from spectrum was used for end point detection of speech in noisy environments [10].

Entropy contour computed on full-band spectrum of clean speech is shown in Fig 1(b). We observe from the figure that full-band spectral entropy can be used as a measure to detect speech and silence.

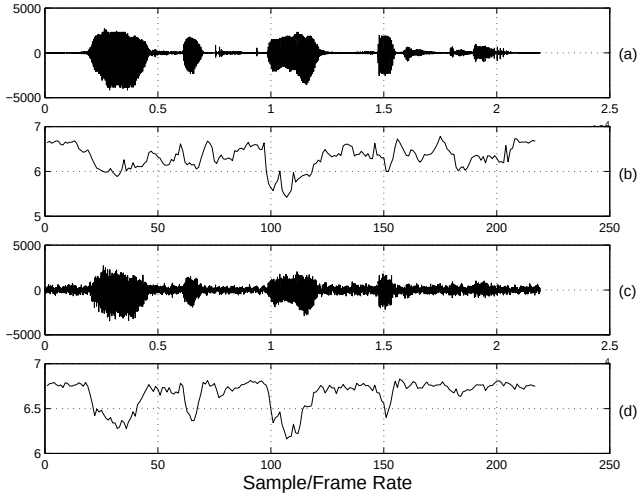

Figure 1: *Entropy computed from the full-band spectrum. (a) Clean speech wave form, (b) Entropy contour for clean speech, (c) Speech corrupted with factory noise at 6 dB SNR, and (d) Entropy contour for speech corrupted with factory noise at 6 dB SNR.*

For noisy speech, similar spectral entropy contour has been plotted in Fig 1(d). The figure shows that the dynamic range of the spectral entropy contour is reduced in presence of noise but it retains its discriminatory properties. This can be attributed to that fact that speech sounds have clear formant structure and formants are relatively insensitive to noise. Thus spectral entropy for speech sounds is low.

The inadequacy of the the full-band spectral entropy is that it can capture only the gross peakiness/flatness of the spectrum but not the position of the formants. To overcome this problem, we suggested [7] the idea of multi-band spectral entropy features which is explained in the following section.

2.1 Multi-band spectral entropy features

The way entropy is computed for the full-band spectrum, similarly we can divide the spectrum into sub-bands and compute entropy in each sub-band. The sub-band entropy can give the absence or presence of formants in a particular sub-band. In [7], the full-band spectrum was divided into sub-bands, where sub-bands could be non-overlapping or overlapping. The sub-band entropy was computed as follows: We normalized the full-band spectrum using Eq 1 and divided the normalized full-band spectrum into **J** non-overlapping sub-bands of equal size. The value of **J** decides the number of sub-bands, which in turn decided the dimension of the entropy feature vector. When $\mathbf{J = 1}$, we work with full-band and extract spectral entropy feature vector of dimension one. For J = 2, we divide the spectrum into two sub-bands and get two dimensional spectral entropy feature vector, one component from each sub-band. In our experiments, we changed the value of **J** from **1** to **32**. Additionally, we did one

experiment with 24 overlapping sub-bands where we used mel-scale [1] for defining the sub-bands. In our studies, we also appended the delta and double-delta features of the spectral entropy features to incorporate the temporal information.

3 Multi-stream Combination

In multi-stream combination, the knowledge or decisions of more than one experts are combined to get an improved performance. The underlying principle of multi-stream combination is to obtain a better estimate of the optimal decision rule by combining several experts with different error characteristics and/or complementary source of information.

Multi-stream combination can help in improving the performance of the baseline system if different streams are corrupted differently under noise conditions and not all of the streams undergo the same kind of degradation.

The two important issues in multi-stream combination are:

1. The features used for every stream should carry complementary information and all the feature streams must not go through the same distortions in presence of noise.
2. The weight given to each stream in combination should be defined such that the reliable streams get more weight while the streams corrupted by noise should be deemphasized. Moreover, the weight adaptation should be dynamic as the useful information content of each stream may change with time.

We have used multi-band spectral entropy features discussed in the previous section along with PLP features. We carried out our studies in the framework of hybrid hidden Markov model/artificial neural network (HMM/ANN) [11] system. Furthermore, we used a special case of multi-stream system which is referred to as full-combination multi-stream (FCMS) [12, 13]. FCMS for HMM/ANN system is depicted in Fig. 2. In FCMS, all the possible combinations of the individual feature representations

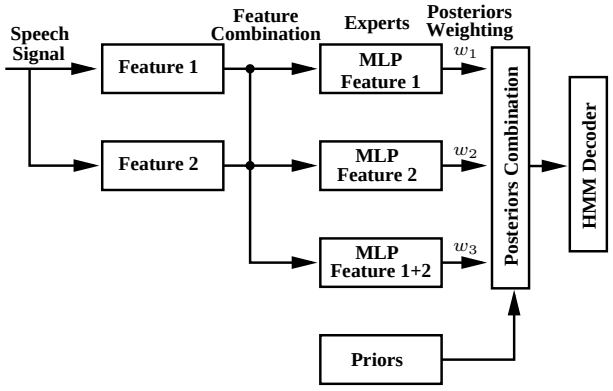

Figure 2: *FCMS for hybrid HMM/ANN system: All possible combinations of the two features are treated as separate streams. An MLP expert is trained for each stream. The posteriors at the output of experts are weighted and combined. The combined posteriors thus obtained are passed to an HMM decoder.*

are treated as separate streams and an multi-layered perceptrons (MLP) model is trained for each such feature stream. In FCMS, for n feature representations, we get $2^n - 1$ feature streams and need to train one MLP for each such feature stream.

In hybrid HMM/ANN system, an MLP with one hidden layer is trained for the given feature representation. The input to the MLP is the feature vector usually with a context of four neighbouring feature vectors on either side and output of the MLP is same as the number of classes (phonemes in case of phoneme based ASR). More explanation about hybrid HMM/ANN system used in the present paper is given in Section 5.

3.1 Inverse entropy based weighting in FCMS

In FCMS hybrid HMM/ANN system, we train one MLP expert for each possible combination of feature representations (Fig. 2). At the time of testing, we obtain posteriors at the output of the MLP classifiers. From these posteriors, for each classifier, we can compute the entropy at the output of the classifier. Before going any further, we would like to emphasize that the spectral entropy feature vector discussed in Section 2 was extracted from the speech signal and is different from the entropy at the output of a classifier. While spectral entropy is a feature, entropy at the output of a classifier indicates the confidence of the classifier. A classifier output with equal posterior probabilities for all the classes doesn't convey any information and has high entropy. On the contrary, a classifier with high posterior for one class and low posteriors for rest of the classes indicates a high confidence and has low entropy. Therefore, entropy at the output of a classifier can be used as a measure to weigh the outputs of a classifier. The output posteriors of a classifier with high entropy should be given less weight and vice-a-versa.

In Fig. 3, we show the relationship between the entropy of a classifier and the probability that the

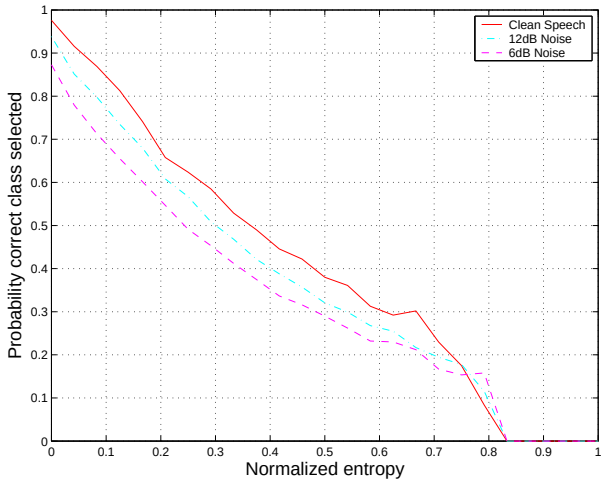

Figure 3: *Normalised entropy (horizontal) Vs probability that the largest probability selects the correct class (vertical). The plot is for the MLP trained on clean data and tested for the following noisy conditions: Clean (–), SNR12 (-.)and SNR6 (- -). Noise is factory noise from Noisex database.*

highest probability selects the correct class. As expected, the relationship is approximately linear and inverse, that is, accuracy is low for high entropy and vice-a-versa. The relationship holds for different noise conditions which were emulated by adding factory noise from Noisex92 database [14] at different signal-to-noise-ratios (SNRs) to the Numbers95 corpus [15].

In [16] and [17], similar weighting approaches were suggested for multi-band and multi-stream combinations, respectively. We have used the inverse entropy based weighting criterion suggested in [17]. The entropy at the output of an MLP classifier is computed by:

$$h_n^i = -\sum_{k=1}^{K} P(q_k|x_n^i, \theta_i) \log_2 P(q_k|x_n^i, \theta_i) \tag{3}$$

where K is the number of output classes or phonemes, x_n^i is the input acoustic feature vector for the i^{th} stream for the n^{th} frame, θ_i is the parameter set of the i^{th} MLP expert, and $P(q_k|x_n^i, \theta_i)$ is the posterior probability estimate for the k^{th} class at the output of the i^{th} MLP for n^{th} frame.

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

to:

$$\hat{P}(q_k|X_n,\Theta) = \sum_{i=1}^{I} w_n^i P(q_k|x_n^i,\theta_i) \tag{4}$$

where I is the number of experts or streams (3 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. In *Inverse entropy weighting with average entropy at each frame level as threshold*, the average entropy of all the streams for a frame is calculated by the equation,

$$\bar{h}_n = \frac{\sum_{i=1}^{I} h_n^i}{I} \tag{5}$$

This average entropy is used as a threshold for the frame and output of all the experts having entropy greater than the threshold are weighted 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 for *Inverse entropy weighting with average threshold* (IEWAT) are:

$$\tilde{h}_n^i = \begin{cases} 10000 & : \quad h_n^i > \bar{h} \\ h_n^i & : \quad h_n^i \leq \bar{h} \end{cases} \tag{6}$$

$$w_n^i = \frac{1/\tilde{h}_n^i}{\sum_{i=1}^{I} 1/\tilde{h}_n^i} \tag{7}$$

4 Tandem System

The hybrid HMM/ANN system does discriminative training, and the output of hybrid systems being posterior probabilities, the system is a good candidate for multi-stream combination. In contrast, HMM/GMM based systems do likelihood based training and it is easier to incorporate techniques like context-dependent modelling and state tying in HMM/GMM system which give an additional improvement in the performance of the system. Tandem [5] is a combination of HMM/ANN system followed by some processing of the MLP outputs before being fed to a HMM/GMM system. The two forms of Tandem system are shown in Figs. 4 and 5. In the first model (*Tandem Model 1 or TM1*),

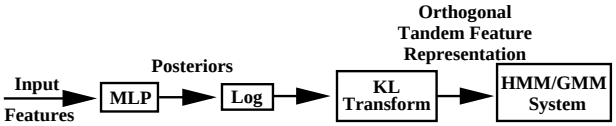

Figure 4: *Tandem Posterior Model (TM1): 'Posteriors' from the MLP are Log scaled and then decorrelated by KL transformation. The transformed posteriors are used as a feature in standard HMM/GMM systems.*

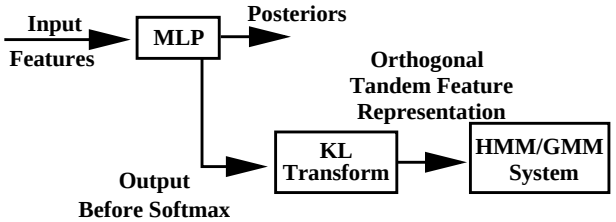

Figure 5: *Tandem Linear Model (TM2): 'Output before softmax' from the MLP are decorrelated by KL transformation and used as a feature in standard HMM/GMM systems.*

logarithmic of the posterior outputs of the MLP is decorrelated by Karhunen-Loeve (KL) transform.

The transformed outputs are given to a standard HMM/GMM system as features and models are created. In the second system (*Tandem Model 2 or TM2*), the outputs are taken from the MLP before the softmax non-linearity of the output layer. These outputs are decorrelated by KL transform and then fed to the HMM/GMM system.

The processing of the MLP outputs before feeding them to HMM/GMM system of the second stage is required for two reasons: 1) The MLP outputs at posterior levels are usually skewed, and 2) The MLP outputs are correlated. The processing ensures that the input to the HMM/GMM system is Gaussian like and uncorrelated and therefore can be modelled by the system.

The relation between the MLP output before softmax and after softmax (posterior estimates) is given by

$$P(q_k|x_n) = \frac{exp(y_k|x_n)}{\sum_k exp(y_k|x_n)} \tag{8}$$

where $y_k|x_n$ and $P(q_k|x_n)$ are the output before and after softmax, respectively, for k^{th} class and feature vector x_n at time instant n. The output after the softmax, $P(q_k|x_n)$, is the estimated posterior probability at the output of the MLP for the k^{th} class. The relation between output before and after softmax is many-to-one mapping and we lose some information in the process. Moreover, the output before softmax is more Gaussian like as compared the output after softmax. In earlier studies, it was observed that the *TM1* is little inferior to *TM2* [18].

4.1 Tandem in multi-stream combination

In this section, we describe the idea of Tandem in the framework of multi-stream combination. Similar ideas were proposed and investigated earlier also [19]. Figs. 6 and 7 illustrate multi-stream combination for *TM1* and *TM2*, respectively.

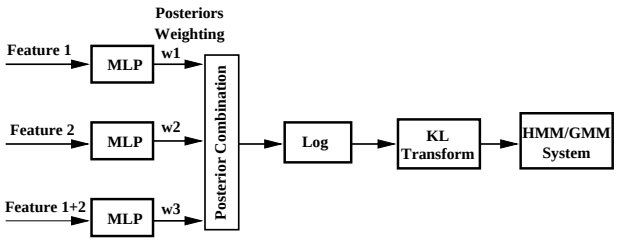

Figure 6: *Multi-stream TM1: 'Outputs after softmax' from different experts are weighted and combined. The combined output undergoes Log scaling followed by KL transform before being fed as features into HMM/GMM systems.*

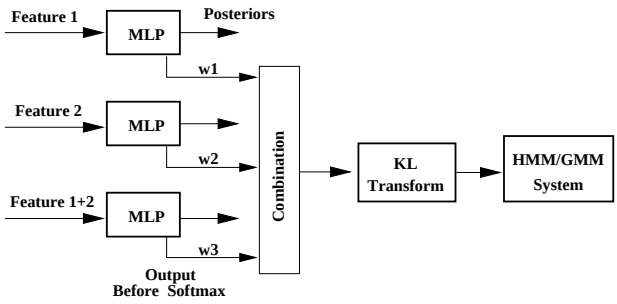

Figure 7: *Multi-stream TM2: 'Outputs before softmax' from different experts are weighted and combined. The combined output undergoes KL transform before being fed as features into HMM/GMM systems.*

Here we extend the idea of inverse entropy weighting introduced in Section 3 for hybrid HMM/ANN system to the Tandem. In case of *TM1*, we weigh the outputs from different MLPs with the weights inversely proportional to their respective entropies computed from the posteriors at the output. Log of the combined outputs is decorrelated and given as features to a HMM/GMM system. While in *TM1*, entropy can be computed from posteriors at the output of the MLP expert, same cannot be done in *TM2* where we have access to the outputs before softmax non-linearity. To circumvent this problem, we use Eq. 8 to convert the linear outputs to posterior estimates and compute entropy for each expert from these posterior estimates. The linear outputs before softmax obtained from each MLP expert are weighted with weights inversely proportional to their respective entropies. The combined output thus obtained is decorrelated and given as features to a HMM/GMM system.

5 Database and Experimental Setup

In this paper, we have used Numbers95 [15] U.S. English connected digit database. The database consists of 30 words represented by 27 phonemes, including silence. In the experiments, we used 3330 utterances for training and 2550 utterances for testing. Training was performed only on clean utterances, and to simulate noisy test conditions, we added noise from Noisex92 database [14] to test utterances at different SNRs.

Our baseline features were 13 PLP [3] cepstral coefficients appended by their first and second order time derivatives. Cepstral mean subtraction (CMS) and variance normalization was applied to the feature vectors on per utterance basis. Multi-band spectral entropy feature vector was extracted by dividing the full-band into sub-bands and obtaining one entropy value from each sub-band. In case of non-overlapping sub-bands, the number of sub-bands were varied from 1 to 32. In the experiment with overlapping sub-bands, 24 sub-bands as defined by mel-scale were used. The multi-band spectral entropy feature vector was developed by appending first and second order time derivatives of the spectral entropy feature vector.

In our first set of experiments, we used hybrid HMM/ANN systems. One MLP with single hidden layer was trained for each feature stream. The number of units in the hidden layer were proportional to the dimension of input feature vector. A context of 9 frames was used at the input of the MLP. The output layer had 27 units, one for each phoneme class. The HMM used for decoding had 1 state mono-phone model for each phoneme and scaled posteriors were supplied as emission likelihoods to it. The minimum duration of each phoneme was modelled by forcing 1 to 3 repetitions of the same state.

The Tandem system was implemented with the MLP of the hybrid system (as discussed above) in first stage followed by HMM/GMM system in the second stage. The outputs of the MLP (27) were either passed through log scale and then KL transformed (*TM1*) or were KL transformed (*TM2*). In both the cases, all the 27 dimensions were retained after the KL transform. The HMM/GMM part of Tandem consists of 80 context dependent phones with 3 left-to-right states per context dependent phone. For each state, emission probabilities were modelled by 12 mixture GMMs. More details about the Tandem system can be found in the literature [5, 18].

6 Results

We used hybrid HMM/ANN system as well as Tandem system to evaluate the ASR performances. In Sections 6.1 and 6.2 we present the results obtained by the two systems.

6.1 Performance: HMM/ANN system

The performance of the multi-band spectral entropy for hybrid HMM/ANN system for different setups is given in this section. Table 1 gives the performance of the multi-band spectral entropy feature appended by its time derivatives. As the number of sub-bands increase, we observe an improvement

Feature	clean	SNR12	SNR6	SNR0
16-bands	15.5%	22.0%	31.9%	53.2%
24-bands	14.0%	20.2%	29.3%	50.1%
32-bands	14.0%	20.4%	28.8%	47.1%
24 Mel-bands	12.8%	18.3%	*27.0%*	*45.1%*
PLP	*10.0%*	*17.7%*	29.6%	51.0%

Table 1: *WERs for spectral entropy features with its first and second order time derivatives appended in hybrid system for noisy speech.* **Only Mel-bands are overlapping**.

in performance, that is, word-error-rate (WER) decreases (the results for number of sub-bands less than 16 are not presented here and can be found in [7]). All the results except the last row are for non-overlapping sub-bands. The last row shows the result when overlapping sub-band as defined by mel-scale are used. The performance of the PLP feature is given for comparison. Further, we have given the performance for different noise conditions generated by adding factory noise to Numbers95 database at various SNRs. The results reveal two things: 1) Among all the sub-bands considered, overlapping mel-scale defines the best sub-bands to generate multi-band spectral entropy feature, and 2) Multi-band spectral entropy feature performs well at low SNRs and poorly at high SNRs as compared to the PLP features.

Finally, we give the results of multi-stream combination approach studied in this paper. Results for individual streams are reproduced from the earlier table. Table 2 also lists the result obtained by appending the PLP feature to the multi-band spectral entropy feature defined by mel-scale. In the

Feature	Clean	SNR12	SNR6	SNR0
PLP	10.0%	17.7%	29.6%	51.0%
24-Mel	12.8%	18.3%	27.0%	45.1%
PLP + 24-Mel	9.6%	15.8%	28.1%	51.7%
FCMS	9.2%	15.0%	24.5%	45.5%

Table 2: *Hybrid system under different noise conditions: WERs for PLP feature, 24 Mel-band spectral entropy feature and its time derivaties (24-Mel), the two features appended (PLP + 24-Mel), and PLP and spectral entropy feature in FCMS with inverse entropy weighting.*

same table, the result of using PLP and multi-band spectral entropy features defined by mel-scale in full-combination multi-stream (FCMS) with inverse entropy weighting are shown. Appending the features improves the performance, but better improvements are observed by FCMS where we model the features first and then combine the outputs of the experts.

6.2 Performance: Tandem system

In this section, we show the results obtained on Tandem system. The Tandem results are better than the hybrid HMM/ANN results but hybrid results were useful to study the spectral entropy features and choose the best candidate among all the spectral entropy features to do the experiments on Tandem system. In rest of the experiments, we have used the overlapping sub-bands defined by mel-scale to extract the spectral entropy features.

The results for *TM1* and *TM2* are given in Tables 3 and 4, respectively. As pointed out earlier, *TM2* performs better as compared to *TM1*. Further, the trend observed in hybrid HMM/ANN system that FCMS performs better than feature combination is visible in both the Tandem systems. This result is on the expected lines. If one of the feature representation gets corrupted by noise and the other remains less affected, it is reasonable to model the two features representations separately and then combine their outputs as compared to concatenating the two feature representations and

Feature	clean	SNR12	SNR6	SNR0
PLP	5.5%	12.0%	22.1%	44.2%
24-Mel	8.6%	13.9%	22.1%	40.8%
PLP + 24-Mel	5.5%	11.9%	22.2%	45.1%
FCMS	5.2%	10.9%	19.6%	39.8%

Table 3: *Tandem system TM1 under different noise conditions: WERs for PLP feature, 24 Mel-band spectral entropy feature and its time derivaties (24-Mel), the two features appended (PLP + 24-Mel), and PLP and spectral entropy feature in FCMS with inverse entropy weighting.*

Feature	Clean	SNR12	SNR6	SNR0
PLP	4.3%	10.3%	20.1%	41.9%
24-Mel	7.1%	12.1%	19.9%	37.7%
PLP + 24-Mel	4.2%	9.7%	18.5%	41.1%
FCMS	4.0%	9.6%	17.6%	37.5%

Table 4: *Tandem system TM2 under different noise conditions: WERs for PLP feature, 24 Mel-band spectral entropy feature and its time derivaties (24-Mel), the two features appended (PLP + 24-Mel), and PLP and spectral entropy feature in FCMS with inverse entropy weighting.*

modelling them jointly.

In short, entropy weighted FCMS consistently gives an improvement as compared to the baseline. Improvement is more significant at lower SNRs when the baseline performance is poor indicating more complementarity between the two feature streams at lower SNRs which FCMS is able to use in a better way.

7 Conclusions

In this paper, we investigated the multi-band spectral entropy features further. The new features were shown to be robust to noise but performed poorly for clean speech when compared to PLP features. The goal of having an ASR system which works well for all conditions was accomplished by using multi-band spectral entropy features along with the PLP features in the framework of multi-stream combination. In multi-stream combination, we showed that full-combination multi-stream gives better improvement in performance as compared to appending the features. Moreover, in FCMS, we studied a combination approach where weights to the output of different MLP experts were inversely proportional to the entropy at the output of the experts. Using the above two techniques, we obtain significant improvements in ASR performance. We validated these findings on two different ASR models, hybrid HMM/ANN and Tandem.

8 Acknowledgements

We wish to thank Dr. Sunil Sivadas, Dr. Andrew C. Morris and Prof. Hynek Hermansky for their 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.

References

[1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 357–366,

1980.

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

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

[4] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[5] H. Hermansky, D. P. W. Ellis, and S. Sharma, "TANDEM connectionist feature extraction for conventional HMM systems," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, (Istanbul, Turkey), 2000.

[6] 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*, (Philadelphia, PA, USA), pp. 426–429, Oct. 1996.

[7] H. Misra, S. Ikbal, S. Sivadas, and H. Bourlard, "Multi-resolution spectral entropy feature for robust ASR," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, (Philadelphia, U.S.A.), Mar. 2005.

[8] M. Padmanabhan, "Spectral peak tracking and its use in speech recognition," in *Proceedings of International Conference on Spoken Language Processing*, (Beijing, China), 2000.

[9] S. Ikbal, M. M. Doss, H. Misra, and H. Bourlard, "Spectro-temporal activity pattern (STAP) features for noise robust ASR," in *Proceedings of International Conference on Spoken Language Processing*, (Jeju Island, South Korea), Oct. 2004.

[10] J. lin Shen, J. weih Hung, and L. shan Lee, "Robust entropy-based endpoint detection for speech recogntion in noisy environments," in *Proceedings of International Conference on Spoken Language Processing*, (Sydney, Australia), 1998.

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

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

[13] A. Hagen and A. Morris, "Recent advances in the multi-stream HMM/ANN hybrid approach to noise robust ASR," *Computer Speech and Language*, no. 19, pp. 3–30, 2005.

[14] 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.

[15] 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.

[16] S. Okawa, T. Nakajima, and K. Shirai, "A recombination strategy for multi-band speech recognition based on mutual information criterion," in *Proceedings of European Conference on Speech Communication and Technology*, (Budapest, Hungary), pp. 603–606, Sept. 1999.

[17] H. Misra, H. Bourlard, and V. 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.

[18] D. P. W. Ellis and M. J. R. Gomez, "Investigations into tandem acoustic modeling for the aurora task," in *Proceedings of European Conference on Speech Communication and Technology*, (Denmark), Sept. 2001.

[19] S. Ikbal, H. Misra, S. Sivadas, H. Hermansky, and H. Bourlard, "Entropy based combination of Tandem representations for noise robust ASR," in *Proceedings of International Conference on Spoken Language Processing*, (Jeju Island, South Korea), Oct. 2004.