IDIAP RESEARCH REPORT

Spectral Entropy Feature in Full-Combination Multi-Stream for Robust ASR

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

IDIAP–RR 05-10

TO APPEAR IN
ISCA European Conference on Speech Communication and Technology (EUROSPEECH), September 2005, Lisbon, Portugal.

[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 05-10

Spectral Entropy Feature in Full-Combination Multi-Stream for Robust ASR

Hemant Misra Hervé Bourlard

TO APPEAR IN
ISCA European Conference on Speech Communication and Technology (EUROSPEECH), September 2005, Lisbon, Portugal.

Abstract. In a recent paper, we reported promising automatic speech recognition results obtained by appending spectral entropy features to PLP features. In the present paper, spectral entropy features are used along with PLP features in multi-stream framework. In our multi-stream hidden Markov model/artificial neural network system, we train a separate multi-layered perceptron (MLP) for PLP features, spectral entropy features and both the features combined by concatenation. The output posteriors from these three MLPs are combined with weights inversely proportional to the output entropies of the respective MLPs. On the Numbers95 database, this approach yields a considerable improvement both under clean and noisy conditions as compared to simply appending the features. Further, in multi-stream Tandem system, we apply the same inverse entropy weighting to combine the outputs of the MLPs before the softmax non-linearity. Feeding the combined outputs after decorrelation to the standard hidden Markov model/Gaussian mixture model system gives a 9.2% relative error reduction as compared to the baseline.

1 Introduction

In automatic speech recognition (ASR), cepstral features obtained from short time Fourier transform (STFT) of the speech signal are used for acoustic modelling, the common ones being Mel-frequency cepstral coefficients (MFCCs) [1], perceptual linear prediction (PLP) [2] and RASTA [3] based cepstral coefficients. Cepstral features capture the absolute energy response of the spectrum.

In [4], we proposed multi-resolution spectral entropy features obtained from STFT of the speech signal (see Section 2). We computed spectral entropy features from the sub-bands of spectrum in order to locate the spectral peaks of the spectrum which are supposed to be more robust to noise. In [5], we suggested improvements to the multi-resolution spectral entropy feature extraction method and showed that the features computed on overlapping mel-scale [1] and appended with first and second order time derivatives are robust to noise. Further, appending the spectral entropy features to the state-of-the-art PLP cepstral features improved the robustness of the ASR system both in hybrid hidden Markov model/artificial neural network (HMM/ANN) [6] as well as Tandem [7] systems. In [5], we observed that the appending of the features does not yield significant improvement in cases when there is a big difference between the performance of the individual feature streams.

In this paper, we study the spectral entropy features in the framework of full-combination multi-stream (FCMS) [8]. In FCMS, all possible combinations of the individual feature representations are used as a feature stream and an expert is trained for each such feature stream.

In the next section we describe multi-resolution spectral entropy feature. In the same section, we explain FCMS and the the Tandem [7] approach. The database and the experimental setup is discussed in Section 3. Section 4 contains the results followed by conclusions in Section 5.

2 Multi-resolution spectral entropy feature in multi-stream

2.1 Spectral entropy feature

Entropy measures can be used to capture the "peakiness" of a probability mass function (PMF). A PMF with sharp peak will have low entropy while a PMF with flat distribution will have high entropy.

The importance of formants is well known and in [9] the author used the location of spectral peaks as an additional feature in ASR. In the same spirit, the central idea in [4], while using multi-resolution spectral entropy as a feature, was to capture the peaks of the spectrum and their position.

To compute the entropy of a spectrum, we converted the spectrum into a PMF like function by normalizing it.

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

where X_i is the energy of i^{th} frequency component of the spectrum, $\mathbf{x} = (x_1, \cdots, x_N)$ is the PMF of the spectrum and N is the number of points in the spectrum (order of STFT). Spectral entropy for each frame was then defined as:

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

Fig. 1(b) shows the entropy contour computed on the full-band spectrum for the clean speech. We observe that entropy computed on full-band spectrum can be used as an estimate for speech/silence detection. In presence of noise, the formants are less affected as compared to the other parts of the spectrum. We can thus reasonably assume that entropy of the spectrum if used for speech/silence detection will be robust to noise, and indeed it is true as shown in Fig. 1(d). Though the dynamic range of the spectral entropy contour is reduced in presence of noise, it retains its discriminatory property. Same properties were reported in [10] where the authors used spectral entropy for end point detection of speech in noisy environments.

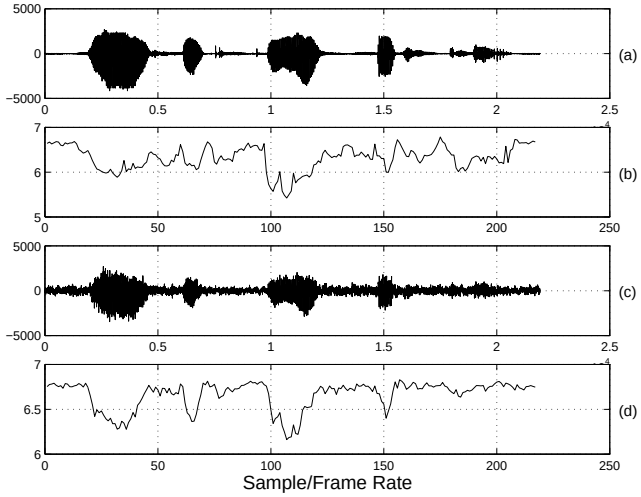

Figure 1: *Entropy computed from the full-band spectrum. (a) Clean speech wave form, (b) Entropy contour for clean speech, (c) Speech corrupted with factory noise at 6 dB SNR, and (d) Entropy contour for speech corrupted with factory noise at 6 dB SNR.*

2.2 Multi-band/multi-resolution spectral entropy feature

The full-band spectral entropy feature can capture only the gross peakiness of the spectrum but not the position of the formants. In [4, 5] we suggested multi-resolution/multi-band spectral entropy features to capture the position of the formants where we divided the spectrum into sub-bands and computed entropy of each sub-band. In [5], we obtained the best results by dividing the normalized full-band spectrum into 24 overlapping sub-bands defined by Mel-scale [1] and computed entropy from each sub-band. This is thus equivalent to *computing the entropy contribution of each sub-band to the full-band entropy.* Further, *we appended the first and second order time derivatives to incorporate temporal information.*

2.3 Entropy based full-combination multi-stream (FCMS)

As illustrated in Fig 2, in FCMS [8, 11], more than one type of feature representation is extracted

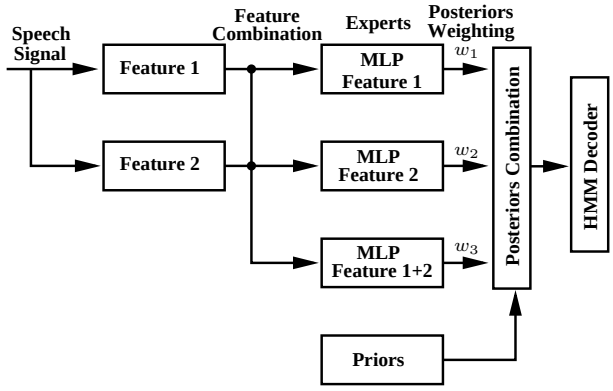

Figure 2: *Full-combination multi-stream: All possible combinations of the two features are treated as separate streams. An MLP expert is trained for each stream. The posteriors at the output of experts are weighted and combined. The combined posteriors thus obtained are passed to an HMM decoder.*

from the speech signal and every possible combination of the feature representations is treated as a

separate feature stream. In hybrid HMM/ANN approach, one multi-layered perceptron (MLP) with one hidden layer is trained for each such feature stream. The posteriors at the output of an MLP classifier (and entropy computed from these posteriors) indicate the confidence of the classifier. Before going any further, we would like to point out that the spectral entropy features we discussed in the Sections 2.1 and 2.2 were extracted from the speech signal and are different from the entropy at the output of an MLP classifier.

A classifier with equal posterior for all the classes has high entropy and does not convey any information. In contrast, a classifier with high posterior for one class and low posteriors for rest of the classes has low entropy and indicates that the classifier has high confidence. 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 [12] and [13], similar approaches were suggested for multi-band and multi-stream combinations, respectively. We have used the inverse entropy based weighting criterion suggested in [13]. 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 and θ_i is the parameter set of the i^{th} MLP expert.

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:

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

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

2.4 Spectral entropy feature in Tandem framework

The simplicity of the HMM/ANN hybrid system is that the input features to the MLP do not require pre-processing like decorrelation as an MLP by itself can learn the correlation among the features. Moreover, HMM/ANN systems perform discriminatory training and the output posteriors of the hybrid systems are well suited for multi-stream combination. In contrast, hidden Markov model/Gaussian mixture model (HMM/GMM) system does likelihood based generative training, and the advantage of HMM/GMM system is that modelling techniques like context dependent phone modelling and state-tying can be easily implemented in it.

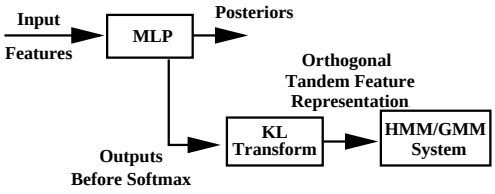

Figure 3: *Tandem Model: 'Outputs before softmax' from the MLP are decorrelated with the help of KL transformation and used as a features in standard HMM/GMM systems.*

In Tandem [7], the advantages of both HMM/ANN and HMM/GMM systems are exploited. As shown in Fig 3, the initial discriminatory modelling is done by an MLP. The output of the MLP (taken before the output layer's softmax non-linearity) is decorrelated by Karhunen-Loeve (KL) transformation. The KL transformed output is modelled by HMM/GMM system. The relation between the output of the MLP before softmax and after softmax is,

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

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.

In second set of experiments, we used Tandem system to ascertain the importance of spectral entropy features when a) used stand alone, b) are appended to the PLP cepstral features, and c) are used in FCMS along with PLP cepstral features as shown in Fig. 4.

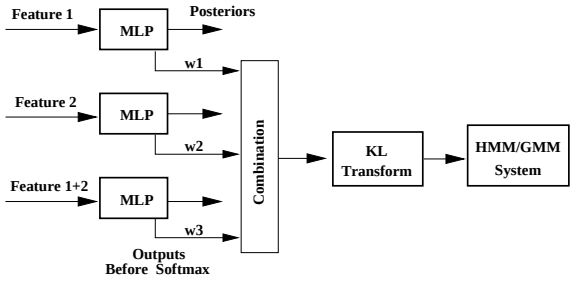

Figure 4: *Multi-stream Tandem: 'Outputs before softmax' from different experts are weighted and combined. The combined output undergoes KL transform before being fed as features into HMM/GMM systems.*

In case c), we have access to the 'outputs before softmax' but not to the posteriors at the output of the MLP to compute the entropy at the output of each MLP expert (Eq. 3). Therefore we cannot use the entropy based weighting directly. To overcome this problem, we converted the 'outputs before softmax' into posteriors using Eq 8. Then, the entropy at the output of each MLP expert is computed (Eq. 3) and is used for weighting the 'outputs before softmax' from the MLP expert. The combined outputs thus obtained are decorrelated and used as a feature in HMM/GMM system.

3 Experimental Setup

In the experiments reported in this paper, Numbers95 database of US English connected digits telephone speech [14] is used. There are 30 words in the database represented by 27 phonemes. Training is performed on clean speech utterances only while testing is performed either on clean speech or speech corrupted by factory noise from Noisex92 database [15] added at different signal-to-noise-ratios (SNRs)

to Numbers95 database. There were 3,330 utterances for training and 2,250 utterances were used for testing the system.

We have used HMM/ANN hybrid system for the first set of experiments. The ANNs used were a single hidden-layer MLPs and the number of units in the hidden layer of an MLP were proportional to the dimension of the input feature vector stream fed to that MLP. The baseline PLP [3] feature vectors used in our system were 13-dimensional cepstral coefficients appended with their first and second order time derivatives. The input layer was fed by 9 consecutive data frames. The HMM used for decoding had 1 state mono phone model for each phoneme for which emission likelihoods were supplied as scaled posteriors [6]. The minimum duration for each phoneme was modelled by forcing 1 to 3 repetitions of the same state for each phoneme.

Multi-band spectral entropy features extracted from 24 overlapping sub-bands and its first and second order time derivatives were used to develop spectral entropy feature based hybrid and Tandem systems. We ran experiments appending the spectral entropy features to the PLP features. Finally, in the frame work of FCMS, we used PLP and spectral entropy features.

The Tandem system was implemented with the hybrid HMM/ANN system discussed above followed by HMM/GMM system in the second stage. The HMM/GMM part of Tandem consists of 80 context dependent phones with 3 left-to-right states per context dependent phone and 12 GMMs per state to estimate emission probabilities within each state. We used HTK to train the system. The features to the HMM/GMM system were the 'outputs before softmax' of the hybrid system (after being decorrelated by KL transform) and were 27-dimensional. The implementation details of the Tandem system can be found in literature [7].

4 Results

The results for HMM/ANN hybrid system, in terms of word-error-rates (WERs), of the different features and their combinations are shown in Table 1. To simulate noisy conditions, we added non-

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 1: *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.*

stationary factory noise from Noisex92 database [15] at different SNRs to the speech signal. Also, it is worth mentioning here that changing the number of parameters in the MLP didn't change the performance of the individual features significantly. We observe that appending the features help to a certain extent, but in very high noise case we don't see any improvement by appending the features. In contrast, a considerable improvement in performance is observed for all conditions when the two features (PLP and spectral entropy) are used in entropy weighted FCMS.

In Table 2 we have shown the results in terms of WERs for Tandem system. When entropy feature vector is appended to the PLP features, more improvements are observed in cases when difference between the performance of PLP and spectral entropy features is not high (SNR12 and SNR6). When the difference between the performance of the PLP and entropy features is high, the gain in performance by appending the two types of feature is not significant. In comparison, the two features in entropy weighted FCMS do significantly better under all conditions as compared to the baseline as well as appending the two features.

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 2: *Tandem 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.*

5 Discussion and Conclusion

In search of new features having complementary information, this paper further investigated the use of multi-band spectral entropy as an additional feature. In this paper, we studied the performance of the proposed feature vector in entropy weighted FCMS. We demonstrated that better performance can be achieved by FCMS as compared to appending the multi-resolution entropy feature vector to the PLP feature vector. In fact, this is in line with the findings reported in the literature. Improved performance is achieved when the feature representations are modelled first and then combined instead of first combined (appended) and then modelled. Moreover, we suggested a method to combine the 'outputs before softmax' of MLPs in FCMS Tandem system.

6 Acknowledgements

We wish to thank Prof. Hynek Hermansky 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.

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] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral entropy based feature for robust ASR," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, (Montreal, Canada), May 2004.

[5] 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.

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

[7] 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.

[8] 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.

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

[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] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.