IDIAP RESEARCH REPORT

Multi-resolution Spectral Entropy Feature for Robust ASR

Hemant Misra [a b] Shajith Ikbal [a b]
Sunil Sivadas [a c] Hervé Bourlard [a b]

IDIAP–RR 04-37

PUBLISHED IN
IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), March 2005, Philadelphia, U.S.A.

[a] IDIAP Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland
[b] EPFL - Swiss Federal Institute of Technology, Lausanne, Switzerland
[c] Nokia Research Center, Tampere, Finland

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 04-37

Multi-resolution Spectral Entropy Feature for Robust ASR

Hemant Misra Shajith Ikbal Sunil Sivadas Hervé Bourlard

PUBLISHED IN
IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), March 2005,
Philadelphia, U.S.A.

Abstract. Recently, entropy measures at different stages of recognition have been used in automatic speech recognition (ASR) task. In a recent paper, we proposed that formant positions of a spectrum can be captured by multi-resolution spectral entropy feature. In this paper, we suggest modifications to the spectral entropy feature extraction approach and compute entropy contribution from each sub-band to the total entropy of the normalized spectrum. Further, we explore the ideas of overlapping sub-bands and the time derivatives of the spectral entropy feature. The modified feature is robust to additive wide-band noise and performs well at low SNRs. In the last, in the frame work of TANDEM, we show that the system using combined entropy and PLP features works better than the baseline PLP feature for additive wide-band noise at different SNRs.

1 Introduction

Acoustic modelling in automatic speech recognition (ASR) is generally accomplished by cepstral features obtained from short time Fourier transform (STFT) of speech signal. The most common ones among existing cepstral features are Mel-frequency cepstral coefficient (MFCC) [1], perceptual linear prediction (PLP) [2] and RASTA [3] based cepstral coefficients. While cepstral features are fairly good representation, they capture the absolute energy response of the spectrum. Further, we are not sure that all the relevant information present in the STFT spectrum is captured by them.

We followed a completely different approach while proposing our multi-resolution spectral entropy feature [4]. Instead of transforming the spectral information into cepstral domain, we suggested computing entropy from the sub-bands of spectrum and trying to locate the spectral peaks of the spectrum which are supposed to be more robust to noise. In [4], we showed that the proposed multi-resolution spectral entropy feature is not very competitive when compared to the state-of-the-art PLP cepstral features, but improves the robustness of the ASR system when appended to the PLP features.

In this paper, we suggest improvements to the feature computation and show that the modified multi-resolution spectral entropy feature along with its time derivatives is noise robust and performs better than the PLP feature under high noise conditions. The spectral entropy feature when used along with PLP feature performs better than the baseline PLP system. Further, when context dependent phone modelling and state tying are used on top of hidden Markov model (HMM)/artificial neural network (ANN) hybrid system, as in the case of Tandem [5] approach, the performance of the feature improves significantly, establishing the usefulness of the new feature.

The remaining paper is organized as follows: In the next section we introduce the basic multi-resolution spectral entropy feature and the improvements suggested over the basic feature. In the same section we explain the Tandem approach. In Section 3, the database used and the experimental setup is discussed. Section 4 contains the comparative results followed by conclusions in Section 5.

2 Multi-Resolution Spectral Entropy Feature

2.1 Motivation

Entropy 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. In a recent publication [6], spectral entropy rate, also known as Wiener entropy, has been used to measure the spectral flatness and explored as one of the feature for detecting stop consonants in continuous speech.

In case of STFT spectra of speech, we observe distinct peaks and the position of these peaks in the spectra are dependent on the phoneme under consideration. The importance of formants is well know and in [7] the author has tried to use the location of spectral peaks as an additional feature in ASR. On the similar lines, the central idea in [4] while using multi-resolution spectral entropy as a feature was to capture the peaks of the spectrum and their location. To compute entropy of a spectrum we converted the spectrum into a PMF like function by normalizing it.

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

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). *Entropy* for each frame was computed from $\mathbf{x}$ by:

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

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 can be used as an estimate for speech/silence detection.

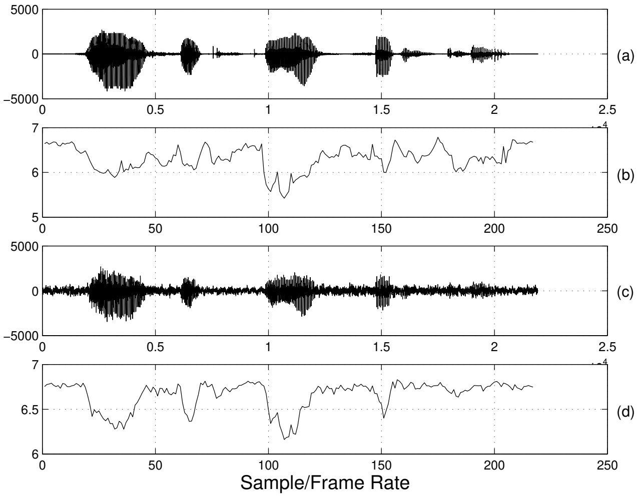

Figure 1: *Entropy computed from the full-band spectrum. (a) Clean speech wave form, (b) Entropy contour for clean speech, (c) Speech corrupted with factory noise at 6 dB SNR, and (d) Entropy contour for speech corrupted with factory noise at 6 dB SNR.*

In presence of noise, the formants are less affected as compared to the other parts of the spectrum. So we can 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 entropy contour is squeezed in presence of noise, it retains its discriminatory property. In [8], authors successfully used entropy for end point detection of speech in noisy environments.

2.2 Multi-band/Multi-resolution entropy

Realizing that full-band entropy can capture only the gross peakiness of the spectrum but not the location of the formants, in [4] we suggested the idea of multi-resolution/multi-band entropy feature to capture the location of the formants. To extract multi-band entropy features, we divided the full-band spectrum into **J** non-overlapping sub-bands of equal size. Entropy was computed for each sub-band and we obtained one entropy value for each sub-band. These sub-band entropy values indicate the presence or absence of formants in that sub-band.

2.3 Modifications

In [4] we advised that the way full-band spectrum is converted into a PMF, each sub-band spectrum should be converted into a sub-band PMF and entropy be computed for each sub-band PMF. Converting each sub-band into a PMF enhances the false local peaks as the normalization was done only for that sub-band. This false relative peak enhancement gives us low entropy values even for peaks which were smaller in the full-band but were observed as high relative peaks in a normalized sub-band.

To overcome this problem, we explored the option of not converting each sub-band into a PMF separately. Instead the full-band was converted into a PMF and then divided into sub-bands to obtain entropy feature from each sub-band. This ensured that the peaks retained their relative strength in each sub-band. This is equivalent to *computing the entropy contribution of each sub-band to the full-band entropy.* The remaining process of feature extraction was kept the same and is explained in the next paragraph.

When $\mathbf{J = 1}$, we work with the full-band spectrum and obtain one entropy value. When there are two sub-bands ($\mathbf{J = 2}$) we obtain two entropy values, one from each sub-band, and so on. In our experiments we changed the parameter **J** from **1** to **32** and obtained the entropy values from each sub-bands. Instead of working with 15 point PLP spectrum as reported in [4], in this paper we worked with 129 point STFT spectrum to have better frequency resolution. We haven't tried the possibility of smoothing the STFT spectrum to remove the pitch affects in this paper.

In this paper, we also explored the idea of overlapping sub-bands of unequal size. To accomplish this, we divided the full-band into 24 overlapping sub-bands defined by Mel-scale [1] and computed entropy from each sub-band. Moreover, in order to incorporate the temporal information, we also used the first and second order time derivatives of the multi-band entropy feature in the system.

The above listed modifications, namely, entropy contribution from each sub-band, overlapping sub-bands and time derivatives, impart robustness to the spectral entropy feature.

2.4 Entropy feature in Tandem framework

The simplicity of the hidden Markov model (HMM)/artificial neural network (ANN) hybrid system is that the features do not need special processing like decorrelation for developing a system as an ANN can learn the correlation among the features on its own. Moreover, an HMM/ANN system does discriminatory training and specially the output of the hybrid system is well suited for multi-stream combination. In contrast, though hidden Markov model (HMM)/Gaussian mixture model (GMM) system does likelihood based training, the advantage of HMM/GMM system is that modelling techniques like context dependent phone modelling and state-tying can be easily implemented in it.

In Tandem [5] approach, where the initial discriminatory modelling is done with the help of an ANN and then the output of the ANN is modelled with the help of HMM/GMM system, the advantages of both the systems can be exploited. In our third set of experiments, we used Tandem system to ascertain the importance of entropy feature when a) Used stand alone and b) Entropy feature is appended to the PLP cepstral features.

3 Experimental Setup

In the experiments reported in this paper, Numbers95 database of US English connected digits telephone speech [9] 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 either clean or corrupted by factory noise from Noisex92 database [10] added at different signal-to-noise-ratios (SNRs) to Numbers95 database. There were 3330 utterances for training and 2250 utterances were used for testing the system.

We have used HMM/ANN hybrid system [11] for the first two set of experiments. The ANNs used were a single layer multi-layer perceptron (MLP) 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 fixed state transition probabilities of 0.5. Each phoneme had a 1 state monophone model for which emission likelihoods were supplied as scaled posteriors [11]. The minimum duration for each phoneme is modelled by forcing 1 to 3 repetitions of the same state for each phoneme.

The new multi-band spectral entropy feature and its first and second order time derivatives were used to develop stand alone spectral entropy feature based hybrid and Tandem systems. Also, we ran experiments appending the entropy feature to the PLP feature in Tandem system.

The Tandem system was implemented with the basic hybrid system discussed above plus 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 GMM 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 linear output of the hybrid system, after being decorrelated with the help of principal component analysis (PCA) and were 27-dimensional. The implementation details of the Tandem system can be found in literature [5] and have not been described here.

4 Results

The results for HMM/ANN hybrid system, in terms of word-error-rates (WERs), of the entropy features alone are shown in Table 1. For example, '2-bands Entropy' feature is obtained by dividing

Word-Error-Rates for spectral entropy features	
Feature	**WER**
Full-band Entropy	91.6%
2-bands Entropy	74.4%
3-bands Entropy	59.5%
4-bands Entropy	42.7%
8-bands Entropy	24.3%
16-bands Entropy	18.6%
24-bands Entropy	16.2%
32-bands Entropy	15.1%
24 Mel-bands Entropy	15.7%

Table 1: *Word-Error-Rates (WERs) for clean speech for multi-band spectral entropy features in hybrid system for different number of sub-bands. Only* **Mel-bands are overlapping**. *Rest of the sub-bands are non-overlapping.*

the normalized full-band into two equal sub-bands and obtaining one entropy value from each sub-band. The two entropy values thus obtained are appended to form a 2-dimensional entropy feature vector used for training and testing the system. Entropy feature vectors are obtained for up to 32 equal-sized non-overlapping sub-bands. The results are shown in Table 1. To consider overlapping sub-bands, we used the 24 overlapping sub-bands of Mel-scale and the result are reported in the same table. WER results indicate as the number of sub-bands are increased, the performance improves. So going from full-band entropy feature to multi-band entropy feature pays rich dividends.

To study the multi-band spectral entropy feature further, we obtained its first and second order time derivatives and appended to the original feature. To observe the performance in presence of noise, we added factory noise from Noisex92 database at different SNRs to the speech signal. We observe (Table 2) that appending the time derivatives of the entropy feature to the entropy feature once again gives an improvement in the performance of the system. The overlapping Mel-scale sub-

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

bands give better performance as compared to non-overlapping sub-bands. Also, changing the number of parameters in the MLP didn't change the performance of the individual features considerably.

In Table 3 we have shown the results in terms of WERs for Tandem system. The results for entropy feature are shown only for the best case, that is for overlapping 24 Mel-bands. The combined feature performs better than the baseline under all conditions. When entropy feature is appended to the PLP feature, better improvements are observed in cases when difference between the performance

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%*

Table 3: *WERs for PLP feature, 24 Mel-band entropy feature and its time derivaties (24-Mel), and the two features appended (PLP + 24-Mel), in TANDEM system under different noise conditions. * indicates that the difference in performance is not significant.*

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 features is not significant.

5 Discussion and Conclusion

In search of new features having complementary information, this paper investigated the use of entropy of the spectrum as an additional feature. It has been shown that entropy of the full-band spectrum can be used as an estimate for speech/silence detection. In this paper, we suggested dividing the normalized spectrum into sub-bands and obtaining contribution to entropy from each sub-band and using that as a feature for ASR. The ideas of overlapping sub-bands and the time derivatives being appended to the feature improve the performance further, specially at low SNRs. Improved performance is obtained when multi-band entropy feature is appended to the usual PLP cepstral features under all conditions.

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. 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] P. Niyogi and M. M. Sondhi, "Detecting stop consonants in continuous speech," *J. Acoust. Soc. Amer.*, vol. 111, pp. 1063–1076, Feb. 2002.

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

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

[9] 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.

[10] 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.

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