IDIAP RESEARCH REPORT

Spectral entropy based feature for robust ASR

Hemant Misra [a] [b]   Shajith Ikbal [a] [b]
Hervé Bourlard [a] [b]   Hynek Hermansky [a]

IDIAP–RR 03-56

PUBLISHED IN
IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2004, Montreal, Canada

[a] IDIAP - Dalle Molle Institute of Perceptual Artificial Intelligence, P.O. Box 592, CH-1920 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 03-56

Spectral entropy based feature for robust ASR

Hemant Misra Shajith Ikbal Hervé Bourlard Hynek Hermansky

published in
IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2004, Montreal, Canada

Abstract. In general, entropy gives us a measure of the number of bits required to represent some information. When applied to probability mass function (PMF), entropy can also be used to measure the "peakiness" of a distribution. In this paper, we propose using the entropy of short time Fourier transform spectrum, normalised as PMF, as an additional feature for automatic speech recognition (ASR). It is indeed expected that a peaky spectrum, representation of clear formant structure in the case of voiced sounds, will have low entropy, while a flatter spectrum corresponding to non-speech or noisy regions will have higher entropy. Extending this reasoning further, we introduce the idea of multi-band/multi-resolution entropy feature where we divide the spectrum into equal size sub-bands and compute entropy in each sub-band. The results presented in this paper show that multi-band entropy features used in conjunction with normal cepstral features improve the performance of ASR system.

1 Introduction

Most of the state-of-the-art automatic speech recognition (ASR) systems use cepstral features derived from short time Fourier transform (STFT) spectrum of speech signal. The most common features used are MFCC [1], PLP [2] and RASTA [3] or some of their variants. Some recently proposed cepstral features like MCMS [4] and PAC [5] have also shown good performance. While cepstral features are fairly good representation, they capture the absolute energy response of the spectrum. Further, we are not sure that all the information present in the STFT spectrum is captured by them. In this paper we suggest to capture further information from the spectrum by computing its entropy.

Entropy plays a central role in information theory as a measure of information, choice and uncertainty (page 11 of [6]). The same entropy can be used to measure the "peakiness" of a spectrum if we convert the spectrum into a probability mass function (PMF). For voiced sounds, spectra have clear formants and entropies of such spectra will be low. On the other hand spectra of unvoiced sounds are flatter and their entropies should be higher. Therefore, entropy of a spectrum can be used as an estimate for voicing/unvoicing decision. In this paper, we extend the idea further and introduce multi-band/multi-resolution entropy feature. Depending upon the phoneme, entropy of the sub-bands where a formant is present will be low and the sub-bands which are flatter will have higher entropy. The important thing is even if the formant is slightly displaced from its position in noisy speech, its entropy will not be affected much. We expect this new feature to capture the "peakiness" of the spectrum and to be different from the usual cepstral features derived from spectral energies.

The remaining paper is arranged as follows: In the next section we introduce the spectral entropy feature and its computation for the present setup. In Section 3 we explain the database used and the experimental setup. Section 4 contains the results followed by conclusions in Section 5.

2 Spectral Entropy Feature

2.1 Motivation

Entropy can be used to capture the "peakiness" of a PMF. A PMF with sharp peaks will have low entropy while a PMF with flat distribution will have high entropy. For this reason entropy is generally associated with a classifier's output posteriors distribution and gives us a measure of the classifier's confidence. In previous work [7], it has been shown that entropy of the posteriors distribution at the output of a classifier can be used for weighting different streams in multi-stream combination.

In this paper, the aim is entirely different and we want to explore the peak capturing property of the entropy to capture peaks (also called as formants) of a spectrum. 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. These formants are the one which characterise a sound. The central idea while using entropy as a feature is to capture the peaks of the spectrum and their location. The problem with computing entropy of a spectrum is that spectrum is not a PMF (the area under the spectrum doesn't sum upto 1). In order to convert the spectrum into a PMF like function we divided the individual frequency components of the spectrum by sum of all the components.

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

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). This ensured that the area under the normalised spectra summed to 1 and this normalised spectra can be treated as a PMF for the purpose of computing entropy. For each frame the entropy was computed from $\mathbf{x}$ by:

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

Fig. 1(b) shows the contours of the entropy computed on the full-band spectrum. From the figure

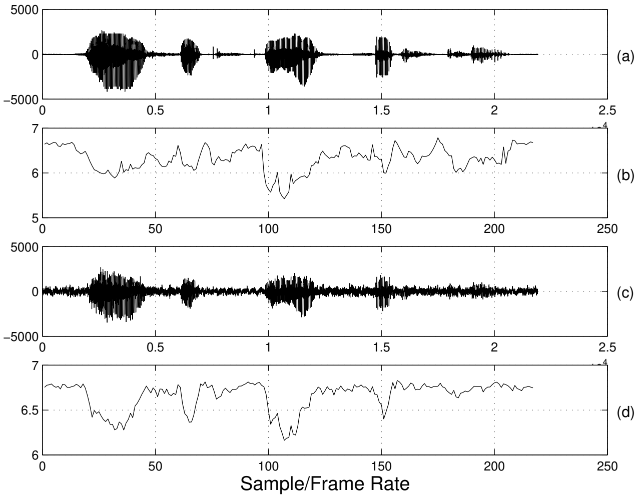

Figure 1: *Entropy computed from the full-band spectrum. (a) Clean speech wave form, (b) Entropy contour for clean speech, (c) Speech corrupted with factory noise at 6 dB SNR, and (d) Entropy contour for speech corrupted with factory noise at 6 dB SNR.*

we observe that entropy computed on full-band can be used as an estimate of voicing/unvoicing decision. Also, we know that in presence of noise the formants are the one which are least affected as compared to the other parts of the spectrum. So intuitively we can assume that entropy of the spectrum if used for voicing/unvoicing decision 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.

2.2 Multi-band/Multi-resolution entropy

We realized that entropy of the full-band spectrum is not a strong feature on its own if we want to capture the formants of the spectrum as well as their location. The reason for this is that entropy of the full-band spectrum cannot resolve the formants location as it captures only the gross peakiness of the spectrum.

To capture the location of the formants we introduced the idea of multi-band entropy features. To extract multi-band entropy features we divide the full-band spectrum into $\mathbf{J}$ non-overlapping sub-bands of equal size. Entropy is computed for each sub-band and we obtain one entropy value for each sub-band. These sub-band entropy values indicate the presence or absence of formants in that sub-band. The way full-band spectrum was converted into a PMF, each sub-band spectrum should be converted into a sub-band PMF. Using (1) and (2) we separately compute entropy for each sub-band PMF.

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 $\mathbf{J}$ from $\mathbf{1}$ to $\mathbf{5}$ and obtained the entropy value from each sub-bands. All the entropy values obtained by varying $\mathbf{J}$ were appended to form a $15(= 1+2+3+4+5)$-dimensional entropy feature vector.

Figure 2 shows the contours of 4^{th} and 5^{th} component of the entropy feature vector. Different components of the entropy feature vector have different dynamic ranges and have different activation points depending upon whether a formant is present in a particular sub-band or not.

Instead of working on the raw spectra, which contain pitch information also, we worked on spectra smoothed by filter bank.

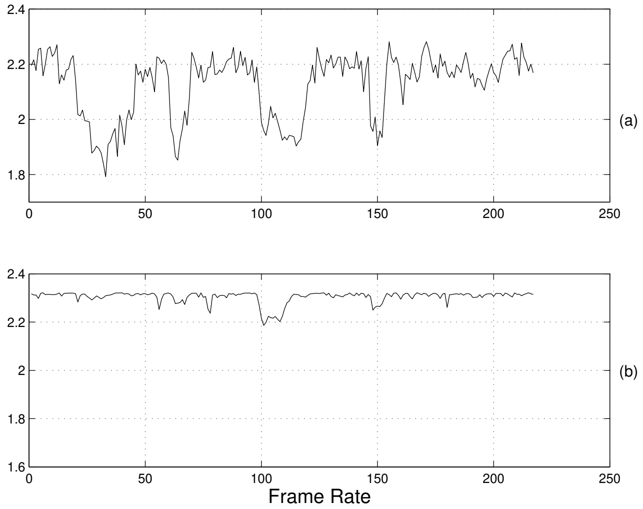

Figure 2: *(a) and (b) are contour of the* 4^{th} *and* 5^{th} *component of the entropy feature vector, respectively, for speech corrupted by 6 dB SNR factory noise.*

3 Experimental Setup

In the experiments reported in this paper, Numbers95 database of US English connected digits telephone speech [8] is used. There are 30 words in the database represented by 27 phonemes. Training is performed on clean speech utterances and testing data, which is different from the training data, is corrupted by factory noise from Noisex92 database [9] added at different signal-to-noise-ratios (SNRs) to Numbers95 database. We ran the baseline experiments using PLP [3] features. There were 3330 utterances for training and 1143 utterances were used for testing the system.

We have used Hidden Markov Model (HMM)/Artificial Neural Network (ANN) hybrid system [10] for performing the 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 feature vectors used in our system were: 12-dimensional raw cepstral coefficients (0^{th} coefficient is not used) appended with 13-dimensional delta and 13-dimensional delta-delta cepstral coefficients. 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 [10]. Many standard techniques like tri-phone modelling and state-tying, which are used in state-of-the-art GMM/HMM systems are not possible in HMM/ANN systems, but HMM/ANN systems don't need any conditioning of the new features as ANN learns the correlation on its own without the need of any fine tuning. The minimum duration for each phoneme is modelled by forcing 1 to 3 repetitions of the same state for each phoneme. *Phone deletion penalty* parameter was empirically optimised for clean speech test database and then it was kept constant for all the experiments.

4 Results

The results in terms of word-error-rates (WERs) of the entropy features alone are shown in Table 1. For example, 'Two-bands Entropy' feature is obtained by dividing the 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 upto 5 sub-bands and their results are shown in the table

Word-Error-Rates for entropy features alone		
Feature	Feature Dimension (J)	WER
Full-band Entropy	1	88.8%
Two-bands Entropy	2	68.4%
Three-bands Entropy	3	57.8%
Four-bands Entropy	4	54.4%
Five-bands Entropy	5	52.3%
Entropies Appended	**1+2+3+4+5=15**	**23.7%**

Table 1: *Word-Error-Rates (WERs) for clean speech for multi-band entropy features alone. The last row result is obtained when all the entropy features ($J = 1$ to 5) are appended to form a 15-dimensional entropy feature vector.*

(Table 1). WER results indicate as the number of sub-bands are increased, the performance improves and sort of starts levelling down. We stopped at 5 sub-bands to keep reasonable number of points in each sub-band for reliable entropy computation. So going from full-band entropy feature to multi-band entropy feature pays rich dividends.

The next experiment was to see how the system performs when these individual multi-band entropy feature vectors from different sub-bands are appended and a big multi-band entropy feature vector is formed. In Fig 3 we show the results when all the 15 entropy vectors obtained above (Table 1)

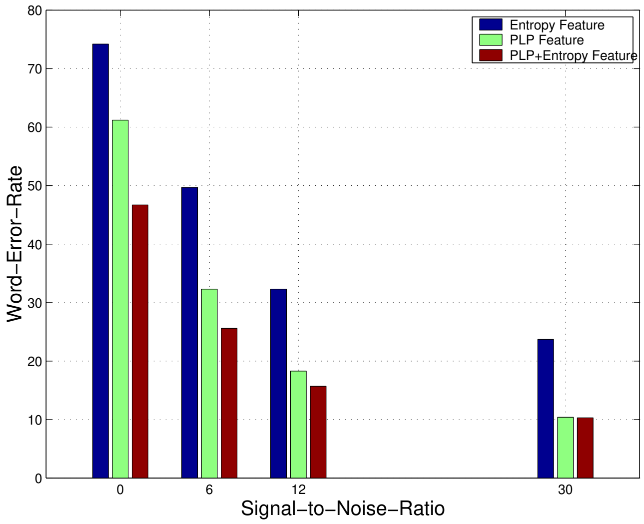

Figure 3: *WERs for Entropy features, PLP features and PLP features appended with ent ropy features at different SNRs. Clean speech is represented by 'SNR 30 dB'.*

are appended to form a 15-dimensional entropy feature vector. WER results are shown for clean and noisy speech. Noise used was factory noise from Noisex92 database added at different SNRs. The last row of Table 1 is same as the entropy feature performance for clean speech in Fig. 3. The entropy feature when used alone doesn't compete with the usual PLP cepstral features used for ASR but the good thing is that it doesn't degrade rapidly in presence of noise. Moreover, in presence of noise, when the entropy feature is appended to the PLP cepstral features, noticeable improvement in the performance of the system is observed. These multi-band entropy features improve the robustness of the baseline system. The relative improvement in performance is 1.0%, 14.2%, 20.7% and 23.7% for clean, SNR 12, SNR 6 and SNR 0, respectively. This result indicates that entropy feature brings more improvement as the noise level increases.

In the last, though the number of parameters of the MLP are more when higher dimensional feature vector (PLP + Entropy Feature = 53-dimension) is used, it has been verified through experiments that the performance of the individual features alone do not change considerably with the increase in number of parameters of their respective MLP models.

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 of voicing/unvoicing. Going one step further we suggested dividing the spectrum into equal sub-bands and obtaining entropy from each sub-band and using that as an additional feature for ASR. Good improvement in performance is obtained when multi-band entropy feature is appended to the usual PLP cepstral features, specially in case of noise. The new feature though doesn't compete with the cepstral features in absolute sense, it seems to be quite robust to noise. The reason for robustness can be attributed to the fact that multi-band entropy feature tries to capture the location of the formants and formant are less affected by noise.

The next step could be to divide the full-band into unequal parts depending upon the a-priori knowledge about the formants in the basic spectra obtained from the phonemes. Also, to automatically find the optimal sub-band boundaries every 10 ms based on minimising the combined entropies of the sub-bands holds promise.

ACKNOWLEDGEMENTS

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] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[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] V. Tyagi, I. McCowan, H. Bourlard, and H. Misra, "On factorizing spectral dynamics for robust speech recognition," in *Proceedings of European Conference on Speech Communication and Technology*, (Geneva, Switzerland), Sept. 2003.

[5] S. Ikbal, H. Misra, and H. Bourlard, "Phase autocorrelation (PAC) derived robust speech features," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, (Hong Kong), Apr. 2003.

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

[7] 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.

[8] 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.

[9] 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.

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