ENTROPY-BASED VARIABLE FRAME RATE ANALYSIS OF SPEECH SIGNALS AND ITS APPLICATION TO ASR

H. You, Q. Zhu and A. Alwan

Electrical Engineering Department, UCLA
Los Angeles CA90095, USA
hyou, qifeng, alwan@icsl.ucla.edu

ABSTRACT

Most speech processing algorithms analyze speech signals frame by frame with a fixed frame rate. Fixed-rate analysis is inconsistent with human speech perception and effectively assigns the same importance or 'weight' to all equi-duration frames. In [1], we proposed a variable frame rate (VFR) analysis technique that is based on a Euclidian distance measure. In this paper, we propose another approach for VFR based on the entropy of the signal. We compare entropy and Euclidian distance measures for VFR in ASR experiments using the Aurora2 and TI46 databases. Better performance is observed for the entropy-based VFR over our earlier VFR approach and over the fixed-rate system.

1. INTRODUCTION

Most speech-processing algorithms window the speech signal into frames and process them sequentially. These algorithms usually process signals with a fixed rate, which results in evenly sampled signals. Fixed-rate processing is inconsistent with speech perception[2].

Several studies have proposed a variable frame rate (VFR) approach to speech analysis. In [3], the Euclidian distance between neighboring frames is measured and compared to a threshold. Only frames with a distance greater than the threshold are retained. In [4], VFR analysis is based on the time derivative of the feature vectors. The Euclidian norm of the derivative feature is computed and compared with a threshold. The study in [1] further improves VFR in [3] by introducing a smaller frame shift, error accumulation, and energy weighting. It is found that the approach in [1] improves ASR of noisy speech. In [5], a comparison between the different VFR algorithms is carried out and results show that the method in [3] outperforms the derivative method[4] in reducing the number of features without a degradation in acoustic modelling. It also verifies that the technique in [1] improves acoustic modelling in noise and that [1] outperforms the other VFR techniques in recognition experiments.

This paper proposes an entropy-based approach for VFR. A speech signal's entropy curve is computed and used to design a frame-picking scheme. Compared with the Euclidian distance approach, our technique provides an information-theoretic framework for frame-picking since spectral changes have a direct connection to entropy. Other advantages of the proposed approach over that described in [1] include: Entropy-based VFR can achieve a more reliable frame-picking decision based on a longer time span spectral information, and is less sensitive to noise.

The remainder of the paper is organized as follows: Section 2 describes entropy computation, while Section 3 specifies the VFR algorithm. Speech recognition experiments are described in Section 4. Section 5 includes a summary and conclusions.

2. ENTROPY COMPUTATION

2.1. Entropy of a Gaussian Random Variable

Assume a random variable v of dimension K. The entropy of the random variable (RV) is computed by first estimating its probability distribution function (pdf). We can compute the pdf either from the RV's histogram or from a parameterized distribution. The latter is used to reduce the computation. Assume the pdf of v follows a K-dimensional Gaussian density,

$$p(v) = |2\pi\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(v-\mu)^T \Sigma^{-1}(v-\mu)} \quad (1)$$

where μ is a K-dimensional mean vector, Σ is a $K \times K$ covariance matrix. The entropy of v is

$$H(v) = -\int p(v) \ln p(v) dv \quad (2)$$

Let the eigen-decomposition of Σ be

$$\Sigma = A \Lambda A^T \quad (3)$$

where A is the unitary matrix of eigen-vectors and $\Lambda = diag[\lambda_1, ..., \lambda_K]$ is the diagonal matrix of eigenvalues. Be-

cause the RV Av has a diagonal covariance matrix, the entropy of Av is simplified as

$$H(Av) = K \ln \sqrt{2\pi} + \sum_{j=1}^{K} \ln \lambda_j \quad (4)$$

$H(v)$ equals $H(Av)$ when A is unitary, so $H(v)$ is expressed in terms of Σ's eigenvalues λ_j, $j = 1, ..., K$. Since $\prod_{j=1}^{K} \lambda_j$ is upper-bounded by $(\frac{\sum_{j=1}^{K} \lambda_j}{K})^K$, we approximate $\sum_{j=1}^{K} \ln \lambda_j$ with $\ln (\sum_{j=1}^{K} \lambda_j)$ and ignore constant difference terms. Thus the entropy of a Gaussian RV is computed from the diagonal elements of its covariance matrix as shown in Eq. 5. Although only on-diagonal elements are used in the entropy computation, we make no assumption of the covariance matrix being diagonal. The approximation also avoids the ill-posed problem when the RV's covariance matrix is not full rank.

$$\begin{aligned} H(v) &= K \ln \sqrt{2\pi} + \sum_{j=1}^{K} \ln \lambda_j \\ \Rightarrow H(v) &\approx K \ln \sqrt{2\pi} + \ln \mathbf{Tr}(\Lambda) \\ \Rightarrow H(v) &\approx K \ln \sqrt{2\pi} + \ln \mathbf{Tr}(\Sigma) \end{aligned} \quad (5)$$

2.2. Implementation

The entropy curve of a signal is computed as follows: First, the speech signal is processed into a feature sequence with a Hamming window of length 25 ms and a frame shift of 2.5 ms. Second, a 30 ms rectangular window is applied to the feature sequence. The signal's local entropy is computed as in Eq. 5, using all the features within the window. Finally, we re-compute the entropy curve every 15 ms over the feature sequence.

There are two time-domain parameters used in computing the entropy curve. The first parameter is the length of the rectangular window. Since speech signals are short-time stationary, a 30 ms window is used in our implementation in order to gather enough information when computing local entropy. When a frame shift of 2.5 ms is used in signal analysis, 12 feature vectors characterize each point in the entropy curve. The second parameter is the window shift parameter, 15 ms. Compared with measures computed frame-by-frame, entropy computation is carried out every 15 ms, which reduces the overall computational cost.

A speech signal can be represented by different features, such as MFCCs, or LPCCs. Different features emphasize different spectral characteristics of the signal. The entropy computed from different feature sequences will, hence, vary. In our entropy computation, we compare the entropy of 13-dimensional MFCC sequences with that of 23-dimensional Mel-filtered spectrum sequences. Figure 1 shows an example. Entropy computed with Mel-filtered spectrum has peaks in the transitional part of speech, smaller values for steady speech, and valleys in background noise. It characterizes the signal's frequency changes better than entropy computed with MFCC sequences. One explanation is that the entropy of an MFCC sequence is sensitive to low-level energy changes caused by logarithmic compression of the Mel-filtered spectrum. Figure 2

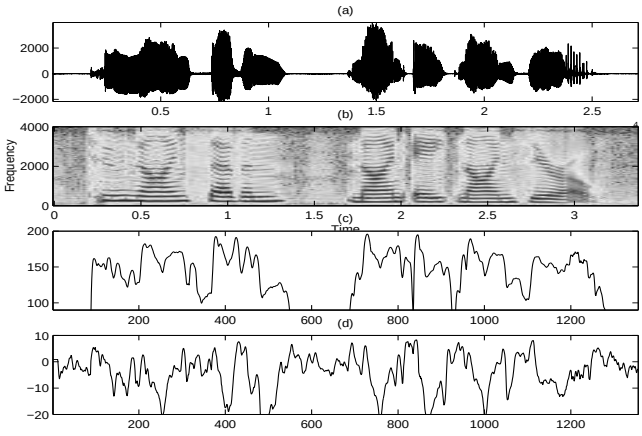

Fig. 1. For the digit sequence '2979890', we show the (a) waveform, (b) spectrogram, (c) entropy of the Mel-filtered spectrum sequence, and (d) entropy of the MFCC sequence.

shows another example comparing entropy measure with Euclidian distance and derivative measures. In this figure, all three measures are computed on a frame-by-frame basis. As shown, the Euclidian distance measure is very sensitive to frame-to-frame feature changes. These changes don't always reflect important spectral changes. The derivative measure doesn't provide detailed spectral change information, while emphasizing transitions between speech and non-speech segments. The entropy method captures spectral dynamic changes over longer time spans and emphasizes spectral transitional regions with high energy which can render it more noise robust than the other measures.

3. VARIABLE FRAME RATE ALGORITHM

Variable frame rate processing is carried out by comparing the signal's entropy to certain thresholds. In the current implementation of the entropy-based VFR algorithm, four frame-picking rates are selected: 5, 7.5, 10, and 12.5 ms. Hence, three frame-picking thresholds, $T1, T2, T3$, need to be optimized. In order for VFR to perform the desired frame picking scheme without prior knowledge of the SNR level, entropy thresholds need to be adjusted per utterance.

Given an entropy curve of a speech signal, $H(v_i)$ with $i = 1, ..., N$, frame-picking thresholds are set as in Eq. 6, using the maximum, median, and minimum entropy values of the entropy curve, represented as M_x, M_d, and M_n, re-

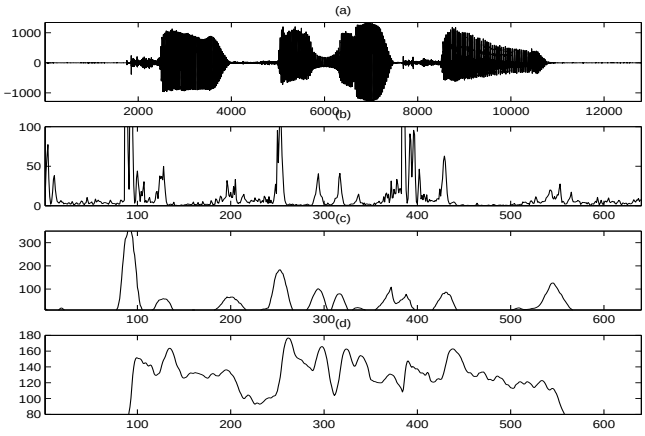

Fig. 2. For the digit sequence '272', we show the (a) waveform, (b) Euclidian distance measure, (c) derivative measure, and (d) entropy measure.

spectively.

$$\begin{cases} T_1 = w_1 M_x + (1 - w_1) M_d \\ T_2 = (1 - w_2) M_x + w_2 M_d \\ T_3 = (1 - w_3) M_d + w_3 M_n \end{cases} \quad (6)$$

Here w_1, w_2, and w_3 are weighting parameters of values 0.7, 0.8, 0.5, respectively. The frame-picking rate is set after comparing entropy with the above thresholds as in Eq. 7. This way, an average frame rate of approximately 7.5 ms is maintained.

$$\text{Frame Picking Rate} = \begin{cases} 5\ ms & \text{if } H(v_i) \geq T_1 \\ 7.5\ ms & \text{if } T_1 > H(v_i) \geq T_2 \\ 10\ ms & \text{if } T_2 > H(v_i) \geq T_3 \\ 12.5\ ms & \text{otherwise} \end{cases} \quad (7)$$

For the Aurora2 database, end point detection is employed

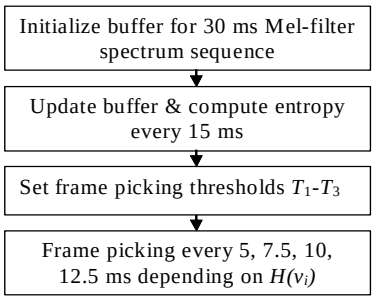

Fig. 3. Algorithm flow-chart for speech segments

first and a 20 ms frame shift is applied to non-speech segments. The framework of entropy-based VFR is summarized is Figure 3. An example of frame-picking using entropy-based VFR is shown in Figure 4.

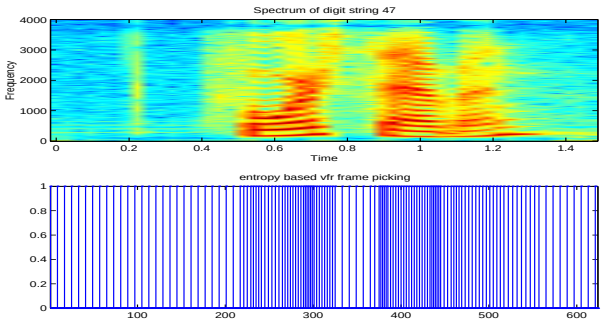

Fig. 4. Upper panel: The Spectrogram of '47', Lower panel: The frame-picking grid of entropy-based VFR in terms of the frame index.

4. RECOGNITION EXPERIMENTS

The performance of entropy-based VFR is tested on the Aurora2 and TI46 databases using HTK 3.0. The feature vectors consist of MFCC features (MFCC_E_D_A) of dimension 39. For the Aurora2 database, 18 states and 3 mixtures per state word HMMs are used. Mismatched ASR conditions are tested. The performance of the entropy-based VFR is compared with that of our previous approach[1] and the MFCC baseline. The same end point detection algorithm is applied to both VFRs, and the average frame rate for both is controlled to be 7.5ms.

Tables 1 and 2 show the average recognition accuracy of the MFCC baseline, Euclidian distance VFR[1], and entropy-based VFR over sets A and B, respectively. VFR

Table 1. Aurora2 Recognition Accuracy(%) (Set A Ave)

SNR	MFCC	VFR[1]	Rel	Entropy	Rel
clean	**99**	**98.58**	**-42**	**98.11**	**-89**
20dB	**95.25**	**95.94**	**14.53**	**96.11**	**18.11**
15dB	**87.33**	**91.57**	**33.46**	**92.57**	**41.36**
10dB	**67.7**	**79.85**	**37.62**	**82**	**44.27**
5dB	**39.47**	**56.23**	**27.69**	**59.92**	**33.78**
0dB	**16.95**	**26.41**	**11.39**	**27.1**	**12.22**
Ave 0-20	**61.34**	**70**	**24.94**	**71.54**	**29.95**

Rel: Relative word error rate reduction over MFCC baseline

Table 2. Aurora2 Recognition Accuracy(%)(Set B Ave)

SNR	MFCC	VFR[1]	Rel	Entropy	Rel
clean	**99**	**98.58**	**-42**	**98.13**	**-87**
20dB	**92.77**	**96.08**	**45.78**	**96.42**	**50.48**
15dB	**81.33**	**92.12**	**57.79**	**93.1**	**63.04**
10dB	**59**	**80.57**	**52.61**	**82.76**	**57.95**
5dB	**31.92**	**56.43**	**36**	**58.68**	**39.31**
0dB	**13.69**	**25.89**	**14.14**	**28.57**	**17.24**
Ave 0-20	**55.74**	**70.22**	**41.26**	**71.91**	**45.6**

Table 3. Entropy-based VFR Performance(%) of Mismatched Aurora2 Test

	Set A						Set B					
SNR	**Subway**	**Babble**	**Car**	**Exhibit**	**Ave**	**Rel**	**Restau.**	**Street**	**Airport**	**Train**	**Ave**	**Rel**
clean	**98.34**	**98.22**	**97.94**	**97.96**	**98.11**	**-92.86**	**98.34**	**98.28**	**97.94**	**97.96**	**98.13**	**-90.82**
20dB	**96.44**	**96.16**	**96.15**	**95.68**	**96.11**	**18.11**	**96.47**	**96.52**	**96.3**	**96.39**	**96.42**	**50.48**
15dB	**92.48**	**92.99**	**92.72**	**92.1**	**92.57**	**41.36**	**93.86**	**93.17**	**92.96**	**92.41**	**93.10**	**63.04**
10dB	**82.25**	**83.1**	**81.36**	**81.3**	**82**	**44.27**	**83.6**	**82.62**	**82.91**	**81.92**	**82.76**	**57.95**
5dB	**60.36**	**62.79**	**56.76**	**59.8**	**59.92**	**36.91**	**57.02**	**60.55**	**60.57**	**56.59**	**58.68**	**39.31**
0dB	**27.79**	**30.65**	**23.56**	**26.41**	**27.1**	**12.22**	**25.42**	**31.74**	**31.79**	**25.36**	**28.57**	**17.24**
-5dB	**9.15**	**9.4**	**8.11**	**8.15**	**8.70**	**0.84**	**8.17**	**10.91**	**10.77**	**7.16**	**9.25**	**1.73**
Ave 0-20	**71.87**	**73.14**	**70.11**	**71.02**	**71.54**	**30.57**	**71.27**	**72.92**	**72.91**	**70.54**	**71.91**	**45.61**

improves the MFCC baseline performance, especially in noise. Compared with our previous VFR technique[1], the entropy-based VFR further reduces the relative word error rate by 4-5% for SNR levels between 0-20dB. The improvement is benefited from using an entropy measure instead of the Euclidian distance measure. In Table 3, detailed results of entropy-based VFR are shown.

As shown in Tables 1 and 2, the entropy-based VFR degrades the MFCC baseline performance in the clean condition. This is caused by the VFR parameter setting. If VFR parameters are optimized for the clean condition, the entropy-based VFR can improve performance over the baseline, while the improvement under noise is less than that of the current setting. In order to optimize VFR performance for all conditions, an SNR-adaptive parameter setting needs to be used.

For the TI46 alphabet database, 8 states and 3 mixtures per state word HMMs are used. The entropy-based VFR is more efficiently implemented here than for the Aurora2 database, since the frame-picking thresholds can be optimized beforehand and fixed for all the utterances as opposed to being SNR dependent. The recognition accuracy of the MFCC baseline, Euclidian distance VFR[1], and entropy-based VFR is 89.6%, 91.3%, and 92.6%, respectively. The entropy-based VFR achieves a relative word error rate reduction of 14.94% over the Euclidian distance VFR in [1].

5. SUMMARY

In this paper we propose an entropy-based VFR approach. Assuming Gaussian distribution of the RVs generating front-end features, we derive an efficient way to compute the signal's entropy. The entropy of Mel-filtered spectrum sequence is found to be advantageous over that of MFCC sequence for improved acoustic modelling. The advantages of the entropy measure over previous distance measures include: first, entropy provides enough detailed spectral change information, while being less sensitive to frame-to-frame spectral changes; second, it is less sensitive to noise. More importantly, entropy-based VFR provides an information-theoretic framework for adaptive frame-picking. Recognition experiments are implemented on two databases: the Aurora2 and TI46 databases. When compared to our previous VFR technique, entropy-based VFR achieves better performance.

For future work, the effect of noise on entropy measurements, and the SNR-adaptive parameter setting will be studied.

6. ACKNOWLEDGEMENT

This work is supported in part by the NSF, and by ST Microelectronics and the state of CA through the UC Micro Program.

7. REFERENCES

[1] Qifeng Zhu and Abeer Alwan, "On the use of variable frame rate analysis in speech recognition," *ICASSP*, pp. 3264–3267, 2000.

[2] James J. Hant and Abeer Alwan, "A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise," *Speech Communication*, vol. 40, pp. 291–313, May 2003.

[3] K.M. Pointing and S.M. Peeling, "The use of variable frame rate analysis in speech recognition," *Computer Speech and Language*, vol. 5, no. 2, pp. 169–179, April 1991.

[4] Philippe Le Cerf and Dirk Van Compernolle, "A new variable frame rate analysis method for speech recognition," *IEEE Signal Processing Letter*, vol. 1, no. 12, pp. 185–187, December 1994.

[5] J. Macías-Guarasa, J. Ordónez, et al., "Revisiting scenarios and methods for variable frame rate analysis in automatic speech recognition," *Eurospeech*, pp. 1809–1812, 2003.