ENTROPY RATE-BASED STATIONARY / NON-STATIONARY SEGMENTATION OF SPEECH

Wolfgang Wokurek

Institute of Natural Language Processing, University of Stuttgart, Germany
wokurek@ims.uni-stuttgart.de, http://www.ims-stuttgart.de/~wokurek

Abstract

This study evaluates the potential of the entropy rate contour to identify stationary and non-stationary segments of speech signals. The segmentation produced by an entropy rate-based method is compared to the manual phoneme segmentations of the TIMIT and the KIEL corpora. Characteristic points, i.e. steepest rises and falls of the entropy rate curve and its maxima and minima are investigated to determine whether they label stationary and non-stationary speech segments. The phonetically labelled speech corpora for American English (TIMIT) and German (Kiel Corpus of Read Speech) serve as references for the corpus-based evaluation.

1. Introduction

Entropy is a quantitative measure of how uncertain the outcome of a random experiment is. Its definition and interpretation was introduced by C. E. Shannon (1948) and N. Wiener (1961) and is an integral part of textbooks on statistics, e.g. Papoulis (1984). Two examples may illustrate the uncertainty concept of entropy.

Throwing a fair die is a random experiment with a high uncertainty of outcome, because each side is equally likely to appear. Hence the entropy is high. If, on the other hand, a random experiment is biased and yields the same outcome most of the time, uncertainty and entropy are low. In particular, the entropy H can be calculated from the probabilities p_i of the n outcomes by $H(p_1,...,p_n) = \Sigma_{i=1}^{n} p_i \log p_i$.

This idea of measuring the uncertainty of a random experiment was extended to a discrete time random signal (i.e. a stationary, discrete time stochastic process) by

Phonus 5, Institute of Phonetics, University of the Saarland, 2000, 59-71.

means of the notion of prediction. What will be the average uncertainty of the next sample, if all past samples are known? The entropy rate of the stochastic process is the answer to that question and the random experiment simply consists in observing the next sample. Depending on the particular properties of the observed stochastic process the uncertainty of the next sample may be influenced or not by the knowledge of the past samples. Hence, the properties of the process will result in a high or low uncertainty of the next one. An example for high uncertainty of the next sample is white noise, where it is impossible to gain insight into the next sample, regardless of how many past samples are already known. Contrary to this, a low-pass filtered noise leaves much less uncertainty for the next sample, even if only the current sample is known.

In the context of speech segmentation the distinction between stationary and non-stationary regions is important. Entropy rate is a candidate to fulfill that purpose due to its sensitivity to statistical uncertainty.

In this paper entropy rate is applied to the task of segmenting speech signals. Section 2 summarizes the estimation of the entropy rate based on spectral estimation techniques. In Section 3 the characteristic points of such an entropy rate estimate contour are discussed with respect to a speech signal. Finally, in Section 4, the relation of the characteristic points of the entropy rate estimate contour to phoneme boundaries and acoustic landmarks (Liu, 1996) is investigated.

2. Entropy rate estimation

The entropy rate $H(\mathsf{x})$ of a stationary, discrete time stochastic process x may be defined equivalently in two different ways (Papoulis, 1984). The first definition formalizes the approach in the introduction. The uncertainty of the unknown next sample x_n assuming the m known past samples $\mathsf{x}_{n-1},...,\mathsf{x}_{n-m}$, is expressed by the conditional entropy $H(\mathsf{x}_n \mid \mathsf{x}_{n-1},...,\mathsf{x}_{n-m})$. The assumption that all past samples are known is expressed by the limit

$$H_c = \lim_{m\to\infty} H(\mathsf{x}_n \mid \mathsf{x}_{n-1},...,\mathsf{x}_{n-m}).$$

The second way of defining the entropy rate $H(\mathsf{x})$ starts with the average uncertainty of a block of m consecutive samples $\mathsf{x}_1,...,\mathsf{x}_m$. Using the joint entropy $H(\mathsf{x}_1,...,\mathsf{x}_m)$, that short term average uncertainty is $\frac{H(\mathsf{x}_1,...,\mathsf{x}_m)}{m}$. Again, the average uncertainty includes all samples and requires the limit

$$\overline{H}(\mathsf{x}) = \lim_{m\to\infty} \frac{1}{m} H(\mathsf{x}_1,...,\mathsf{x}_m).$$

Both definitions result in the same quantity $H(\mathrm{x}) = H_c(\mathrm{x}) = \overline{H}(\mathrm{x})$ which constitutes the entropy rate.

The calculation of both the conditional entropy and the joint entropy requires knowledge of the appropriate probability density functions. Estimation of joint densities is a data-intensive as well as computationally intensive task. Segmentation of speech on such a basis has been proposed in Peterka (1998). It uses the mutual information function and results in a sequence of short term stationary segments. In contrast to this, a spectral approach is used here. In particular if the stochastic process is normal (i.e. Gaussian), the entropy rate results from its power spectrum[1] $S_\mathrm{x}(\Theta)$:

$$\overline{H}(\mathrm{x}) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log S_\mathrm{x}(\Theta)\, d\,\Theta + \log \sqrt{2\pi\, e}, \tag{1}$$

where log is the logarithm to the basis e. That approach requires the knowledge of – or estimates of – the power spectrum $S_\mathrm{x}(\Theta)$ and not the probability density $p_{\mathrm{x}_1,\ldots,\mathrm{x}_m}\,(\mathrm{x}_1,\ldots,\mathrm{x}_m)$ or its estimates.

In fact speech signals are neither stationary stochastic processes nor normally distributed. Hence, it is not justified to apply both the concept of the entropy rate and Eq. 1. However, similar arguments hold true for linear prediction techniques (LPC etc.), which have been successfully applied to speech signals. It may be argued that the speech signal can be approximated locally by a stationary, normal stochastic process. On that local basis, spectral estimation techniques are employed to determine the power spectrum $S_\mathrm{x}(\Theta)$.

In contrast to current probability density based methods the spectrum based method allows computationally efficient implementations. Averaging magnitude spectra computed by the fast fourier transform (FFT) results in a slightly bigger computational load than a spectrogram.

The speech signal is processed using overlapping frames of width W. The signal of each frame is multiplied by a Hanning window $w(n)$ and transformed to the complex-valued short term spectrum

$$\mathrm{X}(m,k) = \sum_{n=0} x([m-1]\,\frac{W}{2} + n)\; w(n) e^{\mathrm{j}2\pi\mathrm{nk/N}} \tag{2}$$

[1] Note that frequency bands with vanishing power make the integral divergent. Such spectra are henceforth excluded.

where *k* is the discrete frequency index and *m* is the frame number. By selecting $N = 2^i$ as an integer power of two a radix two fast fourier transform (FFT) is applicable. Squaring the spectral magnitudes and averaging sequences of *V* consecutive spectra using a Hamming window *v*(*n*) results in a sequence of power spectral estimates

$$S(m,k) = \sum_{l=-\frac{V}{2}}^{\frac{V}{2}} v(\mathrm{l}) \mid X\,(m - l,\, k) \mid^2 \qquad \begin{matrix} m = 0, 1, 2, ... \\ k = 0, 1, ... , N - 1 \end{matrix} \tag{3}$$

Again *m* is the frame number and *k* the discrete frequency. The entropy rate estimate of the *m*-th frame *H*(*m*) is the sum of the logarithms of those power spectral estimates

$$H(m) = \frac{1}{N}\sum_{k=0} \log S(m,\, k). \tag{4}$$

The shape of the entropy rate contour is unaffected by linear filtering of the speech signal *x*(*n*). The effect of such a filter operation is an additive constant. Any reduction of spectral amplitude will result in rate reduction of the entropy. Similarly, mere signal amplification will increase the entropy rate. The gain factor α will add log α to *H*(*m*). To remove this influence of signal amplitude from the entropy rate, a scaled entropy rate *h*(*m*) is proposed which is defined as the subtraction of the logarithm of the energy

$$E(m) = \log\, \frac{1}{N}\sum_{k=0} S(m,\, k) \tag{5}$$

from the entropy rate *H*(*m*),

$$h(m) = H(m) - E(m). \tag{6}$$

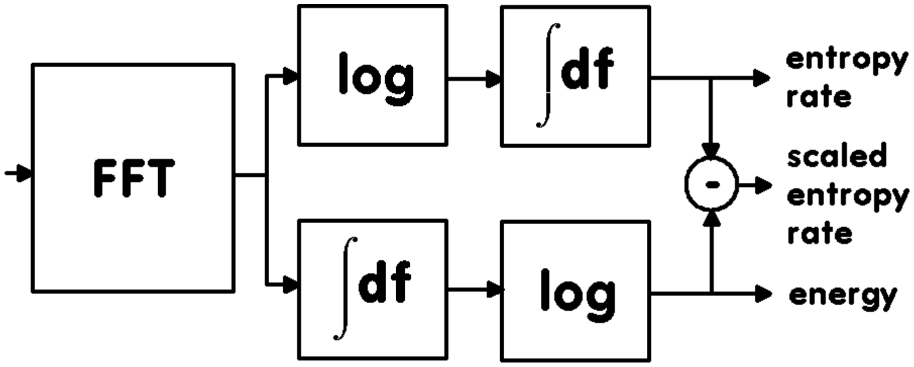

Figure 1. Block diagram of entropy rate estimation

Figure 1 shows a (simplified) block diagram of this algorithm. Entropy rate and energy just interchange the order of logarithm and integration. The scaled entropy rate constitutes the difference between them. The block diagram does not show the multiplication by the pre-FFT window and the post-FFT magnitude squaring and averaging. These operations are part of the FFT box.

Figure 2 shows the entropy rate analysis of a speech signal[2] from the TIMIT database (TIMIT). The sentence uttered by a male speaker is: *Cement is measured in cubic yards*. The displays below the speech waveform show (i) the logarithm of the short term energy Eq. 5, the entropy rate Eq. 4, (ii) the scaled entropy rate Eq. 6 and (iii) the TIMIT phoneme segmentation. The frame size of the analysis is 10 milliseconds. At first glance the entropy rate contour of a speech signal might look similar to a short time energy contour on a logarithmic scale. However, there are important differences concerning sounds of weak intensity, e.g. fricatives and stop consonants. They are represented at the very same entropy rate level as vowels and

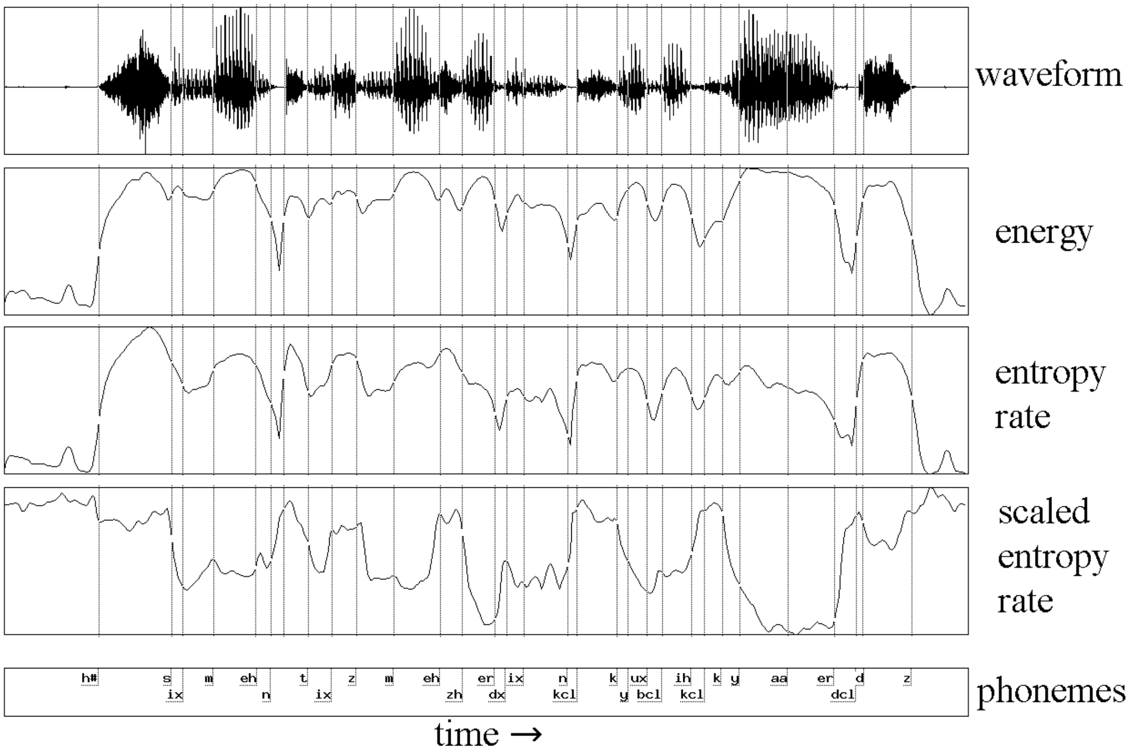

Figure 2. TIMIT example: Cement is measured in cubic yards.

[2] train/dr7/mgaw0/sx85.wav

often show a higher peak. This effect can be observed for the phonemes labelled /s/, /t/, /zh/[3] and the first /k/. The high peaks of these sounds are also observable in the scaled entropy rate contour. Note that no preemphasis high pass filter is used here. Such a filter would not change the shape of the entropy rate contour either, since it results in an additive constant only. This section is concluded with some remarks on the interpretation of entropy rate and scaled entropy rate. First, the entropy rate $H(m)$ can be interpreted as the time marginal of a (smoothed) spectrogram with logarithmic scale. Usually the spectrogram amplitudes are displayed or plotted using a decibel scale, but they are exactly the values log $S(m,k)$ being added in Eq. 4 and resulting in $H(m)$ (aside from the constant factor $\frac{1}{N}$). Hence, viewing the spectrogram and adding the decibel readings at any time results in the entropy rate, not in the energy.

Second, the scaled entropy rate $h(m)$ may be interpreted as a deviation of the spectral shape from the constant (white) spectrum. In that case, e.g. in speech pauses, the scaled entropy rate approaches its maximum value of zero. The scaled entropy rate is a negative quantity, indicating the degree of deviation of the local spectrum from the white spectrum. The more numerous and the lower the spectral valleys, the lower the scaled entropy rate.

3. Characteristic points

To derive a segmentation from the entropy rate contour, the characteristic points of that contour are analyzed. The characteristic points considered here are the maxima and minima of the (scaled) entropy rate contour and the locations of its steepest rising and falling slopes. Figure 3 shows a sample contour and its characteristic points. The maxima and the minima are detected by sign changes of the first difference contour (i.e. the discrete equivalent to the first derivative of a continuous signal). A change of sign from positive to negative indicates a maximum, the reverse process a minimum. This criterion fails for constant plateaus, which are extremely unlikely in most kinds of real world signals, particularly in speech signals. The locations of steepest rising and falling slopes are detected by sign changes of the second difference contour. The steepest rise is indicated by a positive first difference and a change of sign from positive to negative second difference.

[3] TIMIT transcription.

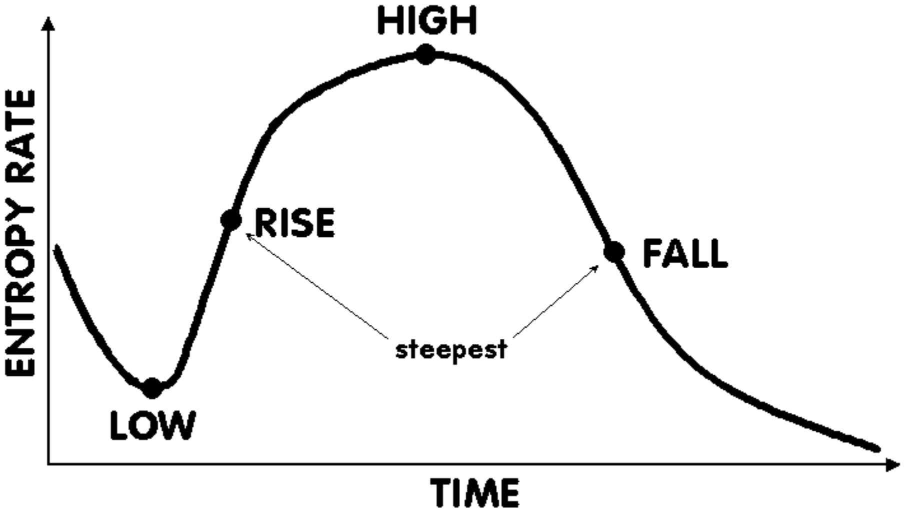

Figure 3. Characteristic points

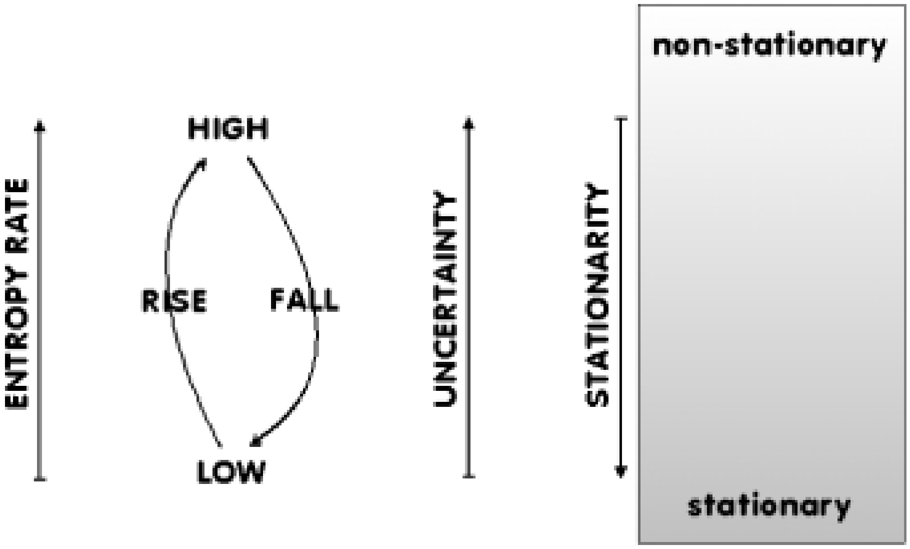

Figure 4. Interpretation of the characteristic points of the scaled entropy rate

Figure 4 summarizes the interpretation of the scaled entropy rate as far as it follows from statistics. HIGH labels correspond to the locally 'whitest' spectra and LOW labels indicate instants of the locally deepest and widest spectral valleys[4]. RISE and FALL report the instants of maximum rate of change of the spectral structure in-between.

[4] Note that the existence of valleys requires some sort of peak between them.

This rigid statistic interpretation is carried over to the terms of stationarity and non-stationarity. Within that framework white noise is considered as a non-stationary signal, which is true from the point of view of predicting the next sample as well as from listening experience. On the other hand, narrow band noise would be an example of a stationary signal. This type of noise is perceived as a tone with a random amplitude modulation for bandwidths of 1 Hz and below. Besides tones (hence formant oscillations) the harmonic structure of voiced sounds will be labelled `stationary' due to low values of the scaled entropy rate.

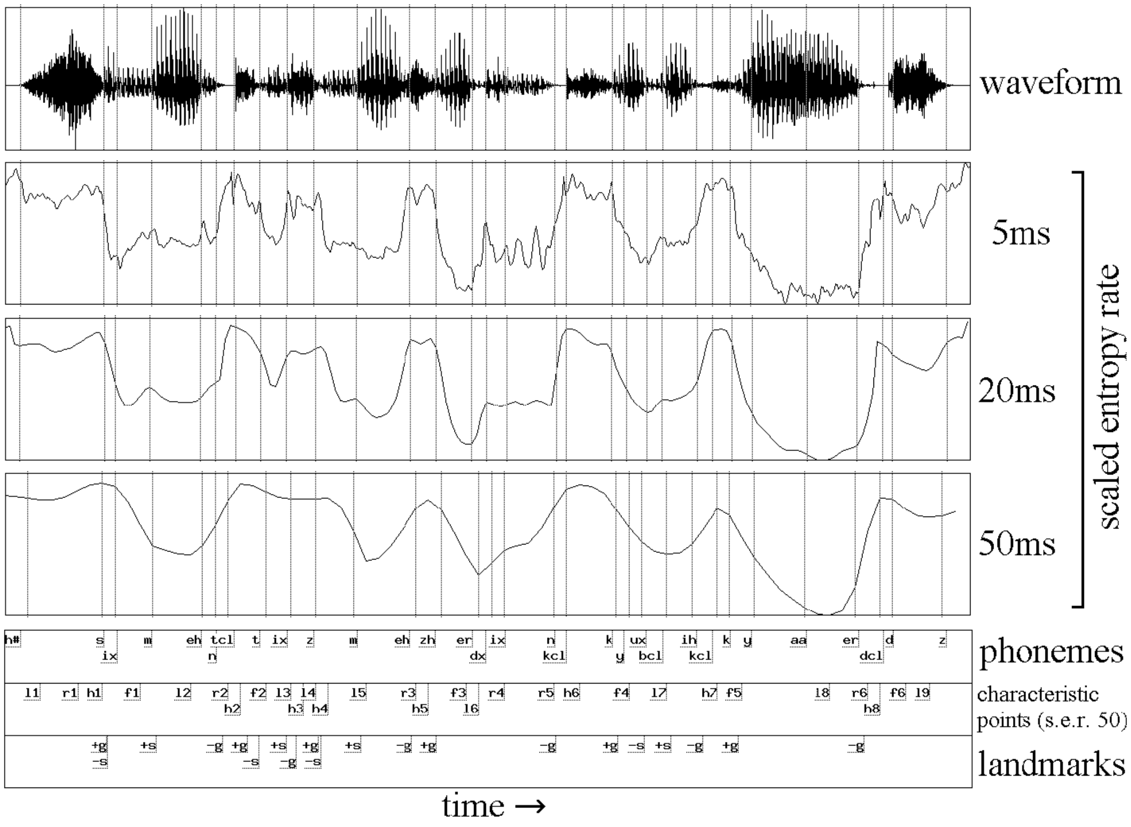

Figure 5. Scaled entropy rate with windows of different duration

Figure 5 depicts the same speech signal as Fig. 2 along with the scaled entropy rate contours for three different frame widths (5ms, 20ms, 50ms). This allows drawing a comparison between the scaled entropy rates of different frame widths. The TIMIT phoneme labels are extended over all scaled entropy rate signal displays. The TIMIT codes of the phonemes are shown in the first label tier. The second label tier corresponds to the characteristic points of the scaled entropy rate with a frame width of 50ms. The third label tier contains the automatically detected acoustic landmarks (Liu, 1996). Currently only two landmark types are implemented, namely the glottal landmark (g) and the sonorant landmark (s). They always come in pairs e.g. a glottal segment extends from +g to -g. Low valleys in the entropy rate contour often correspond to acoustic landmarks of the sonorant type.

The entropy rate contours are being smoothed with increasing frame width and lose their fine structure. Thus, the number of characteristic points decreases with increasing frame width. The characteristic points shown in Figure 5 are derived from the slowest varying contour to limit the number of labels in the track.

A second influence of the frame width on the scaled entropy rate comes from the frequency resolution. As the long term spectrogram displays the harmonics of the fundamental frequency, the spectral structure becomes `visible' to the scaled entropy rate with increasing frame width.

The interpretation of the characteristic points of the scaled entropy rate are examined in Fig. 5:

- **maxima:** *high uncertainty about the next sample:* transitions, non-stationary segments, release transient of stop consonants, phoneme boundary (h1: /s/ → /ix/, h2: /tcl/ → /t/, h3: /ix/ → /z/, h4: /z/ → /m/, h5: /zh/ → /r/, h6: /kcl/ → /k/, h7: /kcl/ → /k/, h8: /dcl/ → /d/)
- **minima:** *low uncertainty about the next sample:* stationary segments, vowels, speech pauses often have a sharp valley (l2: /eh/, l5: /eh/, l7: /bcl/, l8: /er/, l9: /z/), segments marked with the sonorant landmark often contain a deep and wide valley (l2, l6, l8)
- **steepest rises:** *transition from stationarity to non-stationarity:* phoneme boundary (r2: /n/ → /tcl/, r3: /eh/ → /zh/, r4: /ix/ → /n/, r5: /n/ → /kcl/, r6: /er/ → /dcl/)
- **steepest falls:** *transition from non-stationarity to stationarity:* phoneme boundary (f2: /t/ → /ix/, f4: /y/ → /ux/, f6: /d/ → /z/)

4. Corpus evaluation

So far the scaled entropy rate is a time-dependent quantity indicating the amount of uncertainty about a sample, given the statistical properties of the stochastic process. The scaled entropy rate is estimated by averaging the short term spectrum. The frame width or window duration is an important parameter defining the frequency resolution of the short term spectrum and controlling which spectral structures (e.g. harmonics) are 'visible' to the scaled entropy rate estimate. In the previous section the so-called characteristic points were used to facilitate a comparison of the scaled entropy rate contour with a phoneme or other segmentation. This comparison is performed automatically for entire corpora. The main questions answered in this section are (i) what window duration entails which density of characteristic points and (ii) how the characteristic points are distributed within the phonemes.

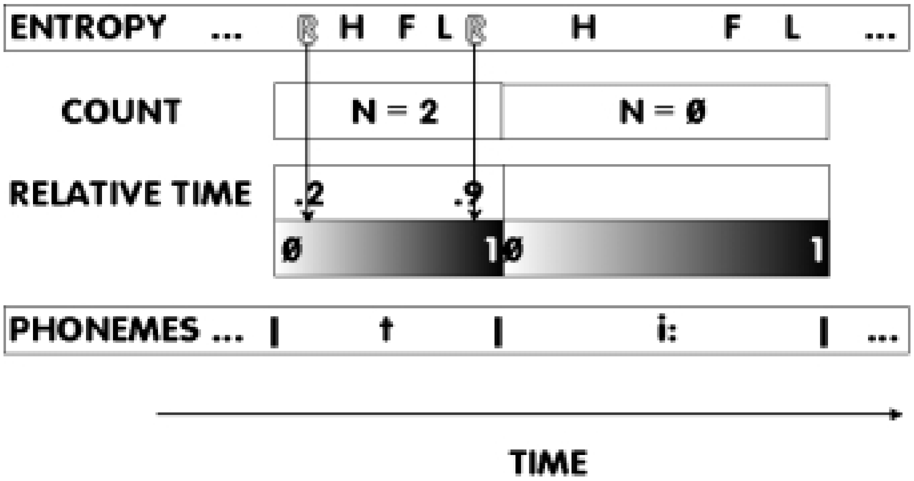

Figure 6. Comparing two segmentations

Figure 6 illustrates the technique of comparing the entropy rate label set to the phoneme labels. The phoneme segments are considered a sequence of reference intervals. Within each reference intervals the number and the relative positions of the characteristic points are recorded. The average number of characteristic points answers question (i). The mean is separately calculated for each type of characteristic point (h,l,r,f). The results for different window durations are plotted[5] in Figure 7. The means and standard deviations of the number of characteristic points per phoneme are given for both the entropy rate *H* (left column) and the scaled entropy rate *h* (right column). The frame duration *w* is listed in milliseconds (5, 10, 20, 50, 100, 200). The resulting density of characteristic points starts at 4 highs, lows, etc. per phoneme and goes down to 0.1 per phoneme, i.e. one label of each type occurs every 10 phonemes.

[5] Cf. Wokurek (1999) for a more detailed representation of this data in tabular form.

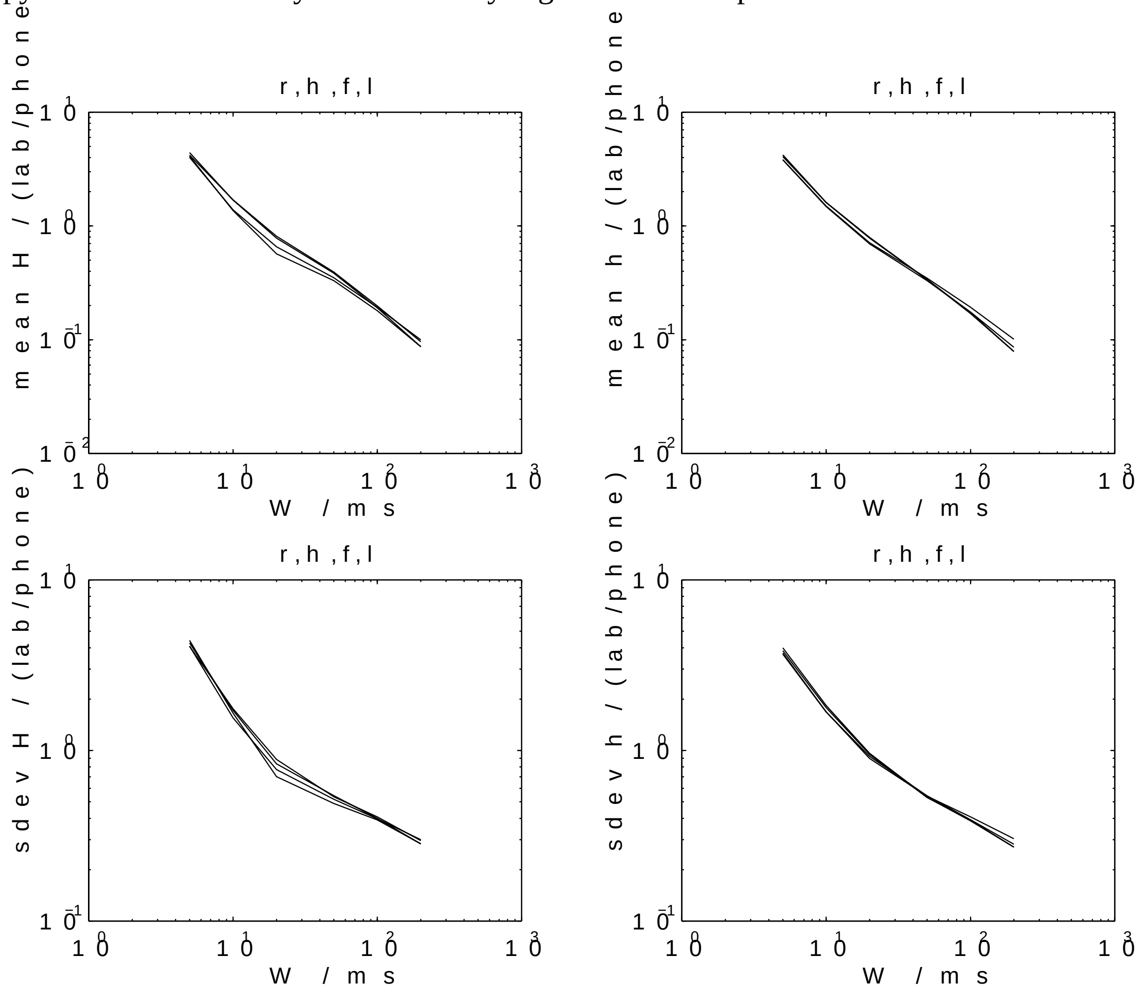

Figure 7. Entropy rate label density vs. window duration

The comparison between the entropy labels and the phoneme labels is made using the American 'TIMIT' database (TIMIT) and the German 'Kiel Corpus of Read Speech' (Kohler, 1992), which are phonetically labelled speech corpora. While Figure 7 shows the results for the TIMIT corpus, the results for the Kiel Corpus are essentially the same. Thus, the density of the characteristic points of the entropy rate contours seems to be language independent. To achieve about one high – fall – low – rise cycle per phoneme a window of about 15ms duration for both the entropy rate and the scaled entropy rate is appropriate.

How are relative positions of the characteristic points distributed within the phonemes? The answer to that question depends on the phoneme class. Figure 8 shows the results for the voiceless fricatives. Two histograms are shown for each type of characteristic point. The upper one displays the label count per phoneme. The window duration is 20ms and the peak indicates a modal value of one label per phoneme for all types of characteristic points.

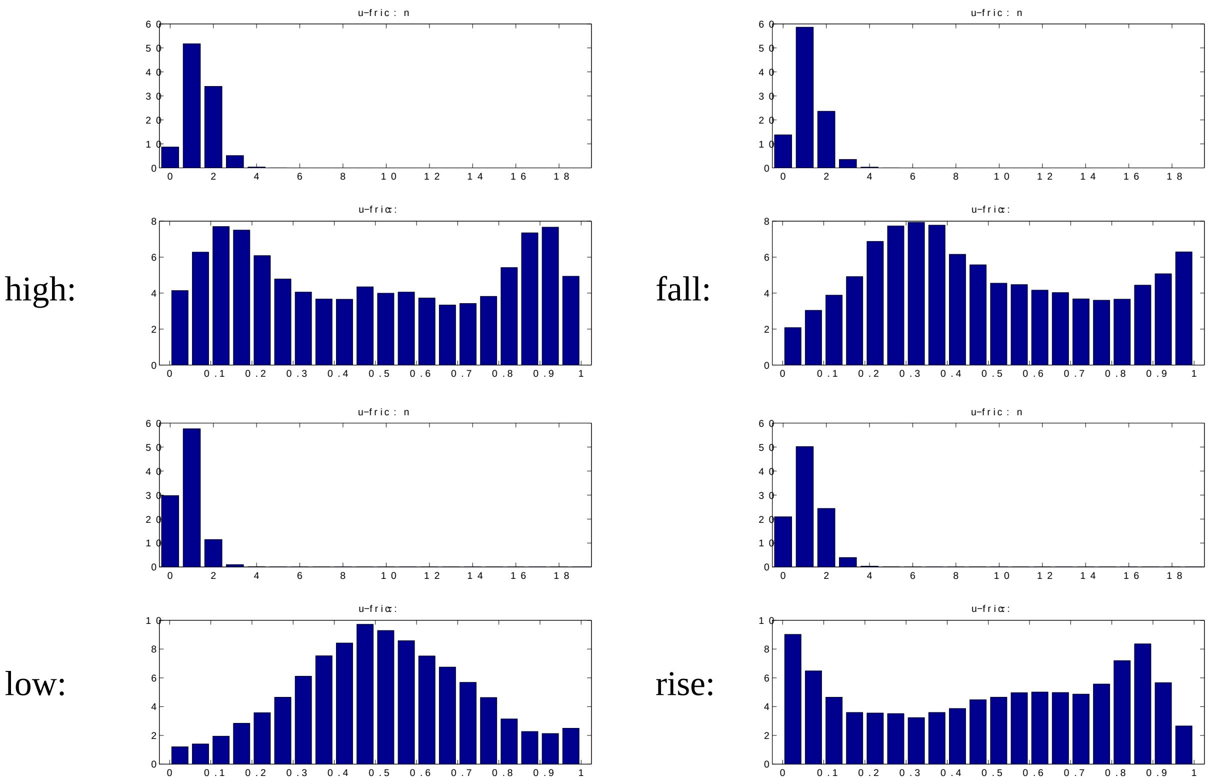

Figure 8. Characteristic points of the voiceless fricatives in TIMIT

The second histogram shows the distribution of the relative label position within the phoneme. While the phoneme class of voiceless fricatives shows a distinct modal structure, not all histograms posess similarly clear peaks. However, up to a window duration of 100ms every sound class has a significantly non-uniform distribution, the significance level being 5%. Uniformly distributed positions occur only for some sound classes and the longest window duration of 200ms. The peaks of the position histograms follow the basic high – fall – low – rise cycle within the phoneme.

Only the scaled entropy rate LOW distribution in Figure 8 displays a single peak. All other types of labels have two[6]. This might indicate either a subphonemic structure in the entropy rate contour of the voiceless fricatives or the influence of adjacent sounds. Also, it could simply originate from those sounds which receive more than one entropy rate label. The causes for the multimodal histograms cannot be distinguished at this state of the analysis.

[6] However, a third minor peak is visible in the center of the HIGH histogram.

5. Conclusion

The entropy rate contour of a speech signal seems to indicate regions of stationary and non-stationary spectral sound structure. Phoneme boundaries are often located near prominent points of the entropy rate contour, i.e. dominant maxima and plateau edges. Maxima, minima, steepest rising and falling slopes of the entropy rate contour are the best candidates for the detection of prominent points. The corpus evaluation quantifies how many characteristic point labels are found within each phoneme. No major differences of these quantities were found between the American English TIMIT corpus and the German Kiel Corpus. Important questions for further research are: Which is the entropy rate label type that is nearest to each phoneme boundary or landmark? and: What is the distribution function of the entropy rate labels in the vicinity of the phoneme boundaries?

6. References

Kohler, K. (1992). Erstellung eines Textkorpus für eine phonetische Datenbank des Deutschen. *Arbeitsberichte Inst. Phonetik Univ. Kiel* **26**, 11-40.

Liu, S.A. (1996). Landmark detection for distinctive feature-based speech recognition. *J. Acoust. Soc. Am.* **100**, 3417-3430.

Papoulis, A. (1984). *Probability, Random Variables and Stochastic Processes*, chapter 15, 500-567. Hamburg: McGraw-Hill.

Peterka, J. (1998). *Automatische Segmentierung von Sprachsignalen* (diploma thesis, TU-Wien).

Shannon, C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379-423, 623-656.

TIMIT (1993). *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*. CD-ROM. LDC.

Wiener, N. (1961). *Cybernetics: Or Control and Communication in the Animal and in the Machine* (2nd ed.). Cambridge: MIT Press.

Wokurek, W. (1999). Corpus based evaluation of entropy rate speech segmentation. *Proc. of the 14th Int. Congress of Phonetic Sciences, (ICPhS‘99),* 1217-1220.