CORPUS BASED EVALUATION OF ENTROPY RATE SPEECH SEGMENTATION

Wolfgang Wokurek
University of Stuttgart, Germany

ABSTRACT

The sequence of estimates of the speech signal's entropy rate is investigated as a potential basis for speech segmentation. Raising and falling edges of that entropy rate curve and its maxima and minima are considered as candidates for segment boundaries. These prominent points are compared to the phonetic segment boundaries and to acoustic landmarks. The comparison is made using the American TIMIT database and the German 'Kiel corpus of read speech' which are phonetically manually labelled speech corpora.

1. Introduction

Entropy is a quantiative measure of how uncertain the outcome of a random experiment is. Its definition and interpretation was introduced by C. E. Shannon [5] and is part of textbooks on statistics e.g. [3]. Two examples may illustrate the uncertainty concept of entropy. Throwing a fair dice is a random experiment with high uncertainty on its outcome, because each side may appear equally likely. Hence the entropy is high. If, on the other hand, a random experiment is biased and gives the same outcome most of the time, the uncertainty and the entropy is low. In particular the entropy H can be calculated from the probabilities p_i of the n outcomes by $H(p_1, \cdots, p_n) = \sum_{i=1}^{n} p_i \log p_i$.

That idea of measuring the uncertainty of a random experiment was extended to discrete time random signals (i.e. stationary stochastic process) by the notion of prediction. What will be the average uncertainty on the next sample, if all the past samples are known? The answer to that question is called the entropy rate of the stochastic process. Now the random experiment is: observing the next sample. Depending on the particular properties of the observed stochastic process the uncertainty on the next sample the knowledge of the past samples may be influenced or not. Hence the process properties will result in high or low uncertainty on the next one. An example for high uncertainty on the next sample is white noise, where we can definitely gain no insight on the next sample, regardless on how many past samples are allready known. Contrary to that a lowpass filtered noise leaves very much less uncertainty on the next sample even if only the current sample is known.

In the context of speech segmentation the distinction between stationary and nonstationary regions is important. Entropy rate is a candidate for that purpose due to the sensibility to statistic uncertainty.

In that paper the entropy rate is applied to the task of segmenting speech signals. In section 2. the estimation of the entropy rate based on spectral estimation techniques is described. The characteristic points of such entropy rate estimate contour are discussed for a speech signal in section 3.Finally the relation of that characteristic points of the entropy rate estimate contour to phoneme boundaries and acoustic landmarks [2] are investigated in section 4.

2. Entropy rate estimation

The entropy rate $H(\mathbf{x})$ of a stationary stochastic process $\mathbf{x}$ may be defined equivalently in two different ways [3]. The first definition formalizes the discussion in the introduction. The uncertainty of the unknown next sample $\mathbf{x}_n$ assuming the m known past samples $\mathbf{x}_{n-1}, \cdots, \mathbf{x}_{n-m}$ is expressed by the conditional entropy $H(\mathbf{x}_n | \mathbf{x}_{n-1}, \cdots, \mathbf{x}_{n-m})$. The assumption that all past samples are known is expressed by the limit $H_c = \lim_{m \to \infty} H(\mathbf{x}_n | \mathbf{x}_{n-1}, \cdots, \mathbf{x}_{n-m})$.

The second way of defining the entropy rate $H(\mathbf{x})$ starts from the average uncertainty of a block of m consecutive samples $\mathbf{x}_1, \cdots, \mathbf{x}_m$. Using the joint entropy $H(\mathbf{x}_1, \cdots, \mathbf{x}_m)$, that short term average uncertainty is $\frac{H(\mathbf{x}_1, \cdots, \mathbf{x}_m)}{m}$. Again, the average uncertainty includes all samples and requires the limit $\bar{H}(\mathbf{x}) = \lim_{m \to \infty} \frac{1}{m} H(\mathbf{x}_1, \cdots, \mathbf{x}_m)$. Both definitions result in the same quantity $H(\mathbf{x}) = H_c(\mathbf{x}) = \bar{H}(\mathbf{x})$ and is the entropy rate.

The calculation of both, the conditional entropy and the joint entropy require knowledge of the appropriate probability density functions. Estimation of joint densities is a data intensive as well as a computational intensive task. Segmentation of speech on such a basis was proposed recently [4] and employed the mutual information function. In contrast to that, a spectral approach is used here. In particular if the stochastic process is normal, the entropy rate results from its power spectrum $S(\Theta)_\mathbf{x}$

$$\bar{H}(\mathbf{x}) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \log S_\mathbf{x}(\Theta) d\Theta + \log \sqrt{2\pi e} \qquad (1)$$

where $\log$ is the logarithm to the basis e. That approach requires the knowledge of the power spectrum $S(\Theta)_\mathbf{x}$.

Unfortunately speech signals are neither stationary stochastic processes nor normal. Hence both the concept of the entropy rate and Eq. 1 may be applied with full right. But similar arguments apply to linear prediction techniques that have been successfully applied to speech signals. It may be argued that the speech signal may be locally approximated by a stationary, normal stochastic process. On that local basis spectral estimation techniques are applied to estimate the power spectrum $S(\Theta)_\mathbf{x}$. Since this is implemented by averaging magnitude spectra computed by the fast fourier transform (FFT), the resulting algorithm is computationally efficient.

The speech signal is processed using half 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

$$X(m,k) = \sum_{n=0}^{N-1} x([m-1]\frac{W}{2} + n) w(n) e^{j2\pi nk/N} \qquad (2)$$

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 is applicable. Squaring the spectral magnitudes and averaging every V consecutive spectra using a hamming window $v(n)$ results in a sequence of power spectral estimates

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

Again m is the frame number and k is the discrete frequency. The entropy rate $H(k)$ was estimated by adding the logarithms of that power spectral estimates

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

		$H(m)$			$h(m)$		
	w	$\overline{n}$	σ_n	N_τ	$\overline{n}$	σ_n	N_τ
h	5	4.010	4.093	967431	3.813	3.665	919929
	10	1.385	1.550	334244	1.482	1.679	357508
	20	0.656	0.772	158192	0.697	0.924	168143
	50	0.351	0.517	84566	0.330	0.542	79598
	100	0.192	0.397	46239	0.175	0.392	42159
	200	0.100	0.300	24040	0.086	0.282	20651
l	5	4.118	4.414	993461	3.827	3.729	923257
	10	1.363	1.656	328827	1.496	1.687	360978
	20	0.569	0.701	137209	0.713	0.896	171938
	50	0.331	0.489	79859	0.348	0.539	83846
	100	0.180	0.391	43483	0.193	0.408	46660
	200	0.087	0.283	20921	0.101	0.304	24314
r	5	4.407	4.274	1063228	4.208	3.999	1015222
	10	1.688	1.714	407233	1.611	1.835	388606
	20	0.781	0.835	188383	0.795	0.961	191836
	50	0.384	0.544	92764	0.339	0.532	81698
	100	0.191	0.400	46044	0.170	0.386	41034
	200	0.087	0.283	21021	0.079	0.271	19120
f	5	4.202	4.084	1013878	4.068	3.846	981506
	10	1.694	1.758	408718	1.604	1.785	386961
	20	0.811	0.885	195582	0.782	0.947	188730
	50	0.393	0.535	94748	0.339	0.530	81809
	100	0.198	0.408	47711	0.173	0.389	41858
	200	0.096	0.296	23098	0.079	0.271	19122

Table 1: Entropy rate labels of the TIMIT corpus.

		$H(m)$			$h(m)$		
	w	$\overline{n}$	σ_n	N_τ	$\overline{n}$	σ_n	N_τ
h	5	3.683	3.939	548371	3.706	6.858	551725
	10	1.481	1.908	220476	1.538	3.013	229010
	20	0.786	1.345	117004	0.775	1.540	115438
	50	0.399	0.688	59471	0.371	0.789	55229
	100	0.215	0.456	31972	0.207	0.508	30769
	200	0.104	0.310	15418	0.109	0.337	16164
l	5	3.784	3.989	563320	3.660	6.866	545000
	10	1.424	1.741	212050	1.503	2.998	223838
	20	0.760	1.281	113208	0.759	1.518	113015
	50	0.386	0.685	57431	0.375	0.779	55804
	100	0.199	0.456	29569	0.214	0.522	31789
	200	0.083	0.285	12420	0.119	0.388	17698
r	5	4.114	4.078	612527	4.106	7.390	611364
	10	1.764	2.082	262574	1.619	2.980	241025
	20	0.897	1.454	133602	0.814	1.526	121219
	50	0.421	0.724	62680	0.361	0.735	53805
	100	0.219	0.480	32615	0.190	0.460	28316
	200	0.107	0.325	15859	0.090	0.302	13343
f	5	4.189	3.984	623651	4.119	7.402	613245
	10	1.887	2.100	280999	1.710	3.036	254538
	20	0.962	1.478	143246	0.866	1.576	128885
	50	0.432	0.733	64271	0.381	0.759	56796
	100	0.218	0.477	32531	0.197	0.477	29365
	200	0.098	0.305	14576	0.091	0.302	13530

Table 2: Entropy rate labels of the KIEL corpus.

The shape of that entropy rate contour is invariant to linear filtering of the speech signal $x(n)$. The effect of any filter operation will be an additive constant. The same applies to mere signal amplification. The gain factor α will add $\log \alpha$ to $H(m)$. To remove the influence of signal amplitude from the entropy rate the scaled entropy rate $h(m)$ is proposed

$$h(m) = H(m) - \log \frac{1}{N} \sum_{k=0}^{N-1} S(m,k) \quad . \qquad (5)$$

Figure 1 shows the speech signal `train/dr7/mgaw0/sx85.wav` from the TIMIT database [6]. The sentence uttered by a male subject is: *Cement is measured in cubic yards.* The tracks below the speech waveform show (i) the logarithm of the short term energy, the entropy rate Eq. 4, (ii) the scaled entropy rate Eq. 5 and (iii) the phoneme segmentation. The frame size of the analysis is 10 milliseconds. At the first glance the entropy rate contour of a speech signal looks similar to a short time energy contour at logarithmic scale. But there are important differences at sounds of weak intensity i.e. fricatives and stop consonants. They are represented at the very same amplitude level as vowels and often get a higher peak than te vowels. That effect can be observed at the phonemes labelled /s/, /t/, /zh/ and the first /k/. The high peak of these sounds is 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.

That section shall be concluded with notes on the interpretation of the entropy rate and on the scaled entropy rate. First, the entropy rate $H(m)$ may be interpreted as a marginal of a (smoothed) spectrogram with logarithmic scale. Usually the spectrogram amplitudes are displayed or plotted using a decibel scale. But that are exactly the values $\log S(m,k)$ beeing added in Eq. 4 and resulting in $H(m)$ (aside from the constant N).

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 the value is zero. Every deviation from the constant specturm leads to negative values of 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 the minima of the entropy rate contour and the locations of its steepest raising and falling slopes. 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, vice versa a minimum. That criterion will fail on constant plateaus which are extremely unlikely in that kind of real world signal. The locations of steepest raising and falling slopes are detected by sign changes of the second difference contour. The steepest raise is indicated by a positive first difference and a change of sign from positive to negative second difference.

Figure 2 shows the same speech signal as Fig. 1 and the scaled entropy rate contours for three different frame widths (5ms, 20ms, 50ms). It allows drawing a comparison between the scaled entropy rates of different frame widths. All contours show the phoneme labels which are named using the phonetic codes from the TIMIT corpus. The second label track corresponds to the characteristic points of the scaled entropy rate with 50ms frame witdh. The third label track contains the automatic detected acoustic landmarks [2]. Currently only two landmark types are implemented the glottal landmark (g) and the sonorant landmark (g). They always come in pairs e.g. a glottal segment reaches from +g to -g.

The entropy rate contours are beeing smoothed with increasing frame width and loose their fine structure. Hence, the number of characteristic points decreases with increasing frame width. The characteristic points shown in Fig. 2 are derived from the slowest varying contour to limit the number of labels in the track.

The following rules for the interpretation of characteristic points may be derived from chapter 1. and chapter 2. and examined in Fig. 2:

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:

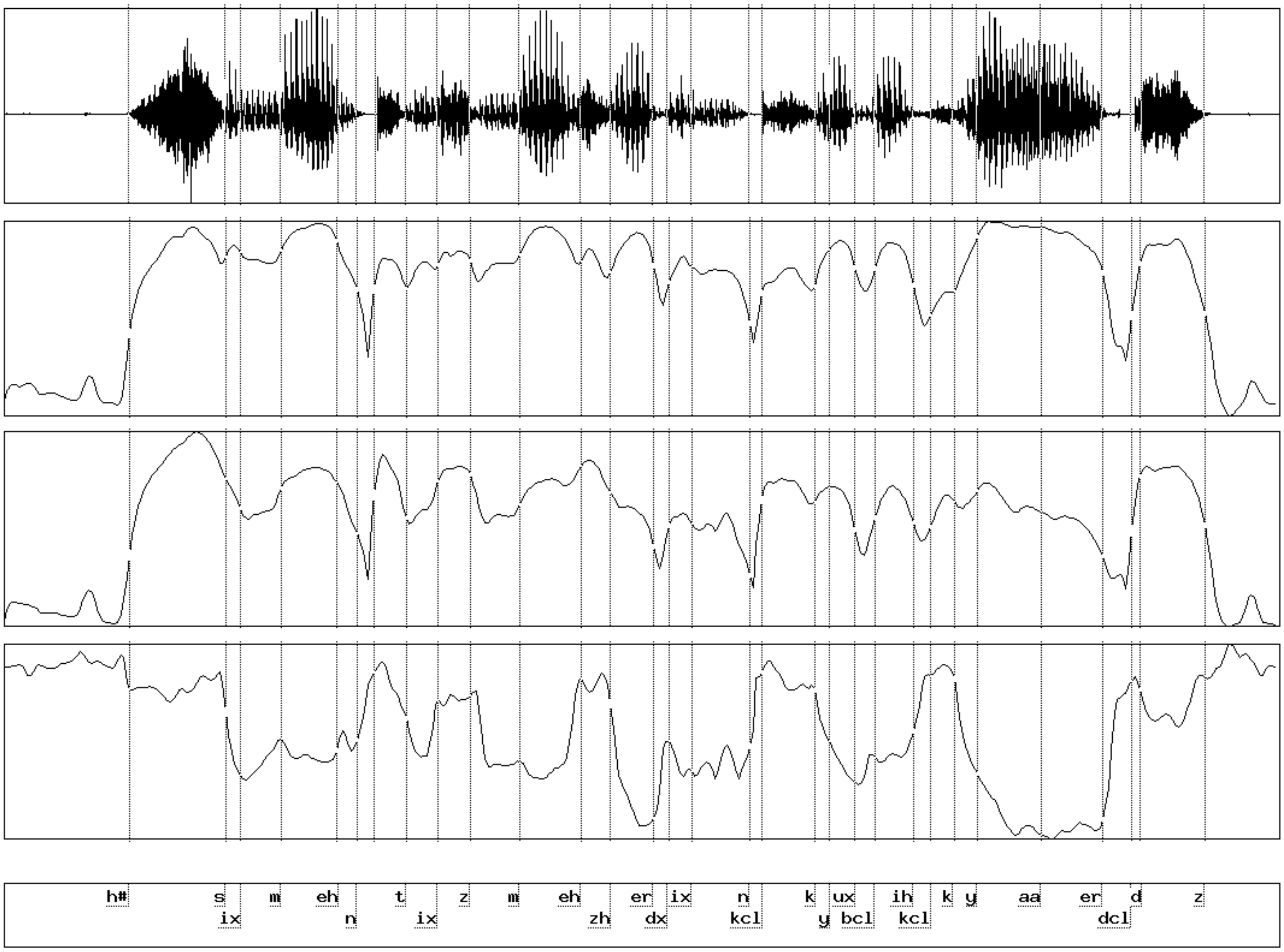

Figure 1: Signal, logarithm of the short term energy, entropy rate, scaled entropy rate. The frame width is 10 milliseconds. The phoneme segmentation of the TIMIT database is shown in the label track.

/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 raises: *transition from stationarity to nonstationarity* phoneme boundary (r2: /n/ → /tcl/, r3: /eh/ → /zh/, r4: /ix/ → /n/, r5: /n/ → /kcl/, r6: /er/ → /dcl/)

steepest falls: *transition from nonstationarity to stationarity* phoneme boundary (f2: /t/ → /ix/, f4: /y/ → /ux/, f6: /d/ → /z/)

4. Corpus Evaluation

The comparison between the entropy labels and the phoneme labels is made using the American 'TIMIT database [6] and the German 'Kiel corpus of read speech' [1] which are phonetically manually labelled speech corpora. The evaluation considers the entropy rate labels against the background of the phone segments. In particular the position and the count of the entropy rate labels within each phone are recorded. The results are shown in tables 1 and 2.

Table 1 answers the question: How many labels are within each phoneme of the TIMIT corpus? The frame duration w is listed in milliseconds (5, 10, 20, 50, 100, 200). The next three columns refer to the characteristic points of the entropy rate contours. Listed is the mean number of labels per phoneme $\overline{n}$, the standard deviation of that number σ_n and the total number of labels N_τ. What follows are the corresponding three columns that refer to the characteristic points of the scaled entropy rate contours.

The TIMIT corpus containes 241262 phoneme segments. The label position statistics are not listed because they are close to $\overline{\tau} = 0.5$ and $\sigma_\tau = 0.28$. Only in two cases the distribution was close to a uniform distribution. This was the case for the minima and for the raises of the entropy rate with 200 milliseconds framewidth and with maximal significance levels of $\alpha_{max} = 0.046$ and $\alpha_{max} = 0.174$ respectively.

The frame duration of the entropy rate estimation controls the smoothness of the curve and hence the label count per phone. That is shown in the $\overline{n}$ columns. The number of entropy rate labels starts with more than 4 labels per phoneme at a frame width of 5ms. It reaches about 1 label per phoneme between 10ms and 20ms. Finally it drops to about one label per 10 phonemes at a frame width of 200ms. For phoneme related entropy rate segmentation the optimum frame width may be found between 10ms and 20ms.

Essentially the same analysis was carried out for the KIEL corpus. Table 2 summarizes the results. It shows no obvious deviation from the results of the TIMIT corpus. Hence it may be concluded the characteristic points of the entropy rate contours behave very much the same way for American English and for German.

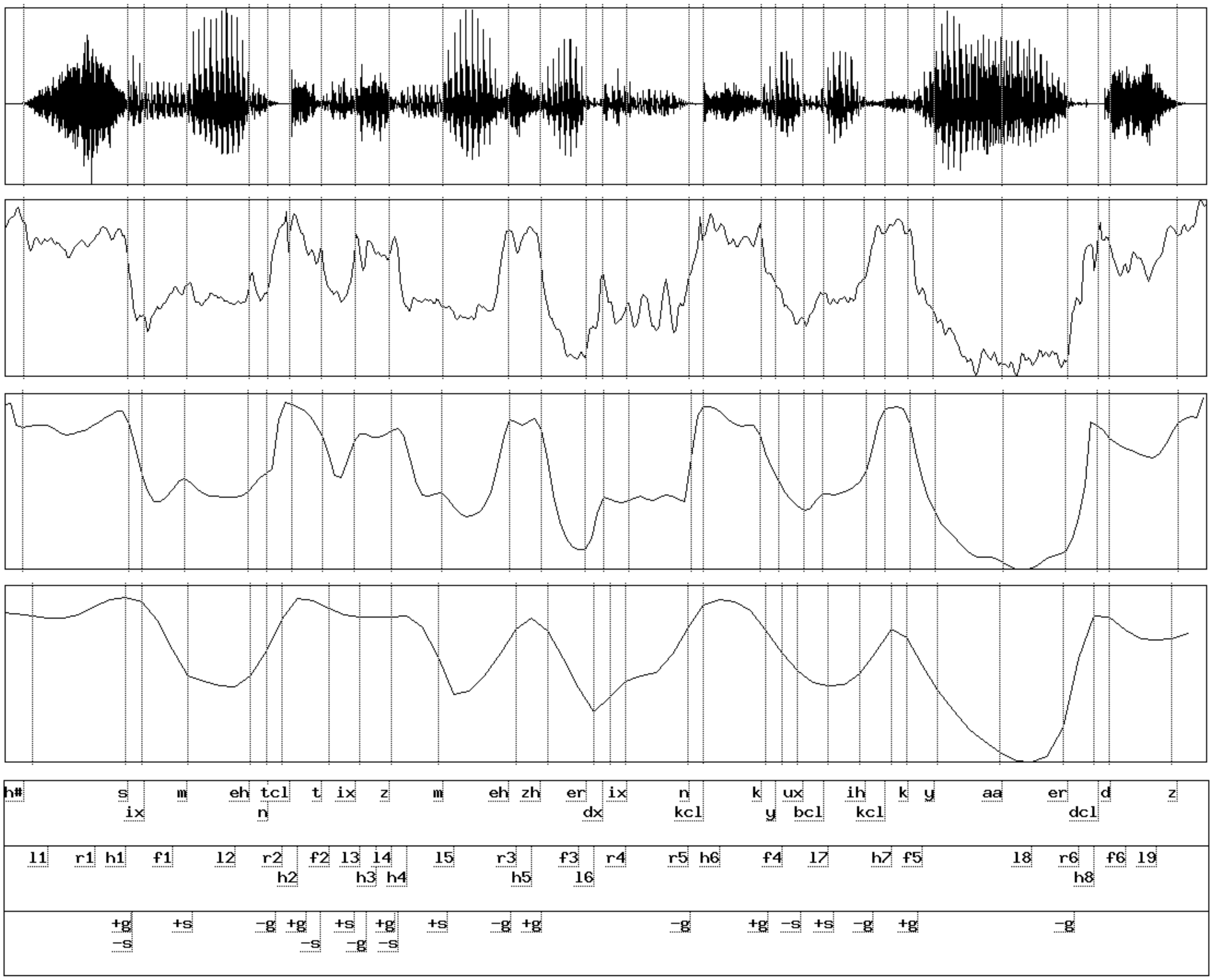

Figure 2: Signal, scaled entropy rate of frame widths: 5ms, 20ms, 50ms, segmentations: phonemes, characteristic points of the scaled entropy rate of 50ms frame width, acoustic landmarks (glottal, sonorant)

The KIEL corpus containes 148888 phoneme segments. The label position statistics are not listed because they are close to $\overline{\tau} = 0.5$ and $\sigma_\tau = 0.28$ but the distribution was never close to uniformity.

5. Conclusion

The entropy rate contour of a speech signal indicates regions of stationary and nonstationarity spectral structure. Phoneme boundaries are often located near prominent points of the entropy rate contour i.e. dominant maxima and the edges of its plateaus. Maxima, minima and steepest raising and falling slopes of the entropy rate contour at best are candidates for the detection of prominent points. Low valleys in the entropy rate contour often correspond to acoustic landmarks of sonorant type. The corpus evaluation quantifies how many characteristic point labels are within each phoneme. No major differences of these quantitys between the American English TIMIT corpus and the German KIEL corpus were found . Important questions for further research are: What is the entropy rate label type that is nearest to each phoneme boundary? and: What is the distribution function of the entropy rate labels around the phoneme boundaries?

REFERENCES

[1] Kohler, K., *Arbeitsberichte Inst. Phonetik Uni. Kiel* **26**, “Erstellung eines Textkorpus für eine phonetische Datenbank des Deutschen”, pp. 11–40 (1992).

[2] Liu, S. A., *J. Acoust. Soc. Am.* **100**, “Landmark detection for distinctive feature-based speech recognition”, pp. 3417–3430 (1996).

[3] Papoulis, A. *Probability, random variables and stochastic processes*, Hamburg: McGraw–Hill, 1984, ch. 15, pp. 500–567.

[4] Peterka, J. *Automatische Segmentierung von Sprachsignalen*, diploma thesis, TU-Wien, 1998.

[5] Shannon, C. E., *The Bell System Technical Journal* **27**, “A Mathematical Theory of Communication”, pp. 379–423, 623–656 (1948).

[6] The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus.