

AUTOMATIC SPEECH SEGMENTATION USING AVERAGE LEVEL CROSSING RATE INFORMATION

Anindya Sarkar and T.V. Sreenivas

Department of Electrical Communication Engineering,
Indian Institute of Science, Bangalore - 560 012, India.
email: anindya@ee.iisc.ernet.in tvsree@ece.iisc.ernet.in

ABSTRACT

We explore new methods of determining automatically derived units for classification of speech into segments. For detecting signal changes, temporal features are more reliable than the standard feature vector domain methods, since both magnitude and phase information are retained. Motivated by auditory models, we have presented a method based on average level crossing rate (ALCR) of the signal, to detect significant temporal changes in the signal. An adaptive level allocation scheme has been used in this technique that allocates levels, depending on the signal pdf and SNR. We compare the segmentation performance to manual phonemic segmentation and also that provided by Maximum Likelihood (ML) segmentation for 100 TIMIT sentences. The ALCR method matches the best segmentation performance without a priori knowledge of number of segments as in ML segmentation.

1. INTRODUCTION

In the 1960's, D.R.Reddy[1] had developed a speech segmentation scheme, using the variation of intensity levels and zero-crossing counts, and other program parameters were obtained by visual inspection of the waveform. More recently, spectrograms [2] have been used for hierarchical acoustic segmentation of continuous speech. Van Hemert [3] has used the intra-frame correlation measure between spectral features to obtain homogeneous segments. Statistical modeling (AR,ARMA) [4] of speech was also used for segmentation, by detecting sequential abrupt changes of model parameters. HMM based automatic phonetic segmentation [5] requires extensive training data but they have reported very high degree of segmentation accuracy. The popularly used feature vector based methods for speech segmentation are Spectral Transition Measure (STM) and Maximum Likelihood (ML) segmentation [6]. Of the spectral domain methods, ML is widely used for phone-level segmentation.

Due to coarticulation effects, the spectral transition across some phoneme boundaries is not clearly defined in the signal and phoneme boundaries start to appear, well within the previous phonemes. The question therefore is whether we are losing out on useful temporal information when we work in the spectral domain. Also, the question remains as to how robust the spectral features are to noise. It was shown in [7] that MFCC serves as a robust parameter set for speech segmentation. We would like to compare our time-domain method with the spectral domain methods (using MFCC) for automatic segmentation.

Since there is more information in the temporal domain, we explore a time-domain approach that uses a feature sensitive to both amplitude and frequency changes of the signal. The level crossing rate (LCR) of a signal is such a feature and is motivated by the human auditory analysis. The neural transduction at the output of the cochlea is believed to be in the form of signal level crossing detectors. This approach was used by Ghitza[8] to develop a noise robust front-end for speech recognition. In his work, the Ensemble Interval Histogram (EIH) model was proposed as an internal representation, from which spectral information is derived. Though EIH model follows a subband approach for spectrum estimation, we will use the fullband signal, in the temporal domain.

2. SEGMENTATION USING ALCR INFORMATION

We define the level crossing rate (LCR) at a sample point, and for a certain level, as the total of all the level crossings, that have occurred for that level, over a short interval around the point, divided by the interval duration. From the 2D information of LCR, we obtain the average level crossing rate (ALCR) at each point by summation of LCR over all levels. The motivation behind using the temporal information for segmentation is that when we change from one phoneme to the next, in most cases, there is a marked change in the vocal tract configuration. We expect this point of change to be expressed through a valley in the ALCR curve. When one phoneme waveform dies down to make way for another, the change will be manifested by substantial change in both amplitude and frequency, and ALCR curve, being sensitive to both, will be able to track the changes. Once the ALCR is computed, we smooth it using a moving-average filter for ease of identifying segment boundaries. Fig.1 demonstrates how ALCR depicts changes in signal waveform properties.

Consider a synthetic signal example of Fig 1. The signal shows both amplitude and frequency changes. By noting the points of change in ALCR curve and then, by applying duration constraints (as is done for speech segmentation), the segment boundaries, marked in black, can be clearly distinguished.

2.1. Overview of ALCR-based Segmentation

We will use ALCR to detect changes in the signal pattern. To apply level crossing on a signal, the level distribution has to be decided upon. In our experiments, we have used both uniform (pdf-independent) and non-uniform (pdf-dependent) level distributions.

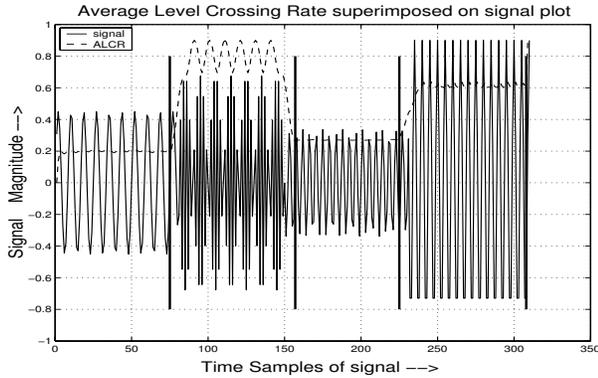


Fig. 1. Average Level Crossing Rate (ALCR) superimposed on signal waveform, brings out points of change in signal

Generally, speech signal amplitude $x[n]$ follows a Laplacian pdf, and phoneme changes are reflected within a certain signal range. Empirically, it is observed that for a signal, amplitude ranges such that the *CDF* $P(X \leq x)$ is within $[0.90, 0.99]$ and $[0.01, 0.10]$ are most useful for detecting phoneme changes. In the pdf-dependent scheme, we allocate more levels in the region of increased phonemic changes for detecting finer signal variations. Also, ALCR will be sensitive to noise, and by performing pdf-estimation on both signal and noise, we try to incorporate a degree of noise robustness in our methodology. To corroborate this observation, we have compared ALCR schemes, with and without the incorporation of noise robustness (Table 1). Once the level distribution and noise robustness issues are decided upon, we detect level crossings, within a window around each sample and for all levels. Thus, we have transformed the time-domain information to a new feature space, represented at each point by its ALCR. For segmentation, we wish to pick valleys from the ALCR curve, which has a rough contour. For removing spurious valleys, we apply mean smoothing on the ALCR curve.

2.2. ALCR-based Segmentation Algorithm:

The segmentation algorithm can be described as follows:

1. Input signal: $x[n]$ normalized to lie within $[-1, 1]$ and then made zero mean;
2. The levels η_j , ($1 \leq j \leq J$), where J = total no. of levels, can be distributed using the following schemes:
 - a) Uniform allocation of levels, to cover the entire dynamic range of the signal
 - b) Pdf-dependent non-uniform allocation of levels

For non-uniform level allocation, we select $\lambda_1, \lambda_2, \lambda_3$ and λ_4 , based on the empirical signal pdf, such that

$$\begin{aligned} F_s(\lambda_1) &= 0.01 \\ F_s(\lambda_2) &= 0.10 \\ F_s(\lambda_3) &= 0.90 \\ F_s(\lambda_4) &= 0.99. \end{aligned}$$

Here, $F_s(x)$ denotes the signal CDF, at value x . The histogram of speech data is taken as an approximation of its pdf.

Normally, the SNR of the speech is estimated using the noise variance of the silence regions. Let min and max denote

the minimum and maximum signal values, respectively. It is found from the TIMIT sentences that phonemic changes rarely occur in the signal of ranges $[\lambda_4, max]$ and $[min, \lambda_1]$. Therefore, we assign less no. of levels for these ranges. For stop consonants and bursts, changes occur mainly for values near 0. So, level changes in $[\lambda_2, \lambda_3]$ are important. However, levels in this range are also sensitive to noise. By putting more levels, we can detect all fine temporal changes, at the cost of false insertions. Signal changes are most prominent in the ranges $[\lambda_1, \lambda_2]$ and $[\lambda_3, \lambda_4]$, and so we allocate more levels here. Also, these levels are more robust to noise.

For SNR=36dB, we allocate 4,32,8,32 and 4 levels for the regions $[min, \lambda_1], [\lambda_1, \lambda_2], [\lambda_2, \lambda_3], [\lambda_3, \lambda_4]$ and $[\lambda_4, max]$, respectively. For SNR=10dB, we allocate 3,20,4,20 and 3 levels for these ranges. Though an optimal level distribution for a certain SNR has not been obtained, we propose that the level distribution should be adaptive, depending on the SNR. For lower SNR, we use less no. of levels to avoid false insertions, at the expense of missing out on some fine temporal changes. For higher SNR, we reduce the level spacing and allocate more levels in $[\lambda_1, \lambda_2]$ and $[\lambda_3, \lambda_4]$ to detect the fine changes. Similarly, for uniform level allocation scheme, large no. of levels (41-65) can be used at high SNR to detect fine changes. For 5-10 dB SNR, we conservatively use 13-25 levels to reduce insertion errors, at the cost of reduced segment match.

3. Noise Robustness Scheme:

We find the pdf of the background noise amplitude from the signal histogram, computed over the initial silence portion.

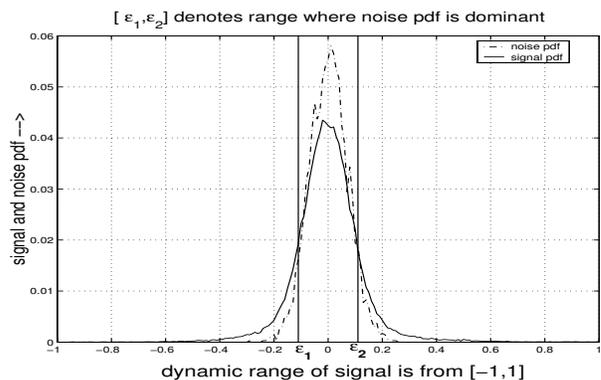


Fig. 2. By comparing the signal and noise histograms, we identify the amplitude range over which the noise pdf is dominant

We select $\epsilon_1 (\leq 0)$ and $\epsilon_2 (\geq 0)$, such that

$$\begin{aligned} f_s(\epsilon_1) &= f_n(\epsilon_1) \\ f_s(\epsilon_2) &= f_n(\epsilon_2). \end{aligned}$$

Here, $f_s(x)$ and $f_n(x)$ denote the PDF of the signal and the noise, respectively, at value x . We will avoid levels in the range $[\epsilon_1, \epsilon_2]$, (Fig. 2) for computing LCR.

4. For every sample, $x[n]$, and level η_j , a level crossing $\ell(j, n)$ has occurred between $x[n-1]$ and $x[n]$, if :
$$(x[n] - \eta_j)(x[n-1] - \eta_j) < 0$$

$$\ell(j, n) = \begin{cases} 1 & \text{if above condition is true} \\ 0 & \text{otherwise} \end{cases}$$

5. We define the level crossing rate $L(j, n)$ for each level η_j , at sample point n as:

$$L(j, n) = \sum_{m=n-\Delta}^{n+\Delta} \ell(j, m) \quad (1)$$

The interval Δ is chosen such that $2\Delta \approx$ one pitch period. Here, no accurate pitch estimation is required. E.g. for a female voice, the pitch is in the range 200-300 Hz. For a signal sampled at 16KHz, 1 pitch period will correspond to 80-53 samples. We have computed level crossings, over an interval of 100 samples (more than 1 pitch period).

6. From the 2D representation of LCR $L(j, n)$, we can obtain the average level crossing rate $E[n]$ over the ensemble of all levels:

$$E[n] = \sum_{j=1}^J L(j, n) \quad (2)$$

Further, we can smooth the average level crossing rate using either a linear or a non-linear filter. A simple moving average filter over $E[n]$ can be defined as:

$$\bar{E}[n] = \sum_{m=-100}^{100} E[n - m] \quad (3)$$

7. Once the vector $\bar{E}[n]$ is obtained, we pick its significant valleys and thus, estimate the number of segments. For picking the valleys, the following speech properties are used:
- Minimum segment duration = 12 ms
 - A valley must serve as a local minimum within a 20 ms interval. From $\bar{E}[n]$ plots, we observed that phonemic changes can occur at points of significant change in the slope of $\bar{E}[n]$, which may not be pure valleys. We set a tolerance limit tol to relax the selection criterion for valleys. If $\bar{E}[n] > (1 + tol) \cdot (\bar{E}[n_i])$ then $x[n]$ is not a valley. Here, n_i is any point in the 20 ms interval, having n as its central point. We used $tol = 0.05$. Higher values of tol will cause false insertions, in a single high energy or high frequency region, which may have high $\bar{E}[n]$, with some abrupt edges in $\bar{E}[n]$ curve.

3. EXPERIMENTS AND RESULTS

It was shown in [7] that MFCC parameters provide the most robust feature for segmentation. To compute MFCC (16 MFCC coefficients per frame), we have used an analysis window length of 20 ms, and window shift of 10 ms. To compare the performance of ALCR based segmentation, we have chosen two spectral domain methods:

- ML segmentation, using MFCC with a symmetric lifter $(1 + A \sin^{1/2}(\pi n/L))$, ($A=4$, $L=$ MFCC dimension=16) proposed in [7]
- Spectral Transition Measure (STM) using the feature vector and lifter combination in (2).

The experiments have been conducted on 100 sentences ($F_s=16$ KHz) - from TIMIT database- from 10 female speakers. The speech data has a SNR of 36dB, on an average, with respect to background noise in silence regions. We repeat the experiment with noisy speech with SNR of 20, 10 and 5 dB. For ML segmentation (without any duration constraint) using MFCC as the feature vector, we have assumed the same no. of segments as provided by TIMIT database. STM requires a global thresholding; to circumvent the problem, we have used STM with the same no. of segments, as given by TIMIT, and only those no. of largest peaks are detected for segmentation. If the obtained boundary is within ± 20 ms of a TIMIT boundary, we call it a 'match'(M). If two consecutive boundaries match, we count it as a 'segment match'(S). Also, insertions(I) and deletions(D) are noted, keeping the ± 20 ms constraint.

Method	SNR (dB)	M%	I%	D%	S%
ML	36	80.8	18.8	19.2	50.8
STM	36	70.1	25.2	29.9	34.1
U-LCR	36	78.6	22.8	21.4	44.2
NU-LCR	36	79.8	24.2	20.2	44.5
NU-LCR1	36	84.4	33.2	15.6	45.0
ML	20	78.8	20.8	21.2	46.8
STM	20	68.1	27.1	31.9	31.5
U-LCR	20	72.0	22.9	28.0	40.2
NU-LCR	20	77.6	25.7	22.4	43.2
NU-LCR1	20	79.0	36.0	21.0	39.8
ML	10	73.3	23.8	26.7	41.6
STM	10	65.0	30.2	35.0	28.0
U-LCR	10	69.7	24.9	30.3	37.9
NU-LCR	10	74.2	29.9	25.8	42.4
NU-LCR1	10	77.9	51.0	22.1	27.8
ML	5	70.6	26.4	29.4	38.5
STM	5	63.1	32.2	36.9	27.4
U-LCR	5	67.1	30.7	32.9	36.9
NU-LCR	5	71.2	39.9	28.8	37.7
NU-LCR1	5	65.9	55.0	34.1	23.8

Table 1. Segmentation performance of ALCR and other spectral domain methods

U-LCR: ALCR method with uniform level allocation

NU-LCR: ALCR method with non-uniform level allocation

(Both U-LCR and NU-LCR employ noise robustness scheme)

NU-LCR1: NU-LCR method without the noise robustness scheme

We see that segmentation accuracy of ALCR method is higher than STM, but lower than that of ML. It may be noted that ALCR technique does not use the number of segments information, unlike ML or STM, and yet provides a performance comparable to the best, *i.e.* ML. The performance is steady even at a low SNR of 5 dB. Also, the importance of the noise robustness scheme is evident by comparing the results of NU-LCR and NU-LCR1. NU-LCR1 shows higher false insertions which increases drastically at low SNR. For lower SNR, since $\bar{E}[n]$ increases, due to increased level crossings near 0 (as evident from Laplacian pdf of signal and also due to noise), the $\bar{E}[n]$ contour peaks up even at points of no distinct phoneme change and spurious valleys get detected. While comparing U-LCR and NU-LCR schemes, we see that match accuracy is more for NU-LCR, but that comes at the cost

of higher insertion errors, which become more prominent at lower SNR. In NU-LCR, we allocate more levels in the ranges $[\lambda_1, \lambda_2]$ and $[\lambda_3, \lambda_4]$, and as inter-level spacing is reduced, extra level crossings may be detected due to noise. In U-LCR, as levels are placed uniformly throughout, the susceptibility to noise is reduced.

We present (in Fig. 3-4), the spectrogram and time-domain plot of a signal ('sa1.wav'), along with the $\bar{E}[n]$ of the signal, which is called MSALCR ('Mean smoothed ALCR') in the plots. The figures show the analysis for two parts of the sentence, 1-2 sec and 2-3 sec, separately. Each figure consists of 3 plots, which are as follows:

upper plot: spectrogram of the signal, with TIMIT labels superimposed

middle plot: $\bar{E}[n]$ of the signal, with ALCR boundaries marked

lower plot: signal waveform, with ALCR boundaries shown

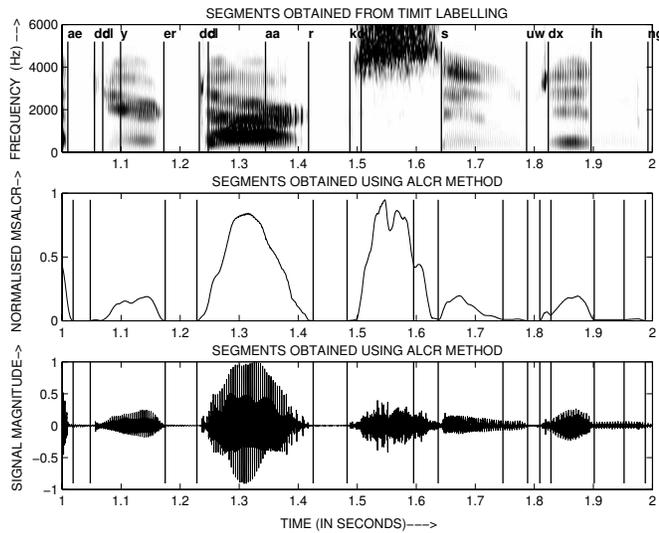


Fig. 3. Comparison of manual and ALCR boundaries over 1-2 sec of signal

As can be seen, the temporal changes in the signal are clearly identifiable through the ALCR method, and the segments are similar to those obtained through a visual inspection of the signal waveform. Vowels and diphthongs have high energy, while fricatives have higher frequency content. Nasals and stops have low energy. Therefore, based on ALCR, these phonemes can be easily segmented. *E.g.* stop consonants get picked up by ALCR method (/d/1.05 s, /d/1.22 s, /k/ 1.49 s, /g/2.08 s). Due to the noise robustness scheme, level crossings near 0 are ignored and so we miss out on detection of closures. Signal variations within a phonemic segment may give rise to insertion errors(/s/1.6 s, /s/2.3 s). In 'greasy' (2.05-2.45 sec), /g/t/iy/s/iy/, the /t/, which is part of a consonant cluster /g/t/, is missed. In 'dark' (1.25-1.45 s), /aa/ and /t/ have been merged into a single segment. Though a labelled boundary exists for /t/, we hear /aa/t/ as a single unit itself.

4. CONCLUSIONS

We have presented an auditory processing motivated time-domain segmentation technique, which performs as good as the best ML

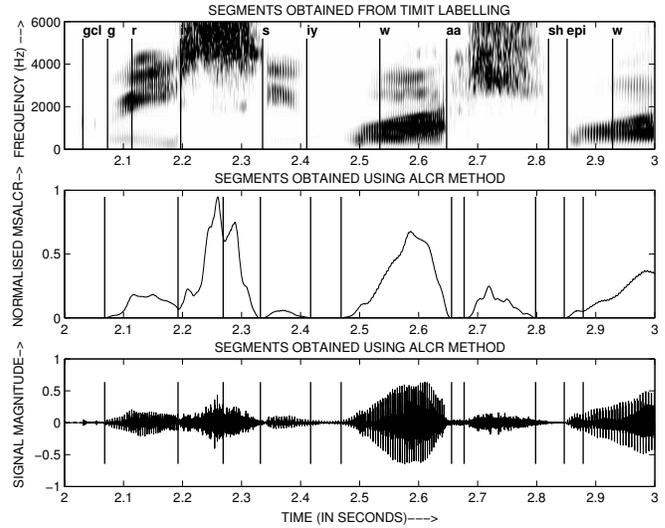


Fig. 4. Comparison of manual and ALCR boundaries over 2-3 sec of signal

segmentation, that we had proposed earlier. Adaptive level allocation and a noise robustness scheme have been incorporated into this technique. A second pass approach is required to detect those phoneme boundaries, which are not associated with a marked change in the ALCR contour. The technique can be further improved using non-linear processing techniques such as median filtering and 2D analysis of level crossing information.

5. REFERENCES

- [1] D.R. Reddy, "Segmentation of Speech Sounds", *J.Acoust.Soc.Am.*-1966, Vol. 40, No. 2, pp-307-312
- [2] James R. Glass and Victor W. Zue, "Multi-Level Acoustic Segmentation of Continuous Speech", *Proc. of ICASSP-1988*, pp: 429-432.
- [3] Jan P. van Hemert, "Automatic Segmentation of Speech", *IEEE Trans. on Signal Proc.*, Vol. 39, No. 4, April 1991, pp-1008-1012.
- [4] R. Andre-Obrecht, "Automatic Segmentation of Continuous Speech Signals", *Proc. of ICASSP-Tokyo,1986*, pp-2275-2278.
- [5] D.T. Toledano, L.A. Hernandez Gomez and L.V. Grande, "Automatic Phonetic Segmentation", *IEEE Trans. Speech and Audio Proc.*, Vol 11, No. 6, Nov. 2003, pp 617-625
- [6] T. Svendsen and F.K. Soong, "On the Automatic Segmentation of Speech Signals", *Proc. of ICASSP-Dallas,1987*, pp: 77-80.
- [7] A.K.V. SaiJayram, V.Ramasubramanian and T.V. Sreenivas, "Robust parameters for automatic segmentation of speech", *Proc. of ICASSP-May,2002*, pp-1-513-1-516
- [8] O. Ghizta, "Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech recognition", *IEEE Trans. Speech and Audio Proc.*, Vol 2, No.1, Jan. 1994, pp-115-132.