SPEECH SEGMENTATION USING EXTREMA-BASED SIGNAL TRACK LENGTH MEASURE

Prasanta Kumar Ghosh

Speech Analysis and Interpretation Laboratory
Department of Electrical Engineering, University of Southern California
Los Angeles, CA 90089
prasantg@usc.edu

ABSTRACT

We introduce a novel temporal feature of a signal, namely extrema-based signal track length (ESTL) for the problem of speech segmentation. We show that ESTL measure is sensitive to both amplitude and frequency of the signal. The short-time ESTL (ST_ESTL) shows a promising way to capture the significant segments of speech signal, where the segments correspond to acoustic units of speech having distinct temporal waveforms. We compare ESTL based segmentation with ML and STM methods and find that it is as good as spectral feature based segmentation, but with lesser computational complexity.

Index Terms— Speech analysis, Speech processing

1. INTRODUCTION

Automatic segmentation of speech is an important problem that is useful in speech recognition, synthesis and coding. There has been many approaches to the problem of automatic segmentation. In 1966 D.R. Reddy [1] has 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 HMM based automatic phonetic segmentation has been reported [2]; it 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 [3]. Of the spectral domain methods, ML is widely used for phone based segmentation.

It was shown in [4] that MFCC serves as a robust parameter for speech segmentation. Due to co-articulation effect, the spectral transition across some phoneme boundaries is not clearly defined and this results in segment boundaries to appear in wrong locations. Hence, temporal domain features have been explored for the problem of segmentation [5], which compare favorably with spectral domain methods.

We introduce a new concept of signal track length measure, which is found to be sensitive to both signal amplitude and frequency. Since acoustic segments have almost distinct waveform structure, which is reflected in amplitude and frequency of the the signal, we utilize signal track length to track the change in either amplitude or frequency of the signal.

2. EXTREMA-BASED SIGNAL TRACK LENGTH MEASURE

Let $x(t)$ be a continuous-time signal. The signal track length (STL) of $x(t)$ over the interval $[t_1,\ t_2]$ can be obtained by integrating the infinitesimal track length dr (as shown in Fig. 1) over time $t = t_1$ to $t = t_2$. Let us consider the infinitesimal

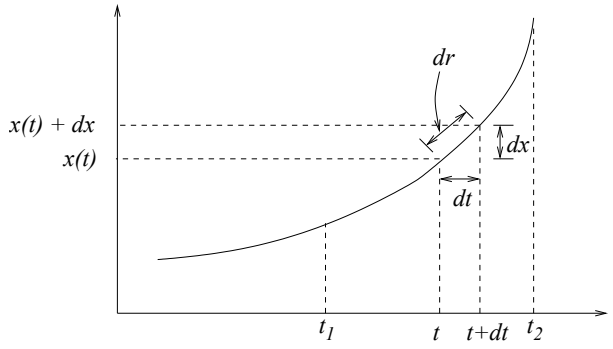

Fig. 1. *Signal track length measure.*

time duration from t to $t + dt$. Also let $x(t + dt) = x(t) + dx$. Then $dr = \sqrt{(dt)^2 + (dx)^2}$. Therefore, STL of $x(t)$ from $t = t_1$ to $t = t_2$ is

$$\begin{aligned} STL(t_1, t_2) &= \int_{t=t_1}^{t=t_2} dr = \int_{t=t_1}^{t=t_2} \sqrt{(dt)^2 + (dx)^2} \\ &= \int_{t=t_1}^{t=t_2} \sqrt{1 + \left(\frac{dx}{dt}\right)^2}\, dt \\ &= \int_{t=t_1}^{t=t_2} \sqrt{1 + (x'(t))^2}\, dt \end{aligned} \quad (1)$$

where $x'(t)$ is the first derivative of $x(t)$. Since the functional form of $x(t)$ or $x'(t)$ is not known, one has to resort to numer-

ical integration using samples of $x(t)$. However, we estimate STL using the extrema information of the signal. Let $t = t_{e_i}$ and $t_{e_{i+1}}$ be the locations of two consecutive extrema of $x(t)$; i.e., $x'(t_{e_i})$=0 and $x'(t_{e_{i+1}})$=0. Then the STL from $t = t_{e_i}$ to $t = t_{e_{i+1}}$ can be approximated as follows

$$STL(t_{e_i}, t_{e_{i+1}}) \simeq \sqrt{(t_{e_{i+1}} - t_{e_i})^2 + (x(t_{e_{i+1}}) - x(t_{e_i}))^2} \quad (2)$$

This we call extrema-based signal track length (ESTL) for the interval [t_{e_i}, $t_{e_{i+1}}$] denoted by $ESTL(t_{e_i}, t_{e_{i+1}})$. ESTL of $x(t)$ from $t = t_1$ to $t = t_2$ can be calculated by finding the extrema points in between t_1 and t_2 and summing all inter-extrema signal track length. If there are M extrema in between t_1 and t_2, $ESTL(t_1, t_2)$ can be calculated as follows,

$$\begin{aligned} ESTL(t_1, t_2) &= \sqrt{(t_{e_1} - t_1)^2 + (x(t_{e_1}) - x(t_1))^2} \\ &+ \sum_{i=1}^{M-1} \sqrt{(t_{e_{i+1}} - t_{e_i})^2 + (x(t_{e_{i+1}}) - x(t_{e_i}))^2} \\ &+ \sqrt{(t_2 - t_{e_M})^2 + (x(t_2) - x(t_{e_M}))^2} \end{aligned} \quad (3)$$

where the first and last term calculate the ESTL from t_1 to first extrema and M^{th} extrema to t_2 respectively.

2.1. Essential property of ESTL

The special property of ESTL is that it is sensitive to both signal amplitude and signal frequency; i.e., $ESTL(t_1, t_2)$ of $x(t)$ is determined by the amplitude and frequency content of the signal over the time interval [t_1, t_2] . This can be easily shown by considering $x(t)$ = Asin($\omega_0 t + \phi$), where A is the amplitude of the sinusoid; $\omega_0 = \frac{2\pi}{T_0} = 2\pi f_0$, where T_0 is the period of the sinusoid; ϕ is the initial phase of the signal.

Since the sinusoid has one extrema for every $\frac{T_0}{2}$ second, the number of extrema of $x(t)$ between t_1 and t_2 is given by

$$M \simeq \frac{t_2 - t_1}{T_0/2} = \frac{2(t_2 - t_1)}{T_0} \text{ [assuming } t_2 - t_1 \gg T_0] \quad (4)$$

By neglecting the first and last term in (3), (since $t_2 - t_1 \gg T_0$) we can write

$$\begin{aligned} ESTL(t_1, t_2) &\simeq \sum_{i=1}^{M-1} \sqrt{\left(\frac{T_0}{2}\right)^2 + (2A)^2} \\ &\simeq M\sqrt{\left(\frac{T_0}{2}\right)^2 + (2A)^2} \quad [M \gg 1] \\ &= (t_2 - t_1)\frac{4A}{T_0}\sqrt{1 + \left(\frac{T_0}{4A}\right)^2} \end{aligned} \quad (5)$$

For $\frac{T_0}{4A} \ll 1$

$$ESTL(t_1, t_2) \simeq (t_2 - t_1)\frac{4A}{T_0} = (t_2 - t_1)4Af_0 \quad (6)$$

The condition under which (5) reduces to (6) can be written as $Af_0 \gg .25$, i.e., $Af_0 \geq 2.5$. For $A = 1$ this means that the frequency of the sinusoid has to be greater than 2.5 Hz, which is a feasible assumption for most practical signals.

From (6) it is clear that ESTL is proportional to Af_0. Thus, short-time ESTL of a signal will exhibit significant change if either amplitude or frequency of the signal changes significantly. However, if the amplitude and frequency change in a signal is such that their product remains constant, the short-time ESTL (ST_ESTL) does not change. This is shown for a synthetic signal in Fig. 2. The Teager Energy Operator

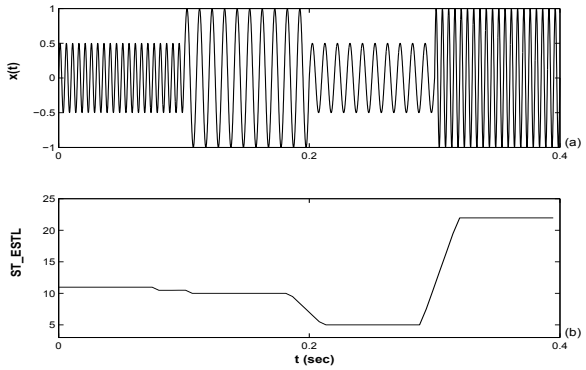

Fig. 2. *ESTL measure for a synthetic signal: (a) Signal consisting of four sinusoidal segments each of 0.1 sec having amplitude and frequency combinations as (0.5, 200Hz), (1, 100Hz), (0.5, 100Hz) and (1, 200Hz) respectively. (b) ST_ESTL computed using short-time window of 30 ms and shift of 5 ms.*

(TEO) of the signal $x(t)$ is proportional to $A^2 f_0^2$ [6]. Thus, the square of $ESTL(t_1, t_2)$ is also a measure of the local signal energy in the sense of TEO.

2.2. ESTL for discrete-time signals

For a discrete-time signal $x[n]$, extrema locations (EL) and extrema amplitude (EA) are estimated as follows

$$\begin{aligned} i^{th}\ EL\ \ \eta_i &= n \\ &\text{if } x[n-1] < x[n] > x[n+1], \\ &\text{or } x[n-1] > x[n] < x[n+1], \end{aligned} \quad (7)$$

$$\text{and } i^{th}\ EA\ \ \xi_i = x[\eta_i] \quad . \quad (8)$$

To minimize error in estimating extrema information in case of discrete-time signals we upsample $x[n]$ to a high sampling frequency ($F_s = \frac{1}{T_s}$, T_s : Sampling period), say 48 kHz for speech. Using these extrema information, we calculate ESTL from $n = n_1$ to $n = n_2$ by replacing t_{e_i} and t_i by $T_s\eta_i$ and $T_s n_i$ respectively in (3).

3. SPEECH SEGMENTATION USING ST_ESTL

Considering speech signal to be a multicomponent signal [7], different speech events have distinguished amplitude and frequency information. Since ST_ESTL changes with the change in both amplitude and frequency, we use ST_ESTL as a means of identifying different speech events. Let us look at a typical ST_ESTL plot of a speech file 'sa1.wav' taken from TIMIT database, shown in Fig. 3. Form Fig. 3(b) it is clear that

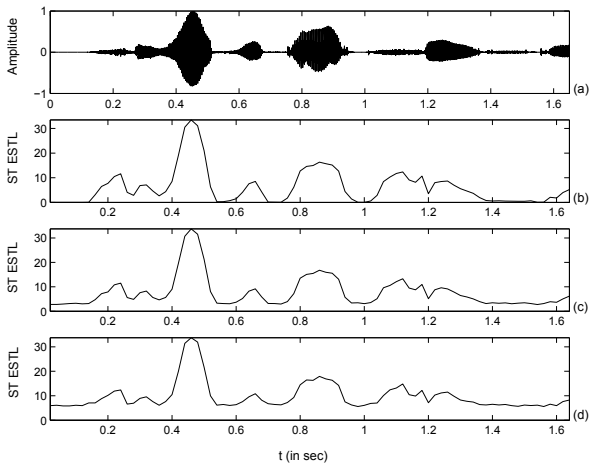

Fig. 3. *ESTL measure on speech data: (a) Speech signal (b) ST_ESTL for 20 ms frames, with 10 ms shift (c) ST_ESTL for 20 dB SNR condition (d) ST_ESTL for 10 dB SNR condition.*

ST_ESTL has very small value during silence region (either at the very beginning of the sentence or in between words). ST_ESTL has significant values when either the signal amplitude is high (0.42-0.5 sec, 0.8-0.9 sec) or signal frequency is high (0.15-0.25 sec, 1.04-1.18 sec). Hence, we pick the location at which ST_ESTL has significant changes and mark them as segment boundaries. The algorithm for finding the segment boundaries (given the required no of segments P for a speech signal) is described below :

- **Step 1:** Given a speech signal $x[n]$, normalize its amplitude to +1 to -1.
- **Step 2:** Use analysis window of 20 msec and shift of 10 msec to calculate ST_ESTL$_m$, where m is the frame index.
- **Step 3:** Let ESTL$_{min}$ and ESTL$_{max}$ denote the minimum and maximum value of the ST_ESTL$_m$. Since the ST_ESTL$_m$ will yield small values for non-speech region of the signal, choose a threshold ESTL$_{th}$ = ESTL$_{min}$ + 0.1{ESTL$_{max}$-ESTL$_{min}$} and mark the frames as segment boundaries, where ST_ESTL plot crosses ESTL$_{th}$. This gives a gross segmentation, which mainly isolates speech from silence. Suppose, we obtain P_1 number of segment boundaries from this step.
- **Step 4:** Consider all valleys in ST_ESTL plot with amplitude A_v such that $A_v > ESTL_{th}$. Let us consider i^{th} valley with amplitude A_{v_i} and the respective left and right peak with amplitude A_{l_i} and A_{r_i}. We arrange all valleys in ascending order according to the valley depth factor defined as $\frac{A_{v_i}}{(A_{l_i}A_{r_i})^{\frac{1}{2}}}$. Out of these arranged valleys we select the first $(P-1-P_1)$ valleys as segment boundaries.

We also explore the noise robustness property of ST_ESTL curve for segmentation at high and medium SNR condition. From Fig. 3(b) and (c) it can be noted that the peak and valley locations of ST_ESTL remain unaltered for signal with additive noise upto 10 dB SNR. Hence the same algorithm is used to find segment boundaries in noisy cases.

4. EXPERIMENTS AND RESULTS

It was shown in [4] that MFCC parameters provide the most robust feature for segmentation. To compute MFCC (16 MFCC coefficients per frame), we have used a analysis window length of 20 ms, and window shift of 10 ms. To compare the performance of ESTL based segmentation, we have chosen two spectral domain methods:
*(1)*ML segmentation, using MFCC with a symmetric lifter (1+$A\sin^{\frac{1}{2}}(\frac{\pi n}{L})$), proposed in [4].
*(2)*Spectral Transition Measure (STM) using the feature vector and lifter combination in (1).

The experiments have been conducted on 50 male and 50 female speakers' sentence (Sampling frequency = 16 kHz) taken from TIMIT database. The experiment was performed on these clean speech sentence as well as on noisy speech with SNR of 30, 20 and 10 dB. The no of manual segments given by TIMIT database is used as input to the ESTL based segmentation algorithm. For ML segmentation (without any duration constraint), we have assumed the same number of segments. STM requires a global thresholding; to circumvent the problem, we have used STM with same number of segments, and only those number of largest peaks are detected for segmentation. If the obtained boundary is within $\pm$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, insertion(I) and deletions(D) are noted, keeping the $\pm$20 ms constraint.

From table 1 we see that the segmentation performance of ESTL method is better than STM but worse than that of ML. The performance is steady at medium SNR due to noise robustness property of the ESTL. As the SNR of the signal decreases, %M for ESTL-based method remains almost same but the insertion rate increases. This is because even though the shape of ST_ESTL curve remain same, with additive noise spurious vallyes appear in the ST_ESTL curve which results in increasing number of wrong segment boundaries. The ESTL based segmentation is *computationally simpler than ML* as we need to pick only the valleys of the one di-

Table 1. *Segmentation performance of ESTL and other spectral domain methods*

Method	SNR(dB)	M%	I%	D%	S%
ESTL	clean	80.1	11.9	19.9	53.4
ML	clean	85.1	11.7	14.9	56.1
STM	clean	69.0	25.4	31.0	48.9
ESTL	30	80.0	12.7	20.0	52.0
ML	30	86.0	12.5	14.0	55.3
STM	30	69.3	25.1	30.7	42.1
ESTL	20	79.4	15.7	20.6	48.3
ML	20	84.3	16.9	15.7	53.7
STM	20	71.7	26.8	28.3	42.9
ESTL	10	78.9	20.8	21.1	43.7
ML	10	76.9	22.1	23.1	49.8
STM	10	68.8	27.7	31.2	40.5

mensional feature contour, unlike the full-search optimization using multi-dimensional spectral features in ML. Simulations show that the cpu time for running segmentation algorithm using the ESTL is almost 25% of that of the ML segmentation.

We present (in Fig. 4 and Fig. 5), the spectrogram and time domain plot of two signal segments along with ST_ESTL of the signal. In both figures part (a) shows the spectrogram with overlaid segment boundaries as in TIMIT database, part (b) shows the time domain plot of the speech signal and part (c) shows the ST_ESTL with segment boundaries obtained from ST_ESTL based algorithm.

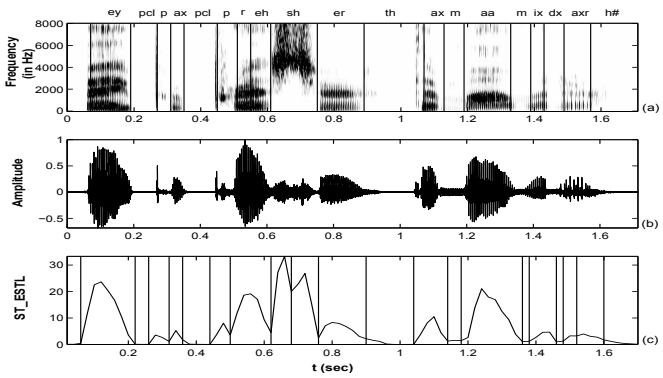

Fig. 4. *Comparison of manual and ESTL boundaries for 'si1271.wav' over last 1.7 sec.*

As can be seen from both the figures, most of the acoustic units in speech can be identified by either amplitude or frequency changes. Vowel and diphthongs have high energy, while fricatives have higher frequency content. Therefore, they are well segmented using the algorithm applied on ST_ESTL plot. Nasal and stops are also segmented easily. Amplitude variation within a phonemic segment has given rise to insertion error (/sh/ at 0.6 sec, Fig. 4). Also in 'programs' (0.82-1.46sec Fig. 5), /p/r/ow/g/r/ae/m/, /r/ phone (part of consonant clusters /g/r/) and /l/ (0.45 sec in Fig. 5) have been missed.

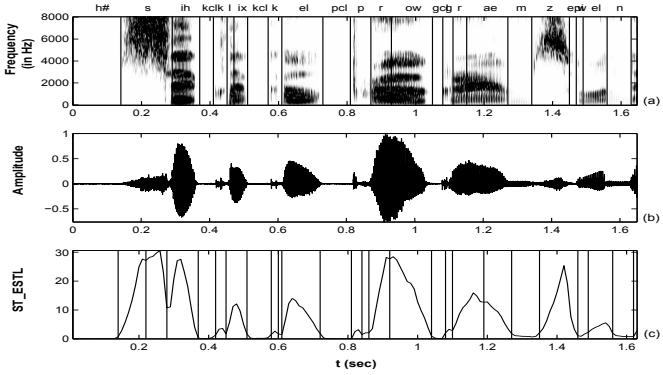

Fig. 5. *Comparison of manual and ESTL boundaries for 'sx146.wav' over 0-1.6 sec.*

5. CONCLUSION

We have presented a temporal domain method for the problem of speech segmentation. ESTL is found to capture both amplitude and frequency changes in the signal and hence is suitable for finding segments of speech which have distinct amplitude or frequency content.

6. ACKNOWLEDGEMENT

The author wishes to thank Prof. T.V. Sreenivas. This work was done under his supervision at speech lab, Indian Institute of Science, Bangalore, India.

7. REFERENCES

[1] D.R. Reddy, "Segmentation of Speech Sounds", J.Acoustic.Soc.Am., 1966, Vol. 40, No. 2, pp 307-312.

[2] 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.

[3] T. Svendsen and F.K. Soong, "On the Automatic Segmentation of Speech Signals", Proc. ICASSP, Dallas, 1987, pp 77-80.

[4] A.K.V. SaiJayram, V. Ramasubramanian and T.V. Sreenivas, "Robust parameters for automatic segmentation of speech", Proc. ICASSP, May 2002, pp I-513-516.

[5] Anindya Sarkar and T.V. Sreenivas, "Automatic Speech Segmentation using Average Level Crossing Information", Proc. ICASSP, March 2005, pp I-397 - I-400.

[6] J.F. Kaiser, "On Teager's energy algorithm and its generalization to continuous signals", Proc. 4th IEEE Digital Signal Proc. Workshop, Sep 1990.

[7] R.J. McAulay and T.F. Quatieri, "Speech analysis/Synthesis based on a sinusoidal representation", IEEE Trans. ASSP, Vol. 34, Issue 4, pp. 744 - 754, Aug 1986.