Speech Segmentation Methodologies for Speech Analysis

M.E. (Signal Processing) Mid-Term Project Evaluation

18 January, 2005

Department of Electrical Engineering,
Indian Institute of Science,
Bangalore - 560 012
Anindya Sarkar

M.E. (Signal Processing),
Dept of Electrical Engg,
Project Guide: Prof T.V. Sreenivas,
Indian Institute of Science, Bangalore 560012, India

email: anindya@ee.iisc.ernet.in, tvsree@ece.iisc.ernet.in
What is Speech Segmentation?

- Speech is a non-stationary signal
- Some degree of *homogeneity* is found in smaller units of speech, like phonemes
- Basic aim of *segmentation* is to capture certain units of speech, such that each unit is maximally *homogeneous* within itself
- Segmentation can be manual or automatic. E.g. An accuracy test for automatic segmentation can be done by comparing with phoneme boundaries set by linguists.
If each speech segment were truly homogeneous, we would have 7 well-defined segments as in this example.

Figure 1: Graphical Explanation of Perfectly Homogeneous Segmentation
Figure 2: Overview of Speech Segmentation Methodologies & Issues
Figure 3: Aspects in Pattern Recognition based Approach to Segmentation
Two-fold Breakup of our work:

- Most of the existing segmentation methodologies work on spectral domain feature vectors. We have explored new methods for *time-domain feature* based segmentation, where we used “Average Level Crossing Rate Information” for Automatic Segmentation.

- In established techniques like ML segmentation, number of segments is *fixed apriori*. We have used a Fischer ratio based Discriminant measure to estimate the optimum number of segments in a given speech utterance.
Past Segmentation Techniques

- Victor Zue: spectrograms were used for multi-scale hierarchical acoustic segmentation of continuous speech
- Van Hemert: intra-frame correlation was used as measure to obtain homogeneous segments
- Andre-Obrecht: Use of statistical models (AR/ARMA) for continuous speech; segmentation done by detecting sequential abrupt change of model parameters
- HMM based automatic phonetic segmentation - performance improves with training
- Svendsen: introduced the following spectral feature based methods: Maximum Likelihood and Spectral Transition Measure
Various Applications of Speech Segmentation

- Speech Segmentation is an important front-end feature in most applications
- Coding Applications, as in Segment Vocoder
- Subword based Isolated Word Recognition
- Segmental Models for Language Identification Work
- Segmental Models for Speech Enhancement
Why use Temporal Domain Features for Segmentation?

- Due to Coarticulation effects, spectral transition is often not clearly defined across phoneme boundaries
- Also the issue remains as to how robust the spectral features are to noise
- From a time-domain viewpoint, signal changes are suggested by change in signal amplitude and frequency
- A feature which is sensitive to both should suffice
Level Crossing Rate (LCR): Auditory System Motivated Feature

- LCR of a signal is sensitive to both *amplitude* and *frequency* changes and is motivated by the human auditory analysis.
- Firing of nerve cells at output of cochlea occurs when signal passes through a series of level-crossing detectors.
- Ghitza had proposed the Ensemble Interval Histogram (EIH) Model and it was a subband approach for spectral estimation.
- Our LCR-based approach is for the fullband signal.
Average Level Crossing Rate (ALCR)- Our Proposed Feature

- LCR at a sample point, for a certain level, is defined as the total of all the level crossings, for that level, over a short interval around the point, divided by interval duration.
- ALCR, at a certain point, is obtained by summation of LCR over all levels.
Motivation Behind using ALCR for Segmentation

- When we change from one phoneme to the next, there is a *marked change* in the vocal tract configuration.
- As one phoneme waveform dies down and another starts, there is substantial change in both amplitude and frequency.
- ALCR, being sensitive to both, will be able to track the changes.
We take a synthetic signal, with both amplitude and frequency changes. Then, based on ALCR, we track the changes.

![Average Level Crossing Rate superimposed on signal plot](image)

**Figure 4:** Average Level Crossing Rate (ALCR) superimposed on signal waveform, brings out points of change in signal
ALCR-Based Segmentation Algorithm

- Input signal: $x[n]$ normalized to lie within $[-1,1]$ and then made zero mean
- The levels $\eta_j$, $(1 \leq j \leq J)$, 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
Philosophy behind Pdf-Dependent Level Allocation

• Generally, speech signal amplitude follows a Laplacian pdf and most phoneme changes occur 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.
CDF of signal $\rightarrow$

$\text{CDF}(x=\lambda_1)=0.01$ and $\text{CDF}(x=\lambda_2)=0.10$ and $\text{CDF}(x=\lambda_3)=0.90$ and $\text{CDF}(x=\lambda_4)=0.99$

Figure 5: From the Laplacian cdf of the speech signal, we allocate more levels in $[\lambda_1, \lambda_2]$, $[\lambda_3, \lambda_4]$, where there is more significant signal change
To incorporate a measure of noise robustness, we do not allocate levels in that portion where noise pdf dominates over the signal pdf. We select $\epsilon_1$ and $\epsilon_2$, such that

$$\epsilon_1 = \max_{x, x \leq 0} \{f_s(x) > f_n(x)\}.$$  
$$\epsilon_2 = \min_{x, x \geq 0} \{f_s(x) > f_n(x)\}$$

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 $[\epsilon_1, \epsilon_2]$, for computing LCR.
Figure 6: By comparing the signal and noise histograms, we identify the amplitude range over which the noise pdf is dominant.
For every sample, $x[n]$, and level $\eta_j$, a level crossing $\ell(j, n)$ has occurred between $x[n - 1]$ and $x[n]$, if:

$$(x[n] - \eta_j)(x[n - 1] - \eta_j) < 0$$

$$\ell(j, n) = \begin{cases} 
1 & \text{if above condition is true} \\
0 & \text{otherwise} 
\end{cases} \quad (1)$$
We define the level crossing rate $L(j, n)$ for each level $\eta_j$, at sample point $n$ as:

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

(2)

The interval $\Delta$ is chosen such that $2\Delta \approx$ one pitch period.
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:

$$E[n] = \sum_{j=1}^{J} L(j, n)$$  \hspace{1cm} (3)

We then smooth $E[n]$ using a moving average filter to obtain $\bar{E}[n]$. Once $\bar{E}[n]$ is obtained, we pick its significant valleys and thus, estimate the number of segments.
Segmentation Experiments

- Automatic Segmentation of 100 sentences ($F_s=16$KHz) - from TIMIT database - from 10 female speakers
- Comparison of ALCR-based technique with two spectral domain methods: 1) Maximum Likelihood (ML) and 2) Spectral Transition Measure (STM)
- Feature Vector used - MFCC (for noise robustness) with 16 coefficients per frame, with Analysis Window=20ms, Window shift=10ms
- Clean speech SNR=36dB; we repeat the experiment with noisy speech with SNR of 20, 10 and 5 dB
Explanation of Different Distortion Terms

The total intra-segment distortion \( D_{\text{intra}(\text{total})} \), over the entire speech utterance, and the distortion, at point of transition from \( n^{th} \) to \((n+1)^{th}\) frame, \( D_{\text{transition}}(n) \), are defined as follows, where we assume the feature vector distribution within each segment is modelled by a multivariate Gaussian: Consider a feature vector set of \( N \) frames, \( \{X_i\}_{i=1}^N \), to be partitioned into \( c \) segments, with segment means : \( \{\mu_i\}_{i=1}^c \), segment boundaries : \( \{b_i\}_{i=1}^c \)

\[
D_{\text{intra}(\text{total})} = \sum_{i=1}^c \sum_{n=b_{i-1}+1}^{b_i} d(X_n, \mu_i)
\]

\[
D_{\text{transition}}(n) = d(X_n, X_{n+1})
\]
Maximum Likelihood (ML) & Spectral Transition Measure (STM)

- In ML segmentation, we use a Dynamic Programming procedure to minimize $D_{intra}(total)$.
- In STM, we find $D_{transition}(n)$ for all $n$, and then allocate segment boundaries at those locations, where $D_{transition}(n)$ is higher than a certain pre-defined threshold.
- We have used norm squared as the distance measure $d$ used in defining the distortion terms.
For judging segmentation accuracy, we have noted segment boundary match (M), full segment match (S), insertions (I) and deletions (D).

Notations used in Tables 1 & 2 are as follows:

- **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
So, we have 8 matches (M), 3 insertions (I), 2 deletions (D), and 2 full segment matches (S).
For our results, we have considered 2 boundaries as matching if they lie within $\pm 20$ ms.

Figure 7: Pictorial representation of concepts of segment boundary match (M), full segment match (S), insertions (I) and deletions (D)
<table>
<thead>
<tr>
<th>Method</th>
<th>SNR(dB)</th>
<th>M%</th>
<th>I%</th>
<th>D%</th>
<th>S%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>36</td>
<td>80.8</td>
<td>18.8</td>
<td>19.2</td>
<td>50.8</td>
</tr>
<tr>
<td>STM</td>
<td>36</td>
<td>70.1</td>
<td>25.2</td>
<td>29.9</td>
<td>34.1</td>
</tr>
<tr>
<td>U–LCR</td>
<td>36</td>
<td>78.6</td>
<td>22.8</td>
<td>21.4</td>
<td>44.2</td>
</tr>
<tr>
<td>NU–LCR</td>
<td>36</td>
<td>79.8</td>
<td>24.2</td>
<td>20.2</td>
<td>44.5</td>
</tr>
<tr>
<td>NU–LCR1</td>
<td>36</td>
<td>84.4</td>
<td>33.2</td>
<td>15.6</td>
<td>45.0</td>
</tr>
<tr>
<td>ML</td>
<td>20</td>
<td>78.8</td>
<td>20.8</td>
<td>21.2</td>
<td>46.8</td>
</tr>
<tr>
<td>STM</td>
<td>20</td>
<td>68.1</td>
<td>27.1</td>
<td>31.9</td>
<td>31.5</td>
</tr>
<tr>
<td>U–LCR</td>
<td>20</td>
<td>72.0</td>
<td>22.9</td>
<td>28.0</td>
<td>40.2</td>
</tr>
<tr>
<td>NU–LCR</td>
<td>20</td>
<td>77.6</td>
<td>25.7</td>
<td>22.4</td>
<td>43.2</td>
</tr>
<tr>
<td>NU–LCR1</td>
<td>20</td>
<td>79.0</td>
<td>36.0</td>
<td>21.0</td>
<td>39.8</td>
</tr>
</tbody>
</table>

Figure 8: Segmentation performance of ALCR and other spectral domain methods (at SNR of 36 & 20 dB)
### Table 1: Segmentation performance of ALCR and other spectral domain methods (at SNR of 10 & 5 dB)

<table>
<thead>
<tr>
<th>Method</th>
<th>SNR(dB)</th>
<th>M%</th>
<th>I%</th>
<th>D%</th>
<th>S%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>10</td>
<td>73.3</td>
<td>23.8</td>
<td>26.7</td>
<td>41.6</td>
</tr>
<tr>
<td>STM</td>
<td>10</td>
<td>65.0</td>
<td>30.2</td>
<td>35.0</td>
<td>28.0</td>
</tr>
<tr>
<td>U–LCR</td>
<td>10</td>
<td>69.7</td>
<td>24.9</td>
<td>30.3</td>
<td>37.9</td>
</tr>
<tr>
<td>NU–LCR</td>
<td>10</td>
<td>74.2</td>
<td>29.9</td>
<td>25.8</td>
<td>42.4</td>
</tr>
<tr>
<td>NU–LCR1</td>
<td>10</td>
<td>77.9</td>
<td>51.0</td>
<td>22.1</td>
<td>27.8</td>
</tr>
<tr>
<td>ML</td>
<td>5</td>
<td>70.6</td>
<td>26.4</td>
<td>29.4</td>
<td>38.5</td>
</tr>
<tr>
<td>STM</td>
<td>5</td>
<td>63.1</td>
<td>32.2</td>
<td>36.9</td>
<td>27.4</td>
</tr>
<tr>
<td>U–LCR</td>
<td>5</td>
<td>67.1</td>
<td>30.7</td>
<td>32.9</td>
<td>36.9</td>
</tr>
<tr>
<td>NU–LCR</td>
<td>5</td>
<td>71.2</td>
<td>39.9</td>
<td>28.8</td>
<td>37.7</td>
</tr>
<tr>
<td>NU–LCR1</td>
<td>5</td>
<td>65.9</td>
<td>55.0</td>
<td>34.1</td>
<td>23.8</td>
</tr>
</tbody>
</table>

**Figure 9:** Segmentation performance of ALCR and other spectral domain methods (at SNR of 10 & 5 dB)
Analysis of Tabulated Results

- ALCR method gives higher accuracy than STM but lower than ML; however, it does not know apriori no. of segments
- Both U-LCR and NU-LCR give steady performance at very low SNR
- Effectiveness of noise robustness scheme shows up in increased insertion errors for NU-LCR1, as compared to NU-LCR
- Among U-LCR & NU-LCR, the latter shows higher match accuracy while the former shows lower insertion errors
Graphical Presentation of ALCR based Segmentation

We present in the next 3 figures, the spectrogram and time-domain plot of a signal (‘sa1.wav’), along with its MSALCR (‘Mean smoothed ALCR’). Each figure consists of 3 plots, which are as follows:

- upper plot: spectrogram of the signal, with TIMIT labels superimposed
- middle plot: $E[n]$ of the signal, with ALCR boundaries marked
- lower plot: signal waveform, with ALCR boundaries shown
Figure 10: Comparison of manual and ALCR boundaries over 1-2 sec of signal
Figure 11: Comparison of manual and ALCR boundaries over 2-3 sec of signal
Figure 12: Comparison of manual and ALCR boundaries over 3-4 sec of signal
Graphical Analysis of ALCR based Segmentation

- Segments obtained with ALCR method correspond to what is *visually obvious*
- Vowels, dipthongs - picked up for their high energy
- Nasals, stops - picked up for low energy
- Fricatives - picked up for high frequency content
Which phonemes are hard to pick?

- Semivowels and consonant clusters are hard to pick
- In /g/r/iy/s/iy/ (2.05-2.45s), the /r/ is part of a consonant cluster and is missed
- In /d/aa/r/k/ (1.25-1.45s), /a/ and /r/ got merged into a single segment
- In /w/aa/dx/er/ (2.95-3.25s), /er/ semivowel is missed
- In /ao/l/ (3.4-3.5s), 2 phonemes got merged
Maximum Fischer Ratio Approach to Segmentation

- F Ratio measure is generally defined from a clustering perspective
- We extend that notion for our segmentation problem
- Having defined an analogous F Ratio, we use a Dynamic Programming approach to find that segmentation which will maximize F ratio for a given no. of segments
- We search for that optimal no. of segments that maximizes the Fischer ratio
Fischer Ratio is a popular discriminant measure for a multi-class problem.

Let a dataset \( \{y\} \) be clustered into \( c \) classes.

Let \( m \) represents the global centroid.

Let \( Y_i \) denotes the \( i^{th} \) cluster, containing \( n_i \) number of data points, whose centroid is given by \( m_i \).

The within and between scatter matrices, denoted by \( S_w \) and \( S_b \), are then defined, based on which Fischer ratio \( F \) is computed.
Introduction to Fischer Ratio Measure (contd..)

\[ S_w = \sum_{i=1}^{c} \sum_{y \in Y_i} (y - m_i)(y - m_i)^T \]  \hspace{1cm} (6)

\[ S_b = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^T \]  \hspace{1cm} (7)

\[ F = \text{trace}(S_b)/\text{trace}(S_w) \]  \hspace{1cm} (8)
analogy between segmentation and clustering problems

clustering problem
(data points arranged in multi-dimensional space)

segmentation problem
(segments should be arranged sequentially in time)

inter-cluster distortion
(to be measured between any 2 segments for clustering problem)

inter-segment distortion
(to be measured only between 2 consecutive segments)

intra-cluster distortion
(for each individual cluster)

intra-segment distortion
(for each individual segment)

Figure 13: Analogy between Clustering and Segmentation Problems
Significance of F ratio & Related Analogy

- Higher the F ratio, higher is the discriminability
- We can draw an analogy between clustering and segmentation problems (difference: segments are placed sequentially in time)
- Inter-segment distortion measure \( \approx \) between matrix scatter, (measure of inhomogeneity between consecutive segments)
- Intra-segment distortion measure \( \approx \) within matrix scatter, (intra-segment homogeneity measure)
**Definition of Intra & Inter Segment Distortions**

Consider a dataset of $N$ frames, $\{X_i\}_{i=1}^N$, to be partitioned into $c$ segments, with segment means: $\{\mu_i\}_{i=1}^c$, segment boundaries: $\{b_i\}_{i=1}^c$

\[
D_{\text{intra}}(\text{total}) = \sum_{i=1}^c \sum_{n=b_{i-1}+1}^{b_i} d(X_n, \mu_i) \tag{9}
\]

\[
D_{\text{inter}}(\text{total}) = \sum_{i=1}^{c-1} d(\mu_i, \mu_{i+1}) \tag{10}
\]
Consider a dataset of N frames, to be partitioned into M segments. Discriminant ratio $F(M)$ is of the form:

$$F(M) = \frac{D_{\text{inter}}(\text{total})/M}{D_{\text{intra}}(\text{total})/N}$$ (11)

- For a given M, we segment $\{X_i\}_{i=1}^N$ such that $F(M)$ is maximized.
- Then, we vary M over a range, depending on N and our prior knowledge of minimum and maximum segment length.
- We find that optimal M ($M_{\text{opt}}$) which maximizes $F(M)$.
- Experiments done with 16 dim MFCC vector, with analysis window=20ms, overlap=10ms.
Finding Optimum No. of Segments for sx307.wav, which has 21 TIMIT segments

Figure 14: Fischer ratio based measure to find optimum no. of segments in sx307.wav (same legend holds good for both subplots)
Finding optimum no. of segments for sa1.wav which has 41 TIMIT segments

Figure 15: Fischer ratio based measure to find optimum no. of segments in sa1.wav (same legend holds good for both subplots)
Estimating No. of Segments from $F(M)$ vs $M$ curve

- In $F(M)$ vs $M$ plot, $F(M)$ tends towards saturation but no proper maximum exists; we try to track onset of saturation to determine $M_{opt}$
- As $F(M)$ has a very rough contour, we smooth it to remove local discontinuities and track global nature
- Smoothing is done by polygonal approximation of order 3, 4, 5 and 6 and also by median smoothing
- $M_{opt} = \min i : \{F_{sm}(i) \geq \alpha F_{sm}(i + 1) \& \& F_{sm}(i) \geq \alpha F_{sm}(i - 1)\}$
  $\alpha = 1$, for pure maxima, $\alpha \approx 1$, in our case
Performance Evaluation of F ratio based Measure

- For sx307.wav, No. of segments = 21 (from manual labeling)
- On smoothing F(M) using polygonal approximation, of order 3,4,5 and 6 we get 22, 21, 20 and 20 segments respectively
- For sa1.wav, No. of segments = 41 (from manual labeling)
- On smoothing F(M) using polygonal approximation, of order 3,4,5 and 6 we get 40, 39, 40 and 40 segments respectively
- We repeat experiment on 15 sentences: accuracy=±4 segments
Analysis of Results obtained through F-Ratio Measure

- For speech, **homogeneity** criterion holds good for most phonemes, as suggested by spectrogram and temporal plots.
- However, for semivowels, when coupled with vowels, **heterogeneity** from one unit to another is often indiscernible.
- For diphthongs, a single unit is often not entirely **homogeneous**.
- Thus, from a pattern discriminability perspective, $M_{opt}$ need not be equal to no. of linguistically transcribed phonemes (though values should lie close to each other).
Figure 16: Plot of manual (with TIMIT phonemic labels) and automatic segment boundaries superimposed on spectrogram of ‘sa1.wav’ (1-2 sec)
Figure 17: Plot of manual (with TIMIT phonemic labels) and automatic segment boundaries superimposed on spectrogram of ‘sa1.wav’ (2-3 sec)
Figure 18: Plot of manual (with TIMIT phonemic labels) and automatic segment boundaries superimposed on spectrogram of ‘sa1.wav’ (3-4 sec)
Plans for Future Work

- Segment Vocoder for very low bitrate speech coding: We can use our segmentation schemes in place of ML segmentation, a commonly used front-end feature in most applications.
- Speech Segmentation and Coding using time-compact basis functions: Technique followed is “Temporal Decomposition of Speech”.
- Subword Based Isolated Word Recognition: the subword units needed for training can be obtained through our segmentation techniques.
Thank You!