Acoustic-Phonetic Analysis of Fricatives for classification using SVM Based Algorithm

Alex Frid and Yizhar Lavner
Department of Computer Science,
Tel-Hai College, Israel

IEEE 26th Convention of Electrical and Electronics Engineers in Israel,
Eilat, November 2010
Phonemes Classification

Classification of phonemes is the process of assigning a phonetic category to a short section of speech signal.

Classification vs. Spotting

/s/ /sh/ /f/ ...
/s/ /p/ /ay/ /d/ /e/ /r/

Frid-Lavner (IEEEI 2010)
Possible Applications

• A key stage:
  – spoken term detection
  – continuous speech recognition.

• In developing technologies for the hearing impaired
  – Many hearing impaired suffer from reduced perception in the high frequency range
  – Fricatives are characterized by a relatively large amount of spectral energy above 2.5kHz.

Frid-Lavner (IEEEI 2010)
The unvoiced fricatives group

The Fricatives group:
/f/; /th/; /s/; /sh/; /v/; /dh/; /z/; /zh/

Unvoiced Fricatives
/f/; /th/; /s/; /sh/

Sibilants
/s/; /sh/

Non-Sibilants
/f/; /th/

Voiced Fricatives
/v/; /dh/; /z/; /zh/

Sibilants
/z/; /zh/

Non-Sibilants
/v/; /dh/

Labiodental
/f/

Linguadental
/th/

Alveolar
/s/

Palatal
/sh/

Frid-Lavner (IEEEI 2010)
Phoneme analysis

• The fricatives were analyzed and various features (in time and spectrum domain) are computed.
Phoneme analysis (cont’)

/she/

Amplitude

Time (Seconds)

Frequency (Hz)

Amplitude

Time (Seconds)

Frequency (Hz)

Frid-Lavner (IEEEI 2010)
The selected features

- More than 40 features analyzed
  - Selected features for classification:

<table>
<thead>
<tr>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral peak locations</td>
</tr>
<tr>
<td>Spectral Rolloff</td>
</tr>
<tr>
<td>Spectral Centroid</td>
</tr>
<tr>
<td>Band Energy Ratio</td>
</tr>
<tr>
<td>Zero Crossing Rate (ZCR)</td>
</tr>
<tr>
<td>Standard Deviation, Skewness and Kurtosis of ZCR</td>
</tr>
<tr>
<td>Mel Frequency Cepstral Coefficients</td>
</tr>
<tr>
<td>Lacunarity β parameter</td>
</tr>
<tr>
<td>Gammatone Maximal Peak Frequency</td>
</tr>
<tr>
<td>Spectral Deformation and Width</td>
</tr>
</tbody>
</table>
Features analysis

- **Spectral Peak Locations**
  (using the LPC spectral envelope and peak picking)

- **Spectral Rolloff (0.5)**

  \[
  \sum_{k=0}^{f_r} M_t(k) = p \cdot \sum_{k=0}^{K-1} M_t(k)
  \]

  \(M_t(k)\) is the magnitude of the Fourier transform
Features analysis (cont’)

- **Spectral Centroid**
  \[ S_t = \frac{\sum_{k=0}^{K-1} M_t(k) \cdot k}{\sum_{k=0}^{K-1} M_t(k)} \]

- **Zero Crossing Rate (ZCR)**
  \[ ZCR = \frac{1}{2} \sum_{n=1}^{N-1} \left| \text{sgn}(x[n]) - \text{sgn}(x[n-1]) \right| \]
Features analysis (cont’)

• Gammatone Maximal Peak Frequency

\[ S_w = \sqrt{\frac{M_2}{M_1}} - \left( \frac{M_1}{M_0} \right)^2 \]

\[ SD = \sqrt{\frac{(M_2 / M_0)}{(M_1 / M_0)}} \]

(Mn is the n’s spectral moment)

• Spectral Deformation and Width
Features analysis (cont’)

• Standard Deviation, Skewness and Kurtosis of zero-crossings
Dimension reduced view of the features
Classification using SVM

A Support Vector Machine maps non-linear separable data to higher dimensional space and performs separation in that space.
Classification using SVM (cont’)

• The kernel ‘trick’
  – Kernel functions transforms features to a linearly separable space.
  – In this work a radial basis function was used.
Classification using SVM (cont’)

• The training stage
  – An equally probable set of phonemes selected
  – Each phoneme divided into 8ms, 50% overlapping windows
  – 15 dimensional feature vector extracted from each window
Classification using SVM (cont’)

• Grid search example
  – At first a logarithmic scale was applied
  – Then a linear space applied on selected locations from the first step

Logarithmic scale grid search

Finer linear scale grid search around maxima points

Frid-Lavner (IEEEI 2010)
Classification Experiments

Separation by phonetic group

Framing

Filtering

Feature extraction

SVM 1

Sibilants

/s/, /sh/

SVM 2

Non-Sibilants

/f/, /th/

SVM 3

/s/  /sh/  /f/  /th/
Classification Experiments (cont’)

“Single step” separation

- Framing
- Filtering
- Feature extraction

SVM

/s/
/sh/
/f/
/th/

Frid-Lavner (IEEEI 2010)
Results

• The tests performed on TIMIT db.
• 432 randomly chosen phonemes were used for the training stage.
• Grid Search was used to find the best parameters.
• 11,848 phonemes from 8 different English dialects was used for the test stage.
• Cross validation procedure was used to prevent ‘over-fitting’.
Results cont’

Binary classification by phonetic group

<table>
<thead>
<tr>
<th>Sibilants vs. Non-sibilants</th>
<th>/s, sh/</th>
<th>/f, th/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/s, sh/</td>
<td>99.04</td>
<td>0.96</td>
</tr>
<tr>
<td>/f, th/</td>
<td>3.85</td>
<td>96.15</td>
</tr>
</tbody>
</table>

Alveolars vs. Palatals

<table>
<thead>
<tr>
<th>/s/</th>
<th>/sh/</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.04</td>
<td>0.96</td>
</tr>
<tr>
<td>2.88</td>
<td>97.12</td>
</tr>
</tbody>
</table>

Labiodental vs. Linguadental

<table>
<thead>
<tr>
<th>/f/</th>
<th>/th/</th>
</tr>
</thead>
<tbody>
<tr>
<td>85.58</td>
<td>14.42</td>
</tr>
<tr>
<td>17.82</td>
<td>82.18</td>
</tr>
</tbody>
</table>
## Results cont’

### Classification using one SVM

<table>
<thead>
<tr>
<th></th>
<th>/s/</th>
<th>/sh/</th>
<th>/f/</th>
<th>/th/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/s/</td>
<td>96.15</td>
<td>0.96</td>
<td>0</td>
<td>2.89</td>
</tr>
<tr>
<td>/sh/</td>
<td>1.92</td>
<td>97.12</td>
<td>0.96</td>
<td>0</td>
</tr>
<tr>
<td>/f/</td>
<td>0.96</td>
<td>2.88</td>
<td>82.7</td>
<td>13.46</td>
</tr>
<tr>
<td>/th/</td>
<td>5.94</td>
<td>1.98</td>
<td>18.81</td>
<td>73.27</td>
</tr>
</tbody>
</table>

### Classification Using LDA

<table>
<thead>
<tr>
<th></th>
<th>/s/</th>
<th>/sh/</th>
<th>/f/</th>
<th>/th/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/s/</td>
<td>72.67</td>
<td>12.98</td>
<td>5.37</td>
<td>8.96</td>
</tr>
<tr>
<td>/sh/</td>
<td>10</td>
<td>79.39</td>
<td>10.34</td>
<td>0.26</td>
</tr>
<tr>
<td>/f/</td>
<td>8.28</td>
<td>4.64</td>
<td>65.65</td>
<td>21.4</td>
</tr>
<tr>
<td>/th/</td>
<td>16.76</td>
<td>2.25</td>
<td>36.52</td>
<td>44.18</td>
</tr>
</tbody>
</table>
Previous Results

• Previous results\(^1\):
  
  – Do not cover the unvoiced sibilants group separation (/f/ vs. /th/).
  
  – The phonemes for separation were chosen ‘manually’ to achieve the rate.
    
    • /s/ - 88%
    • /f/ + /th/ - 83%
    • /sh/ - 88%

Future Work

- Extend to voiced fricatives (preliminary results will be shown).

<table>
<thead>
<tr>
<th></th>
<th>/z/</th>
<th>/v/+/dh/</th>
<th>/zh/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/z/</td>
<td>94.79</td>
<td>1.04</td>
<td>4.16</td>
</tr>
<tr>
<td>/v/+/dh/</td>
<td>7.29</td>
<td>91.66</td>
<td>1.04</td>
</tr>
<tr>
<td>/zh/</td>
<td>1.96</td>
<td>0</td>
<td>98.05</td>
</tr>
</tbody>
</table>

- Separation of /v/ and /dh/ in progress
- Separation of voiced – unvoiced is relatively easy task.
Thanks!
Final decision step

- On each section, the final decision is carried out using majority vote (which can be replaced by a sliding window later on).