Gaussian Mixture Linear Prediction
Jouni Pohjalainen, Paavo Alku
Department of Signal Processing and Acoustics, Aalto University, Finland

Problem

- Digital speech and audio are subject to variation
  - additive noise
  - channel variation
  - source variation (e.g., speaker-related effects)
- Undesirable variation degrades the performance of signal analysis in applications

Hypothesis

- Desired “target” characteristics of the signal can be emphasized by
  - training a mixture autoregressive model by EM with differently initialized “target” and “non-target” states (components)
  - choosing the all-pole spectrum model that belongs to the “target” state
- Different initializations yield differently focused models
- Three variants of this general method are evaluated

Method

- Gaussian mixture linear prediction (GMLP) signal model:
  \[ f(s_n|s_{n-1}, \ldots, s_{n-p}) = \sum_{i=1}^{J} P_i X(u_n), \quad X \sim \mathcal{N}(0, \sigma^2), \]
  with \( u_n \) given by \( u_n = \beta_{i,k} s_{n,k} + u_{n,1} \), \( 1 \leq i \leq J \)
  - probability of being in the \( i \)th state at time \( n \): \( \gamma_n(i) \)
- Target state initialized by \( 1/(1 - \sum_{i=0}^{p} \beta_{i,k} z^{-1}) = 1/(1 - 0.97z^{-1}) \) (“inverse pre-emphasis”)
- Using \( J = 2 \), three different initializations for the other “non-target” state:
  - GMLP-0: \( a_{i,k} = 0 \), \( 0 < k < p \)
  - GMLP-H: \( a_{i,k} = \max(s_i), a_{i,k} = 0, k > 0 \)
  - GMLP-N: \( a_{i,k} = 0, a_{i,k} = b_k, k > 0 \), where \( b_k \) represent the noise spectrum

Results of Speech Event Detection

- GMM speech detection using 12 MFCCs based on different spectrum analysis methods
- Detection based on GMM log likelihoods averaged over high-energy frames within an analysis window:
  \[ \text{max}(L_{\text{normal speech}}, L_{\text{shouted speech}}) - L_{\text{non-speech}} > T \]
- Analysis window of 5 s with 1 s shift
- Unsupervised selection of high-energy frames
- 24 Finnish sentences spoken using normal and shouted voice by 11 male and 11 female speakers, LOSO cross validation
- Additive factory noise (NOISEX-92)
- Red: significantly better than both FFT and LP

<table>
<thead>
<tr>
<th>Method</th>
<th>factory1 SNR (dB)</th>
<th>factory2 SNR (dB)</th>
<th>EER %, Normal speech only</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFT</td>
<td>-5</td>
<td>-10</td>
<td>-15</td>
</tr>
<tr>
<td>LP</td>
<td>2.5</td>
<td>3.5</td>
<td>19.8</td>
</tr>
<tr>
<td>LP</td>
<td>21.0</td>
<td>20.0</td>
<td>19.8</td>
</tr>
<tr>
<td>GMLP-0 (2 it.)</td>
<td>1.4</td>
<td>1.4</td>
<td>14.4</td>
</tr>
<tr>
<td>GMLP-0 (5 it.)</td>
<td>1.1</td>
<td>1.1</td>
<td>7.6</td>
</tr>
<tr>
<td>GMLP-H (2 it.)</td>
<td>2.2</td>
<td>2.2</td>
<td>13.6</td>
</tr>
<tr>
<td>GMLP-H (5 it.)</td>
<td>0.9</td>
<td>0.9</td>
<td>9.5</td>
</tr>
<tr>
<td>GMLP-N (2 it.)</td>
<td>1.9</td>
<td>1.9</td>
<td>6.0</td>
</tr>
<tr>
<td>GMLP-N (5 it.)</td>
<td>0.8</td>
<td>0.8</td>
<td>9.7</td>
</tr>
</tbody>
</table>

Conclusions

- A new stochastic mixture approach to linear predictive spectrum modeling
- Outperforms conventional methods in
  - objective spectral distortion evaluation (log spectral distance)
  - feature extraction for speech detection in acoustic environment monitoring
- Customizable by initialization to distinguish different aspects of the signal
  - e.g., GMLP-H has also been applied to robust feature extraction in speaker verification under vocal effort mismatch
- A wide range of potential applications