Introduction

- Information on the type of distortion corrupting a signal can be used to inform the choice of appropriate enhancement algorithms.
- Most existing methods focused on detecting a single and specific type of distortion in a signal.
- In [1], we proposed a method to classify four major types of distortion in vowels directly from MFCCs extracted from speech signals.
- Limitations of [1]:
  - MFCCs encode not only distortion in signals, but also other variability (speaker, articulation, and disorder).
  - Distortion classification decision is made by majority vote over all frames, and the computation time increases with increasing signal length.
- In this paper, distortion in variable duration recordings is modeled with a fixed-length, low-dimensional vector.

Distortion Modeling

- Channel variability can be produced artificially by corrupting the clean recording by different types and levels of distortion.
- Method:
  - Fitting a Gaussian mixture model (GMM) to the features of a recording.
  - Assuming that the GMM mean vector of the i-th frame from the s-th speaker can be decomposed as:
    \[ M_{s,i} = m + V y_s + U x_{s,i} + D z_{s,i} \]  

Definitions:
- \( m \) is the channel-independent supervector,
- \( V \) is a rectangular matrix of low rank with high speaker variability
- \( y_s \) is the speaker factor
- \( U \) is a rectangular matrix of low rank with high channel variability
- \( x_{s,i} \) is the channel factor containing channel-related information
- \( D \) is a diagonal matrix describing any remaining speaker variability
- \( z_{s,i} \) is the speaker-specific residual factor
- The factors \( x_{s,i}, y_s, z_{s,i} \) are assumed to be independent of each other and have a standard normal prior distribution.

Estimating the matrices \( V, U, D \), and the vectors \( x_{s,i}, y_s, z_{s,i} \):
1. Train \( V \), assuming that \( U \) and \( D \) are zero.
2. Estimate \( U \) given the estimate of \( V \) and assuming that \( D \) is zero.
3. Estimate the residual matrix \( D \) given the estimates of \( V \) and \( U \).
4. \( x_{s,i}, y_s, z_{s,i} \) are then calculated given the estimates of \( V, U \), and \( D \).

Channel Factor and Subspace Estimation

- The channel factor \( x_{s,i} \sim N(\mu_{s,i}, \Lambda_{s,i}) \) and the channel subspace \( U \) are estimated by applying an EM algorithm.
- In the E-step, using a random initialization of \( U \), the posterior distribution of the channel factor is calculated as:
  \[ \mu_{s,i} = E[x_{s,i} | U] = (I + U \Sigma_N U)^{-1} U \Sigma^{-1} x_{s,i} \]  
  \[ \Lambda_{s,i} = E[x_{s,i} x_{s,i}^T | U] = \mu_{s,i}^T \mu_{s,i} + (I + U \Sigma_N U)^{-1} \Sigma \]  
- In the M-step, the channel subspace is updated by solving the equations:
  \[ U^T \Theta = \Psi \]  

Definitions:
- \( \Sigma \) is a block-diagonal matrix whose entries form the covariance matrix of the \( c \)-th mixture of the UBM,
- \( N_{s,i} = \sum_{j=1}^{L} p_r \) and \( f_{s,c,i} = \sum_{j=1}^{L} N_{r,i} p_r (m_c + V y_s) \) are the zero-and first order statistics for each speaker \( s \), recording \( r \) and mixture component \( c \).
- \( p_r \) is the acoustic features of the \( r \)-th frame
- \( I \) is an identity matrix,
- \( N_r \) is a block-diagonal matrix which its entries are \( \{N_r, N_{r,i}\} I \)
- \( f_{s,c,i} \) is a vector constructed by concatenation of \( f_{s,c,i} \)
- \( y_{s,r} \) is the posterior probability of the \( c \)-th mixture generating \( p_r \).
- \( m_c \) and \( V y_s \), are, respectively, the subvector of \( m \) and the submatrix of \( V \) of mixture component \( c \).
- \( \Theta_r = \sum_{c} N_{r,c} A_{s,c} \) \( c = 1, \ldots, C \)
- \( \Psi_r \) is the \( i \)-th row of \( \Psi \) is: \( \Psi = \sum_{c} f_{s,c,i} \mu_{s,i}^T \)

The Proposed Method

Channel Factor Extraction

- The channel factor \( x_{s,i} \sim N(\mu_{s,i}, \Lambda_{s,i}) \) and the channel subspace \( U \) are estimated by applying an EM algorithm.
- In the E-step, using a random initialization of \( U \), the posterior distribution of the channel factor is calculated as:
  \[ \mu_{s,i} = E[x_{s,i} | U] = (I + U \Sigma_N U)^{-1} U \Sigma^{-1} x_{s,i} \]  
  \[ \Lambda_{s,i} = E[x_{s,i} x_{s,i}^T | U] = \mu_{s,i}^T \mu_{s,i} + (I + U \Sigma_N U)^{-1} \Sigma \]  
- In the M-step, the channel subspace is updated by solving the equations:
  \[ U^T \Theta = \Psi \]  

Experimental Setup

- Database:
  - Parkinson’s voice database (sustained vowels, 750 telephone recordings).
- Distortion Classes:
  - Additive noise (white Gaussian, babble, office ambiance noises)
  - Reverberation (8 different real room impulse responses)
  - Peak clipping (clipping level: 0.3, 0.4, 0.5, 0.6)
  - Coding (6.3 kbps, 9.6 kbps and 16 kbps CELP codecs)
- Acoustic features:
  - 39 dimensional vector (12 MFCCs + frame energy + \( \Delta + \Delta \))
- Distortion Modeling:
  - GMM with 256 mixtures
  - Speaker factor dim.: 0
  - Channel factor dim.: 210
- Classifiers:
  - SVM with RBF kernel
  - PLDA

Fig. 2: Performance of different configuration of the FA model.

Results

<table>
<thead>
<tr>
<th>System</th>
<th>Clean</th>
<th>Noisy</th>
<th>Reverb.</th>
<th>Clipped</th>
<th>Coded</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>55 ± 11</td>
<td>97 ± 4</td>
<td>77 ± 4</td>
<td>82 ± 7</td>
<td>85 ± 9</td>
<td>79 ± 3</td>
</tr>
<tr>
<td>PLDA</td>
<td>100 ± 0</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
<td>20 ± 0</td>
</tr>
<tr>
<td>PLDA + LDA</td>
<td>77 ± 4</td>
<td>98 ± 2</td>
<td>86 ± 4</td>
<td>82 ± 2</td>
<td>93 ± 3</td>
<td>87 ± 1</td>
</tr>
<tr>
<td>SVM</td>
<td>28 ± 18</td>
<td>33 ± 5</td>
<td>31 ± 16</td>
<td>35 ± 14</td>
<td>68 ± 12</td>
<td>39 ± 4</td>
</tr>
<tr>
<td>SVM + LDA</td>
<td>78 ± 3</td>
<td>97 ± 2</td>
<td>87 ± 4</td>
<td>85 ± 2</td>
<td>93 ± 3</td>
<td>88 ± 1</td>
</tr>
</tbody>
</table>

Conclusions

- Distortion in variable duration signals is modeled by a fixed-length, low-dimensional vector which is more suitable for classification algorithms.
- Channel vectors are more robust to small changes in signal characteristics than MFCCs, they are more suitable for distortion classification in pathological voices.

References