Combining Five Acoustic Level Modeling Methods for Automatic Speaker Age and Gender Recognition

Ming Li¹, Chi-Sang Jung⁰ and Kyu J. Han¹,²
¹Signal Analysis and Interpretation Laboratory (SAIL), Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089, USA
²School of Electrical and Electronic Engineering, Yonsei University, Korea
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
(mingli@usc.edu, jctoctos@dsp.yonsei.ac.kr, kyunam@usc.edu)

Introduction
Automatic recognition of paralinguistic information from speech is important.
- Speaker identity, gender, age range, emotional state, etc.
- Guide human computer interaction systems to automatically adapt to different user needs.

Our focus is to enhance the performance of acoustic-level approaches for speaker age and gender identification.

The proposed approach by combining five different acoustic level methods:
1. GMM baseline
2. GMM-SVM mean supervector (stacked means of MAP adapted models)
3. GMM-SVM maximum likelihood linear regression (MLLR) matrix supervector [1]
4. GMM-SVM Tandem posterior probability (TPP) supervector [2]
5. SVM baseline using acoustic level features (including prosodic features)

Contribution:
1. The GMM-SVM mean supervector method is extended by two kinds of supervectors, MLLR supervector and Tandem posterior probability (TPP) supervector.
2. Combining these acoustic level methods at the score level can improve the overall performance.

Methods

GMM-SVM mean supervector
- A two stage framework: map the supervectors into discriminative aGender characterization score vectors (DACS), followed by a back end SVM classifier.

GMM-SVM maximum likelihood linear regression (MLLR) supervector
- MLLR adaptation was performed using UBM model and training set.
- Linear Discriminant Analysis (LDA) was employed for dimension reduction.

GMM-SVM Tandem posterior probability (TPP) supervector
- Given a frame of MFCC feature $x_i$ and the GMM-UBM $\lambda = \{w_i, \mu_i, \Sigma_i\}_j$, $i = 1, \ldots, M$.
- The posterior occupancy probability is calculated as follows:
  $$P(\lambda|x_i) = \frac{w_ip(x_i|\mu_i, \Sigma_i)}{\sum_{j=1}^{M}w_jp(x_i|\mu_j, \Sigma_j)}$$
- Thus the TPP supervector for a T frames segment is:
  $$TPP_{\text{supervector}} = [b_1, b_2; \ldots; b_T]$$
  $$b_j = \sum_{i=1}^{T}P(\lambda_i|x_i)$$

Experimental results
Performance of each method and the fusion approach (MFuse and AFuse) denote manually and automatically (inverse entropy) tuned weight in score level fusion.

<table>
<thead>
<tr>
<th>System</th>
<th>Age &amp; Gender</th>
<th>age</th>
<th>gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBM</td>
<td>OK</td>
<td>WA</td>
<td>51.7</td>
</tr>
<tr>
<td>1. GMM baseline</td>
<td>OK</td>
<td>WA</td>
<td>51.7</td>
</tr>
<tr>
<td>2. GMM-SVM mean supervector</td>
<td>42.6</td>
<td>61.1</td>
<td>65.7</td>
</tr>
<tr>
<td>3. GMM-SVM MLLR supervector</td>
<td>39.0</td>
<td>60.1</td>
<td>61.1</td>
</tr>
<tr>
<td>4. GMM-SVM TPP supervector</td>
<td>37.8</td>
<td>60.0</td>
<td>51.5</td>
</tr>
<tr>
<td>5. SVM baseline</td>
<td>44.4</td>
<td>60.2</td>
<td>40.7</td>
</tr>
<tr>
<td>MFuse 1+2</td>
<td>46.0</td>
<td>60.3</td>
<td>47.6</td>
</tr>
<tr>
<td>MFuse 3+4</td>
<td>40.3</td>
<td>60.3</td>
<td>41.5</td>
</tr>
<tr>
<td>MFuse 1+2+3</td>
<td>50.4</td>
<td>60.5</td>
<td>51.7</td>
</tr>
<tr>
<td>AFuse 1+2+3</td>
<td>52.7</td>
<td>60.5</td>
<td>52.7</td>
</tr>
<tr>
<td>AFuse 1+2+3+4</td>
<td>51.2</td>
<td>51.4</td>
<td>56.2</td>
</tr>
</tbody>
</table>

Confusion matrix (left) and accuracy for different valid speech durations (right) for class age and gender task.

Conclusion and future work
Conclusion:
- MLLR supervector and TPP supervector systems perform well in this task.
- Combining different acoustic level methods can enhance the performance.
Future works:
- investigating the GMM-SVM Constrained MLLR supervector method
- combining other prosodic or phonetic level methods

References

Acknowledgement
This work was supported in part by National Science Foundation and United States Army.