**Abstract**

**Background:**
- Singing is a popular performance in entertainment
- Singing varies across styles of singing (genres, languages and cultures)
- Professional music teachers can determine singing quality
  - Listening to hours of songs is a tedious and time-consuming task
- We establish a system to estimate singing quality based on acoustic features, and lip and eye movements

**Data Preparation and Proposed System**

**Database**
- Audiovisual data from videos downloaded from a video sharing website
  - Each video has a duration between 5 and 15 seconds
  - 96 auditions for an American TV talent singing show
- Most candidates sang pop genre
- Participants are not professional singers at the time of the audition
- From different states and cities in The United States
- The ground truth or each candidate is provided by the judges
  - “Qualified” candidates move to next phase
  - “Nonqualified” candidates leave the show

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualified</td>
<td>25</td>
<td>30</td>
<td>55</td>
</tr>
<tr>
<td>Nonqualified</td>
<td>21</td>
<td>20</td>
<td>41</td>
</tr>
</tbody>
</table>

**Audiovisual Feature**
- **Audio:**
  - 12D Mel-frequency cepstral coefficients (MFCCs) with first- and second-order difference (36D in total)
- **Video:**
  - 17 landmarks around lips; 10 landmarks around each eye
  - Calculate the lip and eye areas as visual features

**Classification Scheme**
- Classification on each frame:
  - Logistic Regression linear classifier
  - Naive Bayes non-linear classifier
  - k-NN classifier

**Fusion Scheme**
- Concatenate audio and visual feature
- Fusion applied on each frame

**Recognition Result with Unimodal Feature:**

**Recognition Result with Bimodal Feature:**

**Recognition Result Analysis:**
- K-NN has best performance for both audio and visual features
- Systems with audio features outperform the ones with visual features
- System with audio + eyes + lips features has the best performance

**Recognition Result between Genders:**
- Male
  - 78.25%
  - 76.24%
  - 85.88%
- Female
  - 84.67%
  - 76.24%
  - 82.47%

**Conclusions**

**Conclusion and Future Work:**
- We performed classification of singing skill based on audio, lip and eye features
- It is observed that the performance can be improved (up to 2% absolute) when eyes and lips features are added
- Fusing eyes and lips features provides complementary information
- Other features, e.g. Gabor filter feature, can be incorporated into current system
- The work can be applied to automatic singing skill assessment system