Assessment and Classification of Singing Quality Based on Audio-Visual Features

Marigona Bokshi, Fei Tao, Carlos Busso and John H.L. Hansen

Center for Robust Speech System (CRSS),
The University of Texas at Dallas, Richardson TX

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

	Male	Female	Total
Qualified	25	30	55
Nonqualified	21	20	41

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

Nonqualified singer's samples

Qualified singer's sample

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

Experimental Evaluation

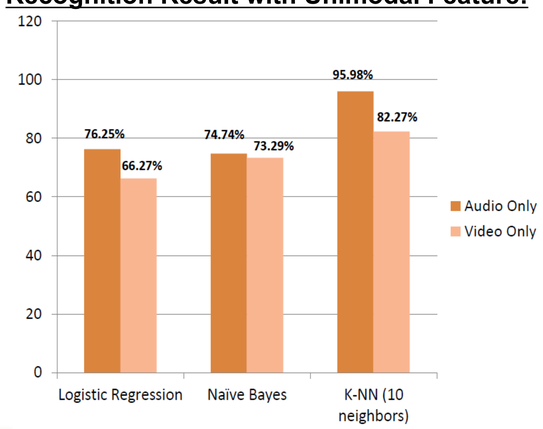
Recognition Task Setting:

- 10-fold cross validation
- Three classifiers are trained
- Unimodal and bimodal features are evaluated

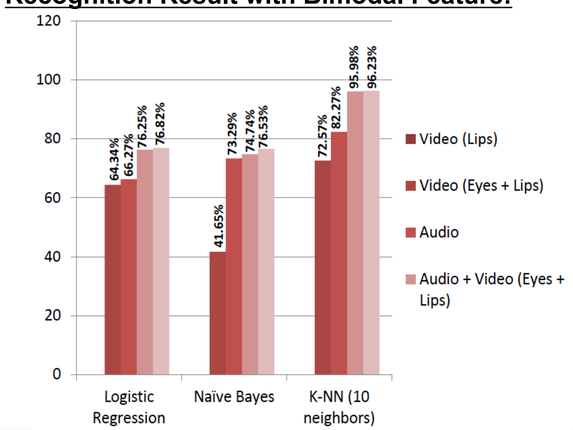
Recognition Result Analysis:

- 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

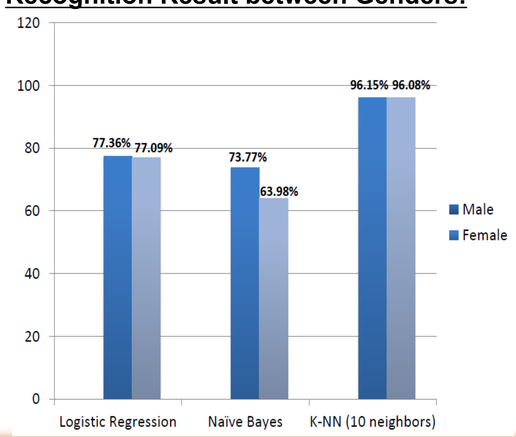
Recognition Result with Unimodal Feature:



Recognition Result with Bimodal Feature:



Recognition Result between Genders:



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