Lip Reading For Song Transcription

Author:Xinghui He , Supervisor: Dr. Jon Barker

Department of Computer Science

Objectives

This project aims at exploiting Audio-Visual Automatic Speech Recognition systems for song transcription with traditional statistic model, HMM-GMM.

- Design Audio Front End
- Design Visual Front End
- Integrate Audio-Visual Feature
- Build Audio-Visual Automatic Speech Recognition System In Music

Introduction

Lip-reading is a technique of understanding speech by observing people's lip movement. And lip-reading recognition technology is a technology that integrates machine vision with natural language processing since it contributes to let machine understand language by analyzing speaker's lip movement. For mankind, singing is an artistic way to express speech, and music is an artistic product of expressing minds. For ASR, recognition in music is more difficult since the same word has various expression in different songs. Moreover, the sound of instrument in song's background also increase the coefficients of difficulty in recognizing. Therefore, it is necessary to take advantage of visual information to recognize the voice in music.

ACOMUS

Acoustic Cover Music Corpus[1] is composed by audio and video isolated segments of the singing parts (utterances) in acoustic covers from amateur artist extracted from YouTube.

- **a**udio: 1 channel , 16kHz sample rate, wav
- **v**ideo: yuv420p, h264, MPEG-4
- 200 songs with single guitar instrument.
- 40 songs with single piano instrument.
- A balance distribution of artist gender.
- No restriction of the music style.
- Songs and artist can have repeated samples.
- In the English language.

Visual Front End Design

1. Face and facial part detection using dlib

Figure 1: Detected face: cover artist M020

2. Estimate affine transform model
3. Estimate lip regions and Extract ROI

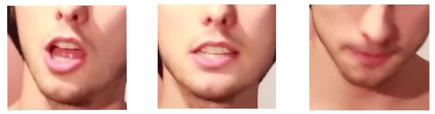
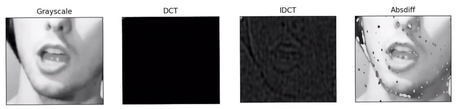


Figure 2: M020_132_01_0101.00.027.mp4: frame001 ,frame089 , frame152: 219x193

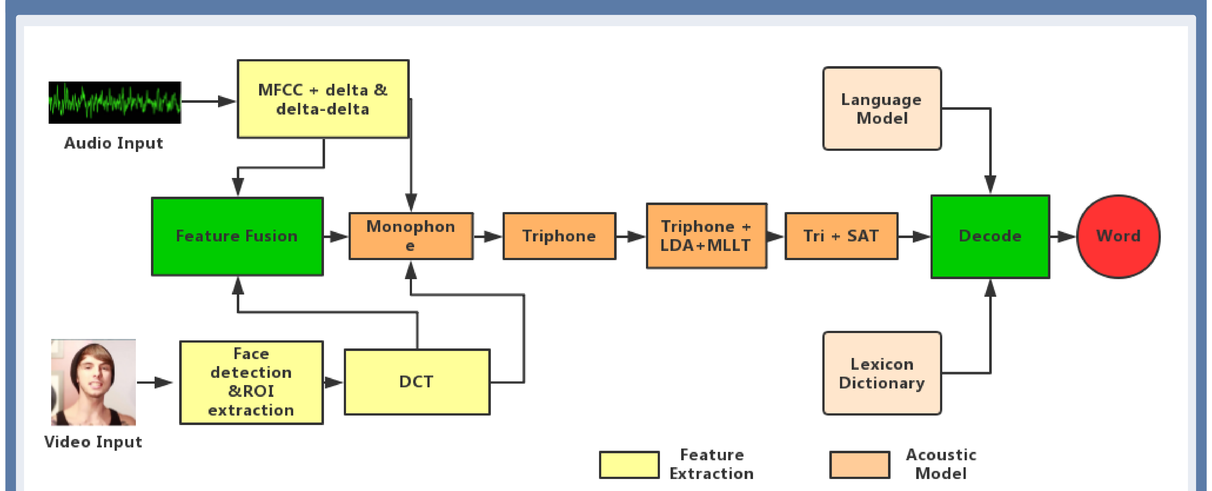
4. Convert to DCT



Bad Lip reading

Bad Lip Reading[2] is a YouTube channel,work to recompose films, TV shows, songs, sports, and political news stories by overdubbing humorous vocal that matches the lip movements of targets. It shows that, even audio-only front end system is weak in robust noise, visual-only front end system is also not reliable. Here are two videos, one is the original MV called **'youŕe beautiful'** from famous singer **James Blunt**, and the other is from this channel.

System implementation in Kaldi



Challenge

- Too much Noise (Instrument Sound) in audio background
- Too much useless sense for shotting instrument in videos
- Multi-faces in one video
- Artists do not frontally face to the camera
- Barriers like microphone or glasses shaded the lip and face
- Adapt visual features into kaldi format and get fusion with audio features

Experimental results

System Front End	Average WER
Audio-only ASR	93.94
Visual-only ASR	
AV-ASR	

Table 1: Tri-phone system

Conclusion

The WER result of audio-only ASR system is really bad. It may caused by the instrument background as noise and the limited size of our music corpus. Besides, as the videos showed in bad lip reading section, it makes a point that lip cues are ambiguous. We can not expect the visual-only system works well. Therefore, it is necessary to develop a AV-ASR system. However, the audio-visual feature integration is a vital problem for our feature level fusion's HMM-GMM model ASR system.

References

Y/[Gerardo Roa Dabike.
Automatic speech recognition in music, 2016.

YO[Bad lip reading, 2011 + present.
frrnq8--cl,ugignchg_,mpe-ugig-@_b]Jgn]Pc_bgle