A Multi-Stream Audio-Video Large-Vocabulary Mandarin Chinese Speech Database

Luhong Liang¹, Yu Luo¹, Feiyue Huang², Ara V. Nefian¹

¹System Technology Labs, Intel Corporation
Santa Clara, CA, 95052, U.S.A
{lu.hong.liang, yu.luo, ara.nefian}@intel.com

²Tsinghua University
Beijing, 100084, China
hfy01@mails.tsinghua.edu.cn

Abstract

In this paper we present the acquisition and content of a multi-stream audio-visual large-vocabulary database in Mandarin Chinese. The database consists of 17,000 utterances spoken by 225 people and captured by a set of seven cameras and 12 microphones. We also provide the label files that describe the end points of the utterances and the script files that represent the actual pronunciation of speech. The database introduced in this paper is can be used in audio-visual speech recognition (AVSR) for both large-vocabulary and small tasks, microphone array based speech recognition, audio-visual speaker identification and 3D face modeling.

1. Introduction

With the increase in the computational complexity of recent computers, audio-visual speech recognition (AVSR) became an attractive research topic that can lead to a robust solution for speech recognition in noisy environment. The AVSR systems [1] require large training and evaluation sets and building a diverse database of large volume is a fundamental step towards facilitating research in this area. Capturing large audio-visual databases is however challenging due large storage requirements and limited availability of subjects. The CMU database [2] consists of 100 video sequences of 10 speakers, each video covered 78 isolated English words. The DAVID database [3] consists of 31 speakers captured over five sessions including digits, alphabets, vowel-consonant-vowel syllable utterances, and some command words used in video conference. The database presented in [4] consists of 1250 isolated confusable words and 1250 connected letter four-tuple utterances, spoken by 50 subjects. The XM2VTS database [5] includes digital video quality sequences of 295 speakers over four sessions captured at one month intervals. The database is designed for speaker identification, but was also used in AVSR research [6]. The CUAVE database [7] includes over 7,000 utterances of both connected and isolated digits. Besides video sequences of 36 speakers, it also includes videos of 20 pairs of speakers. In [1] the IBM ViaVoice™ audio-visual database was introduced, which consists of 24,325 utterances spoken by 290 subjects, covered 10,500 English words. It was used to set up the first English large-vocabulary AVSR system, and the front-end features of the database were available.

While all the above databases are captured in English, in this paper we describe an audio-visual database to facilitate AVSR research on Mandarin Chinese, a language described by non-standard keyboard characters [8] and spoken by almost one fourth of population of the world. In addition, our database includes several audio and video data streams, which enables the research on multi-stream audio-visual fusion techniques and multi-view- or 3D model -based techniques for audio and visual noise reduction.

The database presented in this paper consists of audio-visual data of 225 speakers over two sessions: Session 1 includes 7 streams of video from multiple views and 12 streams of audio data captured by close-talking and in-distance microphone and microphone array, with 483 utterances repeated 15 times by all the speakers. Session 2 includes one video stream of frontal view and 12 streams of audio data, covering more than 17,000 utterances repeated 5 times or more by all the speakers. In addition, label files that mark the start and end points of all the utterances and the script files according to the actual pronunciation of speakers are also provided. The database has been used in our research on Chinese digital string AVSR and 3D face modeling, and can be used in other areas related to speech processing and audio-visual information fusion.

The paper is organized as follows. In the next section we define the database specification. Section 3 introduces the data capturing system. Section 4 describes the methods of data compression and labeling. Section 5 describes the potential usage of the database. Section 6 presents the conclusion of this work.

2. Database specification

The main design criteria of the database described in this paper is to create a multi-stream audio-visual data corpus of
The speaker was framed including the shoulders and head and recorded reading the scripts displayed in a LCD in front of the subject. Since there has been some database focusing on the illumination variety such as [11], we utilized normal lighting in the whole data capturing system. Recognizing it is hard to design artificial noise that covers several application scenarios and it is common to utilize additive noise in AVSR experiments, we recorded all the data in a clean acoustic environment. Section 3.1 and 3.2 gave the details of the data capturing system.

Considering the requirement of large-vocabulary AVSR research, we selected a set of Mandarin Chinese scripts built by Intel for LV-ASR research. The script set has more than 17,000 utterances in total, in which more than 3,750 Chinese characters appear. 81.1% of the Mandarin Chinese tri-phones (without tone) are covered, and 40% of the tri-phones occur more than 10 times. Among the scripts we also selected a subset that covers all the mono-phones (without tone) and has a relative small mean square error of the frequency of the mono-phone occurrence.

The database acquisition commenced with a population of 225 speakers of age 20–40. All of them are Mandarin Chinese native speakers without heavy local accents. The speaker was also chosen so that there is an even representation of male and female individuals. Several shots were designed for each speaker, in which the speaker read the given scripts or make several facial expressions, with capturing devices recording the audio and video data. Session 3.3 gave the details.

3. Capture apparatus and procedure

3.1 Video capturing system

The basic audio-visual stream was acquired using a Sony DCR-VX2000E 3CCD miniDV camcorder mounted on a tripod horizontally pointing to the speaker’s head. The camcorder recorded video (720×576, 25fps) to miniDV tapes and saved to the computer using the IEEE 1394 interface. Besides the camcorder, 6 Sanyo VCC9572P cameras with 12 mm lens working in PAL mode (768×576, 25fps) were also employed to simulate two kinds of scenarios. As shown in figure 1, camera #1, #2, and #6 were mounted in a vertical plane pointing to the speaker in order to simulate future desktop or laptop with cameras embedded into the corners of the monitor. Leveraged with the requirements of 3D face modeling [14] and the limitations in cost and in storage space, 3 more cameras (#3, #4, and #5) on tripods were placed in right side around the speakers in order to simulate the applications in vehicle where there may be cameras around the driver to obtain videos of different views. All the cameras and camcorder were placed about 1.5 meters away from the speaker, so that the head and shoulders of the speaker could be framed and the head of speakers appeared large enough in videos. Figure 1 showed the frames captured.

In order to generate the normal lighting condition on the speaker, three 150W solar lamps were utilized, where two lamps were placed at the left and right side in front of the speaker respectively, and another lamp was at the right side behind the speaker. The illuminations and the positions of the lamps were adjusted so that strong reflection did not appear on the glasses of the speaker.

3.2 Audio capturing system

The data was captured in a room of size 16.7ft (length) × 10.5ft (width) × 8.2ft (height), with the sound insulation materials affixed on the wall. We placed all the computers outside the room to keep the level of the acoustic noise below that of a well-designed conference room.

Besides the camcorder, two 8-channel microphone array (MA) developed by Intel were also utilized in audio capturing. Each MA has 8 microphones collocated linearly with same intervals, providing 8 streams of synchronized 48Hz 16bit audio data via USB interface. As shown in figure 1, MA #1 was mounted at the underside of camera #1 pointing to the mouth of the speaker. As for MA #2, only 3 channels were used to connect to a closing-talking microphone, a throat microphone and a microphone in-distance respectively. In total, there were 11 streams of audio data captured by MAs, besides the audio stream captured by the camcorder.

All the computers that control the cameras and MAs were connected via Ethernet. A client-server software was developed to control all the devices in the computers and provide the monitoring interface that facilitates the data collection work.
3.3 Data collection procedure

The data for each speaker was recorded in two sessions. The first session consists of shorter shots recorded using the seven video and 12 audio channels described in the previous section. In the second session longer shots were recorded using only the camcorder and the MAs.

As shown in Figure 2, session 1 includes 8 speaking shots and one expression shot and each shot is about 20 sec long. The scripts consist of 483 utterances that were designed to cover all the mono-phones and had a small mean square error of the frequency of the mono-phone occurrence. This part of scripts were divided into 15 groups, each of them consisting of 25–35 utterances. In shots 1-8 each speaker read one group of utterances, so that the scripts were covered 15 times by all the 225 speakers. In shot 9, the speaker made 6 expressions such as happy, sad, angry, fear, disgusted and surprised.

In session 2, we recorded 8-9 speaking shots of 5 minutes each. All the scripts of more than 17,000 utterances were divided into 40 groups, 400–500 utterances each, by one speaker. Therefore these scripts were spoken by all the subjects more than 5 times. Moreover, from speaker 186 to 225, 30 utterances of digital strings and from speaker 206 to 225 about 40 command sentences for digital home application were added into shot 9 in order to enhance the training data for the small and medium sized task such as digital string and command-and-control.

Before data recording the lens focus and the white-balance of all the cameras were calibrated, and these parameters together with the positions of cameras, camcorder, solar lamps and MAs were kept constant during the data capturing period. In addition, we have also recorded images of an 8×6 Chess Board and a Color Target everyday in order to provide data for camera calibration and chromatic rectification.

4. Postprocessing and labeling

4.1 Data Compression

All the video sequences were compressed to MPEG-2 files with bit-rate of 9.8Mbps (data captured by camera) or 10Mbps (data captured by camcorder, together with audio data) respectively. All the audio data were stored in uncompress WAVE format.

4.2 Data labeling

We also built utterance label files and script files for each speech shot according to the audio data captured by the close-talking microphone. All the start and end points of the utterances were manually labeled in millisecond to facilitate the data segmentation as well as the research on speech detection. In order to obtain the start and end points in other audio streams, the algorithms described in [12] were utilized to estimate the time delay between the close-talking data and other streams, so that the labels could be easily mapped to those streams captured by camcorder, MA #1 and other microphones.

Since the occasional vocalized mistakes made by the speaker were not corrected or removed in data capturing, copies of original scripts for each speech shot were made and manually corrected in terms of the actual pronunciation in speech. As shown in Figure 3, in the final script file each utterance corresponds to one line with two parts. First part denotes the Chinese characters with labels that show the mistakes such as insertion, deletion and substitution with different symbols. The second part represents the Pinyin (the phonetic notation system of Mandarin Chinese) corrections that represent the actual pronunciations.
5. Potential usage

The database presented in this paper, mainly designed for large-vocabulary tasks, can also be used in small and medium sized tasks such as digital string and command-and-control were also considered. We are currently using the database described here for speaker independent audio-visual continuous speech recognition (AVCSR) using coupled hidden Markov models (CHMM)[6], for the definition of the Chinese phoneme and viseme [13] and for the improvement in the visual front-end of the AVSR system[10]. Future use of the database will include the 3D modeling of the faces using the approach described in [14].

Besides AVCSR and 3D face modeling, the current database can be used in the microphone array based speech recognition, audio-visual speaker identification and other audio-visual speech enhancement techniques. Preliminary experiments in utilizing the audio-visual data together with the utterance labels in audio-only and audio-visual speech detection showed encouraging results.

6. Conclusion

In this paper we introduced a multi-stream audio-visual large-vocabulary Mandarin Chinese speech database. The database is designed for several research areas such as audio-visual speech recognition for both large and small vocabulary tasks, microphone array based speech processing, 3D face modeling and multi-stream speech information fusion.

Future work in data collection will include the capture of audio and visual noise which is important for the in-depth evaluation of any audio visual speech processing system. At the application level we are currently considering medium-sized speech recognition tasks beyond command-and-control.

Acknowledgements

The authors would like to thank all the engineers at Intel China Research Center in Beijing involved in the development of the database. We would like to give our special thanks for the comments and suggestions regarding the design of the database to Xiaobo Pi, Xiaoxing Liu, Ying Jia, Yiming Zhang, Yongge Hu, Zhaorong Hou, Jian Li, Jia Sheng, Ming Yang, Wei Ning, Xingwu Li and Liang Wang. And the partners from Tsinghua University included: Prof. Guangyou Xu, Prof. Dingji Shi, Huijun Di and Qiang Wang.

References