Computers and Electrical Engineering 47 (2015) 35–50

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Multiple camera in car audio–visual speech recognition using phonetic and visemic information

Astik Biswas [a,*], P.K. Sahu [a], Mahesh Chandra [b]

[a] Department of Electrical Engineering, National Institute of Technology, Rourkela, Odisha, India
[b] Department of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, Jharkhand, India

ARTICLE INFO

Article history:
Received 4 December 2014
Revised 12 August 2015
Accepted 13 August 2015
Available online 31 August 2015

Keywords:
AVICAR corpus
AVSR
Speech recognition
AAM
Phoneme
Viseme

ABSTRACT

This paper presents a phonetic and visemic information-based audio–visual speech recognizer (AVSR). Active appearance model (AAM) is used to extract the visual features as it finely represents the shape and appearance information extracted from jaw and lip region. Consideration of visual features along with traditional acoustic feature has been found to be promising in the complex auditory environment. However, most of the existing AVSR systems rarely faced the visual domain problems. In this work, a real world multiple camera corpus audio visual in car (AVICAR) is used for the speech recognition experiment. Texas Instruments and Massachusetts Institute of Technology (TIMIT) corpus sentence portion is used to study the performance of bimodal audio–visual speech recognizer. To consider “Mc-Guruk” effect, acoustic and visual models are trained according to phonetic and visemic information, respectively. Phonetic–visemic AVSR system shows significant improvement over phonetic AVSR system.

1. Introduction

Automatic speech recognition (ASR) has a significant role in the field of computer vision and robotics. In the recent decades, a wide range of applications of ASR systems for human–computer communication has motivated the researchers. Thus, it is essential to improve the performance of speech recognizer under real world adverse conditions. Nowadays car voice navigation system is one of the most popular applications of speech recognition. A reliable voice navigation system can provide extra safety to drivers by minimizing the chance of driver distraction. However, the success of this existing systems is restricted to relatively controlled environments or under quite condition. The system is severely affected by the ambient background noise such as traffic noise, engine noise or wind noise. On the other hand, we humans can often compensate uncertainty in speech information by integrating other sources of speech information such as the visible body or face gestures of the speaker. Therefore, it is necessary to look for other sources of complementary information that could minimize these problems. As visual articulators such as lip and jaw are rigid to sound noise thus inspection on visual articulators (speech-reading) can provide a feasible solution in the realistic automotive environment. The other applications of visual speech recognition (VSR) are communication with hearing impaired people, smartphone/tablet applications, and speech recovery from corrupted/mute movie clips.

The notable progress of audio–visual speech recognition (AVSR) was achieved in the recent years, and continuous researches are going on to develop more robust AVSR [1–4]. In spite of this progress, we are facing some real time visual domain issues to implement a robust real-world AVSR system. However, most of the existing visual modality speech recognition systems have

[*] Corresponding author. Tel.: +918018358059.
E-mail addresses: astikbiswas@live.com (A. Biswas), pksahu@nitrkl.ac.in (P.K. Sahu), shrotriya@bitmesra.ac.in (M. Chandra).

http://dx.doi.org/10.1016/j.compeleceng.2015.08.009
0045-7906/© 2015 Elsevier Ltd. All rights reserved.

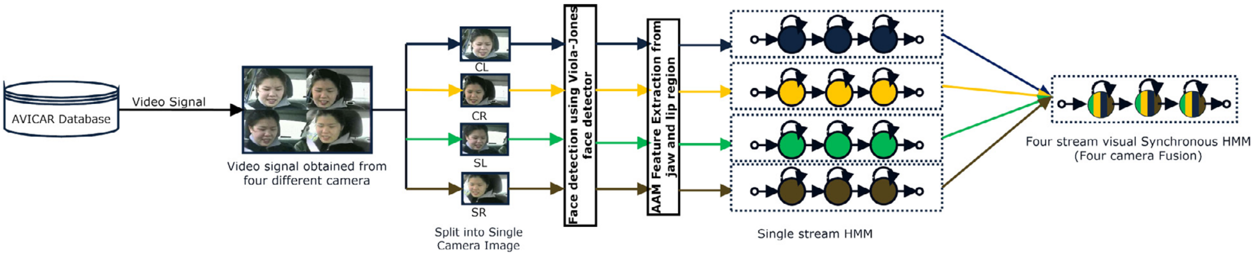
AVICAR Database
Video Signal
Video signal obtained from four different camera
CL
CR
SL
SR
Split into Single Camera Image
Face detection using Viola-Jones face detector
AAM Feature Extraction from jaw and lip region
Single stream HMM
Four stream visual Synchronous HMM (Four camera Fusion)

Fig. 1. Overview of forming four-stream visual synchronous Hidden Markov Model (SHMM) .

been developed by assuming that speakers keep their head in steady position. In the real time scenario, this is very difficult to maintain the head under unmoved condition. The other visual constraints are the light condition, poor resolution, view angle, head position, occlusions, etc. [4,5]. Thus, a better solution is needed to address this issues and to find a possible solution to develop a flexible AVSR system where speakers will have more liberty to move their head naturally. It is not possible to have all movement and pose related information with a single camera. The multiple cameras can be a possible solution to this problem by capturing speaker's natural movement-related information. Some researchers have used multiple camera protocols to improve the flexibility of AVSR. Estellers and Thiran [6] have used multi-camera protocol to implement multi-pose lip reading AVSR system. They have tried to overcome the non-ideal visual effect such as pose change where the speaker does not maintain a frontal view with respect to the mounted camera. However, experiments were conducted in the controlled environment with synthetic conditions where other visual disturbances have not considered. Recently Navarathna et. al [7] reported a notable research on multiple camera AVSR. They have worked with a real world in car database (AVICAR) to address the real world visual domain disturbance. Availability of standard multiple camera audio–visual database is the main burden to researchers in this field. However, there is one standard challenging multiple camera audio–visual corpus AVICAR is available [8]. AVICAR database was prepared in a car cabin environment. This database was designed in such a systematic way, that it could cover different types of driving scenario. The variety of types of data makes it a challenging corpus to evaluate the performance of new AVSR systems. Navarathna et.al [7] proposed a framework to make full use of four camera stream to build five stream (four visual and one acoustic) AVICAR AVSR system. They have shown that the performance of real-world AVSR system significantly differs compared to the ideal laboratory controlled AVSR systems. Poor results indicate that further researches are needed to improve the performance of real world environment AVSR system.

We humans often use visual and acoustic information to perceive speech. This effect is known as the "*Mc-Gurk effect*" [9]. They have demonstrated that sometimes perceived bimodal (audio and visual) speech can be completely different from single modal speech. As an example, in visual domain humans can easily distinguish and perceive bilabial consonants /p/, /b/, and /m/ from their velar and alveolar counterparts /k/, /d/, and /n/, by their distinctive places of articulation [10]. On the other hand, some English phonemes (e.g., /k/ and /g/, /p/ and /b/) are created by changing the tongue position , where mouth movement cannot be noticed visually [10,11]. In simple words some phoneme pairs are having the same appearance in the visual domain. However by using the both modality we can perceive the sound correctly. Thus, it is important to train the visual recognizer according to the visual speech unit (viseme) and acoustic recognizer according to the speech sound unit (phoneme) for better performance.

In this paper, Here we are motivated to extend the research [7] to enhance the performance of AVICAR AVSR by employing robust visual features that are rigid against visual disturbance. However, to the best of our knowledge, there has been no work reported on bimodal AVICAR speech (TIMIT Sentence portion) recognition with detailed phoneme and viseme score analysis. Discrete cosine transform (DCT) features are not very robust against visual disturbance, resulting poor performance in the outdoor environment. Here we use statistically powerful AAM [12–15] to model mouth portion and extracted active appearance model (AAM) visual features to train the visual recognizer. The AAM combines the shape and textural variations of objects to form the statistical model. Later this model is fitted to target image by a gradient-based searching and fitting technique. The reliable and simple framework of AAM makes it widely used technique in many areas of computer vision. AAM features are extracted, and single stream Hidden Markov Model (HMM) is formed for each camera. A series of VSR experiments are carried out to find the relative contribution of side and central faced camera. Overview of forming four stream visual synchronous Hidden Markov Model (SHMM) is illustrated in Fig. 1. Finally, four stream visual modality is combined with single stream acoustic modality by using of late integration technique. The performance of AAM is compared with DCT visual features. Significant improvement is achieved with AAM features over DCT features. Traditional Mel frequency cepstral coefficient (MFCC) [16] is used as the acoustic feature extraction technique.

The rest of paper is organized as follows: Section 2 gives a brief overview of AVICAR corpus and organization of database used in this experiment. Section 3 describes the acoustic and visual feature extraction. Audio–video synchronization is also described in this section. Section 4 gives an idea of audio–visual speech modeling, training, and performance evaluation metrics. Experimental results and discussions are given in Section 5. Finally Section 6 draws the conclusion.

2. AVICAR database evaluation protocol

AVICAR database is a great approach to meet the real-world challenges while evaluating speech recognition metrics. The AVICAR database was prepared by the researchers of the University of Illinois. This multi-channel database was developed in