Audio Visual Automatic Speech Recognition in Vehicles

Rajitha Navarathna, David Dean, Patrick Lucey, Sridha Sridharan

Speech, Audio, Image and Video Technology,
Queensland University of Technology,
GPO Box 2424, Brisbane 4001, Australia.

{r.navarathna, d.dean, p.lucey, s.sridharan}@qut.edu.au

Abstract

Acoustically, car cabins are extremely noisy and as a consequence, existing audio-only speech recognition systems, for voice-based control of vehicle functions such as the GPS based navigator, perform poorly. Audio-only speech recognition systems fail to make use of the visual modality of speech (eg: lip movements). As the visual modality is immune to acoustic noise, utilising this visual information in conjunction with an audio only speech recognition system has the potential to improve the accuracy of the system. The field of recognising speech using both auditory and visual inputs is known as Audio Visual Speech Recognition (AVSR). Continuous research in AVASR field has been ongoing for the past twenty-five years with notable progress being made. However, the practical deployment of AVASR systems for use in a variety of real-world applications has not yet emerged. The main reason is due to most research to date neglecting to address variabilities in the visual domain such as illumination and viewpoint in the design of the visual front-end of the AVSR system. In this paper we present an AVASR system in a real-world car environment using the AVICAR database [1], which is a publically available in-car database and we show that the use of visual speech conjunction with the audio modality is a better approach to improve the robustness and effectiveness of voice-only recognition systems in car cabin environments.

1. Introduction

Vehicles are an essential commodity in daily living activities and driver distraction is a consequence of the fast paced and complex driving environment. The increase in complexity of in-vehicle navigational and other operating is one of the factor that contributes to driver distraction. Therefore, there is a strong need to reduce driver distraction. The use of voice recognition technology has the potential to provide solutions to this problem. It achieves this by providing voice based control for the operation of such in-vehicle systems. Unfortunately, the robustness and effectiveness of voice recognition systems in the car cabin environment are still poor.

One of the main reason for this is that rely solely on the audio channel for input which corrupted by a number of environmental factors, such as acoustic and engine noise. As most real-world applications involve some type of noise, these voice recognition systems are of limited use in these applications due to their poor performance. These audio-only systems fail to make use of the bimodal nature of speech. Visual information such as lip movements is immune to these acoustic environmental factors. Utilizing this visual information in conjunction with the voice recognition technology has the potential to overcome the problems with audio-only voice recognition systems. The field of recognizing speech using both audio and visual inputs is known as Audio Visual Automatic Speech Recognition (AVASR) [2].

AVASR is by no means a new research field. In actual fact, the first work in the field was conducted over fifty years ago [3] and continuous research in this field has been ongoing for

the past twenty years with notable progress being made. Over this time, the advantages of the visual modality in automatic speech recognition (ASR) systems have been established theoretically and several of the issues involved with AVASR have been resolved. Prototype systems have been built that have demonstrated improved performance over audio-only systems under laboratory conditions. However, the practical employment of AVASR systems which will be useful in a variety of real-world applications such as car environment has not yet emerged.

The majority of research conducted in AVASR has not dealt with real-world environment. The main aim of this paper is to address the effect of the visual aspect for AVASR, in a real-world environment (i.e the car cabin). Section 2 describes the research data. Section 3 presents the AVASR system in a car environment and followed by the experimental results in Section 4. Conclusion and future works are reported in Section 5.

2. Research Data

The AVICAR database is a publicly available in-vehicle speech corpus containing multi-channel audio and video recordings [1]. It was recorded by researchers at the University of Illinois.

The AVICAR database consists of audio and video files for 100 speakers [1]. Most of the speakers are American English speakers and the recorded speech is in English. Each recording session contains speech under five different driving conditions; i.e. idling (IDL), driving at 35mph with windows open (35D) and close (35U), and driving at 55mph with windows open (55D) and closed (55U).

3. AVASR System

The AVASR system can be seen as a combination of an audio-only and visual-only speech recognition system. This section describes the audio-only, visual-only and AVASR systems. Figure 1 shows a block diagram of a complete AVASR system.

3.1 Audio-only speech recognition system

A 9 state left-to-right hidden Markov model (HMM) was trained for speaker-independent speech recognition. Each HMM state was represented using 8 mixtures. The acoustic word models (i.e. sil, zero, oh, one, ..., nine) were trained using 39-dimensional Mel-Frequency Cepstral Coefficient (MFCC) vectors (13 MFCC plus delta and acceleration coefficients).

3.2 Visual-only speech recognition system

An efficient visual front end system which is able to track and locate the speaker's face and mouth region-of-interest (ROI) was developed using the Viola-Jones algorithm [4]. Following the ROI extraction, a feature mean normalisation step was used to remove any redundant information, such as illumination or speaker variances. A two-dimensional separable discrete cosine transform (DCT) was then applied to the mean-removed image. The top 100 features energy components were selected to capture the static information. Subsequently, in order to incorporate dynamic speech information, seven of these neighbouring static feature vectors were concatenated, and were projected via an inter-frame linear discriminant analysis (LDA) step to yield a 40-dimensional "dynamic" visual feature

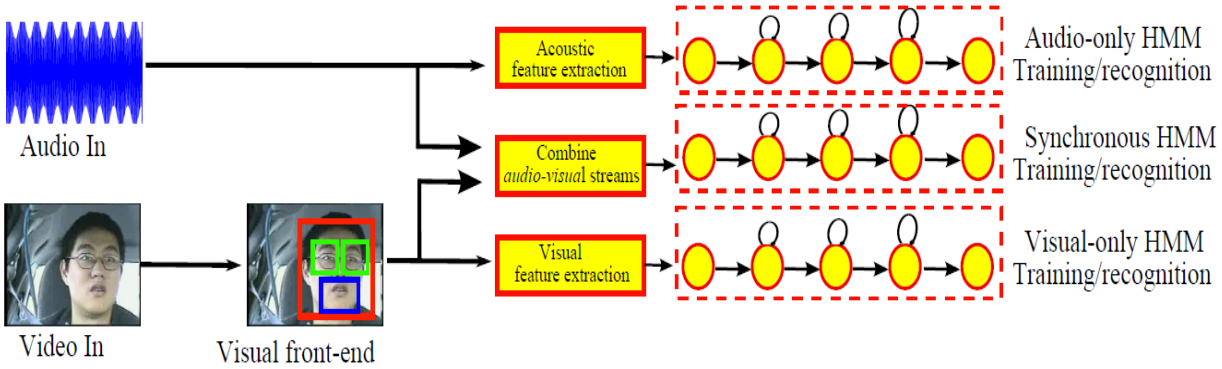

Figure 1: Block diagram of an AVASR system, which is a combination of audio-only and visual-only speech and audio-visual recognition systems

vector, extracted at the video frame rate of 30 Hz. These features were then used to train up a 9 state left-to-right HMM word models using 8 mixtures.

3.3 Audio-Visual speech recognition system

Following the audio and visual feature extraction, the visual features were upsampled to 100Hz using nearest neighbour interpolation to synchronise with the audio signal. This combined feature vector was then used to train a synchronous multi-stream HMM, using 9 states and 8 mixtures for both audio and video. The stream weighting parameter was varied (0-1) depending on the acoustic noise condition).

4. Results

The overall results of the AVASR system are shown in Figure 2. It presents that the accuracy of audio-visual speech recognition is higher compared with the audio-only speech recognition system in any driving condition. All speech recognition results quoted in this paper are HTK-style word accuracies (in %) [5].

4.1 Audio-only speech recognition results

Results are collated by noise condition, with the average results reported. Having the windows open appears to have more significant effect the recognition accuracy than simply increasing the vehicle speed. The word accuracy of windows down condition is less compared with the windows up in the same speed condition, as expected. With windows open, greater decreases in accuracy occur as the speed increases. The Idle condition shows reasonable performance accuracy (64.69%) due to there being less acoustic noise in IDL condition. In the 55D condition the word accuracy is poor (23.68%). This is mostly due to increases in engine noise and wind effects as vehicle speed increases. Note that accuracies across the board are generally low because in the speech recognition system. (ie: The HTK). However as this research is only interested in the relative variations in accuracies, this does not affect the validity of the results.

4.2 Visual-only speech recognition results

In literature [6], all reported visual-only speech recognition results were the same in every noise condition due to the visual data being the same (i.e. only the acoustic noise condition changed due to the additive noise). Our results presented here differ from those as the visual data varies with the noise condition, although they are approximately similar in accuracy

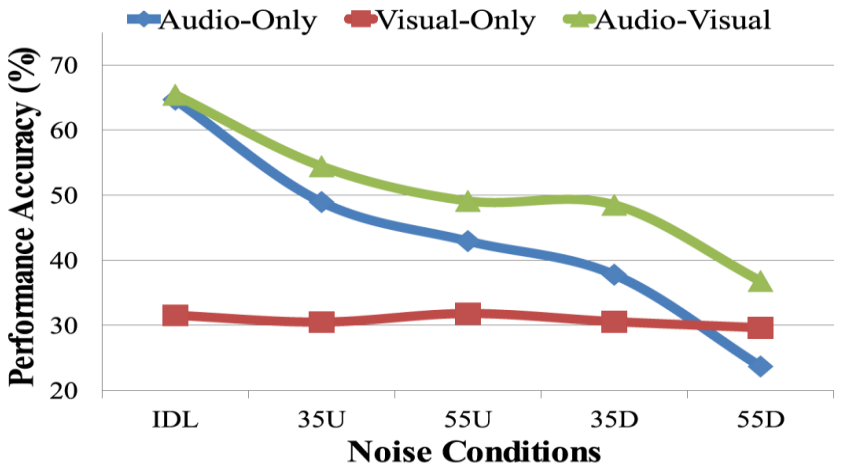

Figure 2: Audio-only, visual-only, audio-visual speech recognition results

(29.63% to 31.89%). In terms of improvement over the audio results, only the 55D condition shows gain performance (improvement of 5.95%).

4.3 Audio-visual speech recognition results

The experiments showed that, in every noise condition the word accuracy is increased compared with the audio-only speech recognition results. In the 55D condition, it shows a greater improvement of 13.12% compared with the audio-only results. This reports that visual information has a greater effect in noisy conditions. In the IDL condition, the improvement is less than 1% compared with the audio-only results. AVASR results for the other noise conditions also showed an improvement over audio-only speech recognition results.

5. Discussion and Future work

This paper describes the speaker independent AVASR system in a real-world environment (i.e car cabin). This paper shows that the use of visual speech conjunction with the audio modality is a better approach to improve the robustness and effectiveness of voice-only recognition systems in car cabin environments.

The broad scope of this research in the AVASR field is to develop a robust visual front-end for the AVASR system. Within this broad scope, this research will focus on: implementing algorithms which will be able to locate and track the driver's face and facial features across many variables (i.e. illumination and head pose); extracting visual features which incorporate both static and dynamic modes; exploring various visual action classifiers and filters which can detect voice activity accurately and efficiently; investigate the effect of 3D lip information for a AVASR system.

6. References

[1]. B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys,M. Liu, and T. Huang, "AVICAR: An audiovisual speech corpusin a car environment," Jeju Island, Korea, 2004, pp. 2489–2492.

[2]. G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in Issues in Visual and Audio-Visual Speech Processing, MIT Press, 2004.

[3]. W. Sumby and I. Pollack, "Visual contribution to speech intelligibility," Journal of the Acoustical Society of America, vol. 26, pp. 212-215, 1954.

[4]. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Computer Vision and Pattern Recognition, 2001. CVPR 2001, vol. 1, pp. 511-518, 2001.

[5]. S. Young, G. Everman, T. Hain, D. Kershaw, G. Moore, J. Odell, V. V. Ollason, D. D. Povey, and P. Woodland, The HTK Book (for HTK Version 3.2.1), Entropic Ltd, 2002.

[6]. P. Lucey and G. Potamianos, "Lipreading using profile versus frontal views," in Proc. Int. Works. Multimedia Signal Process. (MMSP), pp. 24–28, 2006.