VOL. 10, NO. 23, DECEMBER 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

FEATURE-FUSION BASED AUDIO-VISUAL SPEECH RECOGNITION USING LIP GEOMETRY FEATURES IN NOISY ENVIROMENT

M. Z. Ibrahim[1], D. J. Mulvaney[2] and M. F. Abas[1]
[1]Faculty of Electrical and Electronics Engineering, University Malaysia Pahang, Pahang, Malaysia
[2]School of Electronic, Electrical and Systems Engineering, Loughborough University, United Kingdom
E-Mail: zamri@ump.edu.my

ABSTRACT

Humans are often able to compensate for noise degradation and uncertainty in speech information by augmenting the received audio with visual information. Such bimodal perception generates a rich combination of information that can be used in the recognition of speech. However, due to wide variability in the lip movement involved in articulation, not all speech can be substantially improved by audio-visual integration. This paper describes a feature-fusion audio-visual speech recognition (AVSR) system that extracts lip geometry from the mouth region using a combination of skin color filter, border following and convex hull, and classification using a Hidden Markov Model. The comparison of the new approach with conventional audio-only system is made when operating under simulated ambient noise conditions that affect the spoken phrases. The experimental results demonstrate that, in the presence of audio noise, the audio-visual approach significantly improves speech recognition accuracy compared with audio-only approach.

Keywords: lip geometry, feature fusion, audio-visual speech recognition, OpenCV.

INTRODUCTION

People with hearing impairments successfully used visual lip movements to aid the understanding of speech, promises the potential of being able to improve the robustness of automatic speech recognition in environments where substantial audible ambient noise is present. Studies available in the literature have shown a close correlation between the information present in lip movements and speech signals, and consequently the addition of visual information has been a line of investigation followed by a number of researchers in their efforts to improve machine perception of the spoken word [1].

Whenever two modalities are to be considered jointly, the question arises as to the processing stage at which the modalities' information content should be fused. In the case of speech and visual speech modalities, fusion can take place either at the feature level (often termed early integration) or the decision level (often termed late integration).

Decision fusion

In the decision fusion approach, recognition is performed separately for each of the modalities, with the partial results from each sub-process being combined to produce the final classification [2]. As the models may often deliver different partial classification decision outcomes, the decision-fusion approach must provide a suitable method for their ranking and collation. A major drawback of this approach is that fusion itself normally only takes place after the complete utterance has been recognized, which, compared to the feature-fusion approach, can lead to a delay in generating the classification result and so make interactive sessions appear unnatural. The main advantage of this approach is that each of the classifications performed is specific to that modality, allowing specifically tailored methods to be adopted.

Feature fusion

In this paper, the feature fusion approach is adopted and the features extracted for each modality are combined into a common vector to be used by the recognition system. A main drawback with this approach is that, due to the large number of visual features often acquired, the combined feature vector often becomes considerably longer. The advantage of this type of fusion is the straightforward extension of techniques already developed for audio-only speech recognition to include the visual aspects, although in a practical implementation, the modalities need to be synchronized and interpolation used to correct for the different frame rates. In order to reach satisfactory convergence, a substantial increase is required in both the number of training vectors and the training time of the recognition models. In the literature, this problem is known as 'the curse of dimensionality' [3], [4]. The contribution of this paper is twofold. Firstly, experiments have been carried out to demonstrate that the geometrical features established in this work have information content that is highly relevant for the recognition task when classification is carried out in the presence of acoustic noise. Secondly, the investigation of a propose system applied to digit recognition under noisy conditions is performed and the results showed that the recognition of individual digits exhibits a performance that depends substantially on the magnitude of the movement of the lips required in its articulation.

METHODOLOGY

The software used in this work was developed using Microsoft Visual C# 2010 [5] and utilized both the open source image processing library, OpenCV [6] and the Hidden Markov Model Toolkit (HTK) speech processing