International Journal of Innovative
Computing, Information and Control
Volume **6**, Number **2**, February **2010**

ICIC International ©2010 ISSN 1349-4198
pp. 711–723

MANDARIN AUDIO-VISUAL SPEECH RECOGNITION WITH EFFECTS TO THE NOISE AND EMOTION

Tsang-Long Pao[1], Wen-Yuan Liao[2], Yu-Te Chen[1] and Tsan-Nung Wu[1]

[1]Department of Computer Science and Engineering
Tatung University
No. 40, Zhung-Shan N. Rd., Taipei City, Taiwan
tlpao@ttu.edu.tw; d8906005@ms2.ttu.edu.tw; g9606005@ms.ttu.edu.tw

[2]Department of Computer Science and Information Engineering
DeLin Institute of Technology
No.1, Lane 380, Qingyun Rd., Tucheng City, Taipei County 236, Taiwan
andres@dlit.edu.tw

Received July 2008; revised December 2008

Abstract. *This paper presents a Mandarin audio-visual recognition system dealing with noisy and emotional speech signal. In the proposed approach, we extract the visual features of the lips. These features are very important to the recognition system especially in noisy condition or with emotional effects. In this recognition system, we propose to use the weighted-discrete KNN as the classifier and compare the results with two popular classifiers, the GMM and HMM, and evaluate their performance by applying to a Mandarin audio-visual speech corpus. The experimental results of different classifiers at various SNR levels are presented. The results show that using the WD-KNN classifier yields better recognition accuracy than other classifiers for the used Mandarin speech corpus.*

Keywords: Audio-visual recognition, Feature extraction, Gaussian mixture model, K-nearest neighbour, Hidden Markov model, Weighted-discrete KNN

1. Introduction.

Automatic speech recognition (ASR) by machine has been a goal and an attractive research area for past several decades. Most previous ASR systems make use of the acoustic speech signal only and ignore the visual speech cues. They all ignore the auditory-visual nature of speech. The limited performance of ASR in the presence of background noise still restricts its usability. Different researches have been proposed to increase the robustness of ASR systems [3-5,11,21,25].

In recent years, there have been many automatic speech-reading systems proposed, that combine audio and visual speech features [7,11,12,15,20]. For all such systems, the objective of the audio-visual speech recognizer is to improve the recognition accuracy, particularly in difficult condition. It is well known that the movement of lips plays an important role in speech perception. Thus, for an ASR system, the lip features are believed to be vital especially in noisy condition. Most of the previous researches concentrated on the visual feature extraction and audio-visual fusion problems. Thus, the audio-visual speech recognition is a work combining the disciplines of image processing, visual/speech recognition and multi-modal data. Recent reviews can be found in Mason [5], Poamianos [27], Goldschen [13] and Chen [3,4]. However, most of these researches deal with utterances in English or other languages. In this paper, our focus is on the Mandarin bimodal speech recognition with the audio signal at low Signal to Noise Ratio (SNR).

In addition to study the effects of various Mandarin audio and visual information for speech recognition in noisy condition, we also studied the effects of emotions to the speech