A ROBUST AND REAL-TIME VISUAL SPEECH RECOGNITION FOR SMARTPHONE APPLICATION

MIN GYU SONG\textsuperscript{1}, MD TARIQUZZAMAN\textsuperscript{1}, JIN YOUNG KIM\textsuperscript{1}
SEONG TAEK HWANG\textsuperscript{2} and SEUNG HO Choi\textsuperscript{3}

\textsuperscript{1}School of Electronics and Computer Engineering
Chonnam National University
300 Yongbong Dong, buk-Gu, Gwangju 500757, South Korea
tareq@moiza.chonnam.ac.kr; smg686@icdsp.jnu.ac.kr; beyondi@jnu.ac.kr

\textsuperscript{2}Communication Research Center
Multimedia Lab, IT Center
Samsung Electronics
Suwon, Gyeonggi-do 442600, South Korea
shwang@samsung.com

\textsuperscript{3}Department of Computer Science
Information Center
Dongshin University
Kunjae-Ro, Naju, Chonnam 520714, South Korea
shchoi@dsu.ac.kr

Received November 2010; revised March 2011

Abstract. Visual speech recognition (VSR) is one prospective complementary approach for speech recognition under very noisy environments, especially in mobile phone circumstances. In implementing visual speech recognition on a smartphone, the two main issues of real-time responsiveness and robustness conflict with each other. In this paper we proposed and implemented a robust visual speech recognition system that performs in real-time. First, we devised a robust and fast lip detection method based on eye-detection, which is not vulnerable to changes in illumination. The pair of eyes was determined based on image binarization and a coupled-eye validation method. Then the lip region was estimated by geometric lip candidate detection and k-means clustering. Second, to cope with the problem of lighting-dependent visual speech recognition performance, we combined the previous methods of lip-folding and RASTA filtering and introduced a modified histogram equalization, in which a mapping function was calculated for the first frame image and fixed through the following images. Third, the visual speech recognition system with 32 control words was implemented on a smartphone with code optimization. It was shown to work in real-time with promising results.

Keywords: Visual speech recognition, Lip localization, K-means clustering, Histogram matching, Lip folding, RASTA filter

1. Introduction. With the advent of smartphones in our mobile society, the importance of natural human machine interaction (HMI) is continuously increasing. However, state-of-the-art smartphones still adopt conventional keypads or touchpads for HMI. The touchpad is replacing the keypad or mouse due to the small size of mobile phones. There is no doubt that automatic speech recognizers (ASRs) are installed on smartphones. Nevertheless, ASRs have not been fully used, as typical ASR performance is severely degraded when used in noisy environments. Therefore, ASR has limited commercial applications in games, door locks, call-centers, or remote controls. To cope with the problem of ASR performance degradation in noisy surroundings, three broadly different approaches have