Chapter 8
Enhancing Robustness in Speech Recognition using Visual Information

Omar Farooq
Aligarh Muslim University, India

Sekharjit Datta
Loughborough University, UK

ABSTRACT
The area of speech recognition has been thoroughly researched during the past fifty years; however, robustness is still an important challenge to overcome. It has been established that there exists a correlation between speech produced and lip motion which is helpful in the adverse background conditions to improve the recognition performance. This chapter presents main components used in audio-visual speech recognition systems. Results of a prototype experiment conducted on audio-visual corpora for Hindi speech have been reported of simple phoneme recognition task. The chapter also addresses some of the issues related to visual feature extraction and the integration of audio-visual and finally present future research directions.

1. INTRODUCTION
Speech is a complex signal which has variability not only from one speaker to another, but also a lot of variations within the same speaker. The variations of speech signal within a speaker may be attributed to factors like age, stress, emotional state or biological reasons (such as sore throat, flu, etc.) and is also called intra-speaker variability. Pronunciation of same word differs among people from different geographical backgrounds due to variations in accents. The recent advances in the signal processing algorithms and availability of fast computational machines have enabled practical implementation of speaker independent automatic speech recognition (ASR) systems feasible.

In an ASR system, attempts are made to imitate human speech recognition which has tremendous recognition capabilities; however, the knowledge of exact mechanism of human speech recognition is still limited. Due to this reason, the current ASR can perform equally well as humans in the case of quite background conditions, but their performance degrade severely when there is a
mismatch between training and test conditions. In practical scenario, these mismatch conditions are frequently encountered because of the difference in background conditions in which the speech is to be recognized. Due to this reason realizing an ASR system which matches human speech recognition capabilities under adverse conditions has been a big challenging task.

To achieve robustness in ASR various techniques have been proposed, which could be grouped into the following four categories:

- Robust feature extraction
- Compensation techniques
- Noise filtering during pre-processing
- Audio visual speech recognition

The first approach is based on the extraction of the features that are inherently resistant to noise. The techniques used under this category are RASTA (RASTA) processing (Hermansky & Morgan, 1994), one-sided auto-correlation LPC (You & Wang, 1999) and auditory model processing of speech (Kim, Lee, & Kil, 1999). The assumption made here is that the noise is additive and white with Gaussian distribution. The second approach is based on the compensation model, which tries to recover the original speech from the corrupted speech in the feature parameter domain or at the pattern-matching stage. Methods using the second approach are cepstral normalization (Acero & Stern, 1990), probabilistic optimum filtering (Kim & Un, 1996; Neumeyer & Weintraub, 1994) and parallel model combination (Gales & Young, 1996).

Spectral enhancement techniques like spectral subtraction and Wiener filtering have been used resulting in improved recognition performance. To reduce the effect of noise present the speech signal, robust amplitude modulation, frequency modulation (AM-FM) features in combination with MFCCs have shown considerable error rate reduction for mismatched noisy conditions (Dimitriadis, Maragos, & Potamianos, 2005). Specialized order statistics filters which work on the sub-band log-energies also have been implemented for noise reduction (Ramirez, Segura, Benitez, Torre, & Rubio, 2005). Denoising process based on soft and hard thresholding of wavelet coefficients has also been proposed (Donoho & Johnston, 1995; Mallat, 1998).

Although the above techniques improves the recognition performance, but the improvement is limited to low background noise levels only. It is well documented that humans use visual information from speaker’s lip region for speech recognitions particularly in presence of noise or by people with hearing impairment (Chibelushi, Deravi, & Mason, 2002). A strong correlation has also been reported between face and speech acoustics (Grant & Braida, 1991; Williams & Katsaggelos, 2002).

The idea of using visual information along with the audio for ASR applications started only recently due to the fact that higher computational and storage was required for video processing. Lips are one of the most important regions of interest which is of particular interest. The movement of lips conveys dynamic information complimentary to the information obtained from the acoustic speech. Since this visual information from the lips is not affected by the poor acoustic conditions; hence, its use is logical along with the audio information. Many experimental results have shown that modelling visual dynamics along with the speech, can improve speech and speaker recognition performance not only in noisy environment (in the SNR range of 15dB to -10dB) but also in noise-free environments (Chaudhari, Ramaswamy, Potamianos, & Neti, 2003; Potamianos, Neti, Gravier, Garg, & Senior, 2003).

An audio visual speech recognition system (AVASR) uses both audio as well as visual information and integrates them to achieve rec-