Chapter XI
Hidden Markov Model Based Visemes Recognition, Part I: AdaBoost Approach

Say Wei Foo
Nanyang Technological University, Singapore

Liang Dong
National University of Singapore, Singapore

ABSTRACT

Visual speech recognition is able to supplement the information of speech sound to improve the accuracy of speech recognition. A viseme, which describes the facial and oral movements that occur alongside the voicing of a particular phoneme, is a supposed basic unit of speech in the visual domain. As in phonemes, there are variations for the same viseme expressed by different persons or even by the same person. A classifier must be robust to this kind of variation. In this chapter, the author’s describe the Adaptively Boosted (AdaBoost) Hidden Markov Model (HMM) technique (Foo, 2004; Foo, 2003; Dong, 2002). By applying the AdaBoost technique to HMM modeling, a multi-HMM classifier that improves the robustness of HMM is obtained. The method is applied to identify context-independent and context-dependent visual speech units. Experimental results indicate that higher recognition accuracy can be attained using the AdaBoost HMM than that using conventional HMM.

INTRODUCTION

Brief Review of Research in Lip-Reading

The technique of retrieving speech content from visual clues such as the movement of the lips, tongue and teeth is commonly known as automatic lip-reading.
It has long been observed that the presence of visual cues such as the movement of lips, facial muscles, teeth and tongue may enhance human speech perception (Sumby, 1954). It has also been shown (Petajan, 1984; Morishima, 2002; Adjoudani, 1996; Silsbee, 1996; Tomlinson, 1996; Chen, 1998; Finn, 1988) that the performance of a purely acoustic based speech recognition system will improve with additional input from the visual speech elements, especially when the speech sound is swarmed by environmental noise. Visual speech processing can also be applied to areas such as speaker verification, multimedia telephony for the hearing impaired, cartoon animation and video games.

In 1984, Petajan developed probably the first visual speech processor. In this system, the distance of geometric measures among different mouth shapes was computed for identifying the visual representations of word productions. In 1993, Goldschen extended Petajan’s design by using Hidden Markov Model as the visual classifier. Subsequent researches on implementing visual speech processing/visual-audio integration include Neural Network (Yuhas, 1989), time-delayed Neural Network (TDNN) (Stork, 1992; Bregler, 1995), fuzzy logics (Silsbee, 1996) and Boltzmann zippers (Stork, 1996).

Among the various techniques for visual speech processing studied so far, Hidden Markov Model (HMM) holds the greatest promise due to its capabilities in modeling and analyzing temporal processes. In Goldschen’s system, HMM classifiers were explored for recognizing a closed set of TIMIT sentences based on speech sounds (Goldschen, 1994). In 1990, Welch et al explored audio-to-visual mapping using HMM for building speech-driven models. Silsbee and Bovik (1993) applied HMM to identify isolated words based on sounds alone. Tomlinson et al (1996) suggested a cross-product HMM topology, which allows asynchronous processing of visual signals and acoustic signals. Luettin et al (1996) used HMMs with an early integration strategy for both isolated digit recognition and connected digit recognition. In recent years, coupled HMM, product HMM and factorial HMM are explored for audio-visual integration (Zhang, 2002; Gravier, 2002; Nefian, 2002; Dupont, 2000).


**Visemes and Viseme Classifiers**

The most conspicuous element in visual speech perception is the oral movement. The sequence of movement of the lips for a phonetic sound is regarded as the most representative feature of a viseme (Owens, 1985). It indicates a short period of lip movement repeatable for the same sound. Like phonemes, which are the basic building blocks of sound of a language, visemes are the basic constituents for the visual representations of spoken words. A visual speech recognition system may be designed to recognize words by recognizing the visemes that constitute the words. For notational convenience, we shall identify the visemes by the names of the phonemes they represent. As our focus is on oral movement, we shall refer to the movement of mouth when voicing a particular phoneme as a viseme.

However, the relationship between phonemes and visemes is not a one-to-one but a many-to-one mapping. For example, although phonemes /b/, /m/, /p/ are acoustically distinguishable sounds, they are grouped into one viseme category as they are visually confusable. An early viseme grouping was suggested by Binnied et al (1974) and was applied to some identification experiments (Greenwood, 1997). Viseme groupings suggested by Owens et al (1985) are obtained by analyzing the stimulus-response matrices of the perceived visual signals. The MPEG-4 multimedia standard adopted the same viseme