Recognition of Visual Speech Elements Using Hidden Markov Models

Say Wei Foo[1] and Liang Dong[2]

[1] School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore 639798
eswfoo@ntu.edu.sg
[2] Department of Electrical and Computer Engineering
National University of Singapore, Singapore 119260
engp0564@nus.edu.sg

Abstract. In this paper, a novel subword lip reading system using continuous Hidden Markov Models (HMMs) is presented. The constituent HMMs are configured according to the statistical features of lip motion and trained with the Baum-Welch method. The performance of the proposed system in identifying the fourteen visemes defined in MPEG-4 standards is addressed. Experiment results show that an average accuracy above 80% can be achieved using the proposed system.

1 Introduction

Lip reading, which is also referred to as speech reading, is the technique of retrieving speech content from visual clues. As early as 1970's, researchers had studied the bimodal aspects of human speech. The "McGurk effect" indicated that the perceived sound existed in both audio signal and visual signal[1]. And even earlier, Sumby and Pollack proved that visual clues could lead to better perception of speech especially under noisy environment[2]. However, the ability of lip reading was long regarded as the privilege of our human being because of the complexity of machine recognition. Only in recent years did lip reading become an interested area of multimedia processing due to the development of pattern recognition tools and modern computing techniques. In 1988, Michael Kass *et al* developed snake-based method to dynamically track lip boundaries[3]. Tsuhan Chen and Ram R. Rao studied the audio-visual integration in multimodal communication[4]. Bregler *et al* used the time-delayed neural network (TDNN) for visual speech recognition[5]. The efforts made by the researchers chiefly serve two objectives: i) Providing an informative description for the lip motion, and ii) Designing a sequence recognition algorithm with strong logic capacity. The first task is associated with image processing and feature extraction, whose purpose is to obtain sufficient features for speech analysis. The second task involves configuration and training of certain mathematic tool, e.g. neural network or HMM. However, the progress in both areas is not smooth due to the difficulties in information extraction from lip motion. In many bimodal speech

Y.-C. Chen, L.-W. Chang, and C.-T. Hsu (Eds.): PCM 2002, LNCS 2532, pp. 607–614, 2002.

processing systems, the visual channel only serves for performance enhancement for an existing acoustic speech recognizer. In this paper, we introduce a novel HMM-based subword classifier and investigate the possibility of visual-only speech analysis. The dynamics of lip motion is systematically studied and the HMMs are configured accordingly. Experiment results show that if the HMMs are well tuned, high recognition rate of individual visual speech element can be achieved.

2 Features of Lip Motion

While we are speaking, our lip is driven by some facial muscles to move in continuous 3D space. In most cases, only the frontal projection is processed for speech analysis. The boundary of the lip (2D) or the surface of the lip (3D) can be very complicated under fine resolution. However, not all of the details are helpful for speech reading. For example, a human speech reader knows what a person is speaking from a distance. What he sees is merely a coarse shape. Computer-aided lip reading should also not pay too much attention on the details but focuses on the "skeleton" of the lip, such as the width and height.

Lip motion is chiefly the up-and-down shift of the upper lip and lower lip. The movement of the other parts, e.g. the lip corner, is somewhat subject to it. Such motion is simple at the first glance. However, it is difficult to be applied to speech analysis because of the following reasons.

i) the movement of the lip is slight compared with its geometric measures during natural speaking. For example, if the width of a speaker's mouth is 6cm in its relaxed state, the variance is usually between 5.5 to 6.5cm during speaking. This fact indicates that the statistical features of lip motion concentrate around some stable states.

ii) the movement of the lip varies slowly over time. Compared with the speech signal, which has significant frequency components up to 4kHz, the lip motion is a very low-frequency signal. It indicates that the information conveyed by lip motion is limited.

iii) the basic visual speech elements corresponding to English phonemes, namely visemes, have too many similarities with each other. Most visemes experience the same process during production: starting from closed mouth, proceeding to half-opened mouth and ending with closed mouth again. Such similarity is easy to observe in our daily experience and is reported in many experiments.

iv) the visemes are liable to be distorted by their context. For example, the visual representation of the vowel /ai/ is very different in the words *hide* and *right*. The preceding letter and the posterior letter will both influence the lip states of the studied viseme to certain extent. A viseme will therefore demonstrate polymorphism under different context.

The above factors make it a challenging job for visual-only speech recognition. In the following sections, we will discuss the measures taken in our system to solve some of the issues.