VISUAL SPEECH SEGMENTATION AND RECOGNITION

Carol Mazuera and YingLi Tian
The City College of New York
cmazuer00@citymail.cuny.edu, ytian@ccny.cuny.edu

1. INTRODUCTION

This project is motivated by the difficulties blind people and deaf people have to face in order to be able to communicate effectively with others. In everyday life, an enormous amount of information is communicated with others through speech. If a conversation is held in a noisy environment, visual information comes as a very useful tool in the business of efficiently maintaining that conversation. This hearing and visual task may, easily, be taken for granted by sighted and hearing individuals; the same cannot be said about blind & visually impaired people, and deaf & hard of hearing people. The first can hardly depend on their vision (or none at all) and the second rely solely on it (or partially). These individuals can potentially benefit from the development of speechreading technologies to: help blind people hold a conversation in a noisy environment, and assist deaf people with speech learning.

In this paper, we propose a visual speech recognition system based on the analysis and comparison of lip movements between two pre-recorded speakers. A word utterance of one speaker is evaluated against a word utterance of a second speaker to identify weather both speakers are speaking the same word. Accordingly, the classifier of this scheme is trained by correct/incorrect utterance patterns. The main structure of our proposed system can be divided into two stages: segmentation and recognition. Segmentation performs word fragmentation of a visual speech sequence by identifying frames with moving or neutral lip shapes. Recognition determines whether two speakers are saying the same word or not.

2. METHODOLOGY

2.1. Low-Level Features

To model the lip movement, we compute two types of dynamics-based features: stretch dynamics and point dynamics. Both feature types employ a lip tracker to extract 19 coordinate points to define the outer and inner contours of the lips shape. Stretch dynamics employ only 12 of these points (outer contour). Distances between selected pairs of top and bottom of these points are calculated (35 in total). The 35 distances from each frame are concatenated as the feature representation of stretch dynamics. Unlike stretch dynamics, point dynamics are based on the definite coordinate points not the distance between them, making it susceptible to head motion while speaking, for this reason, rotation and alignment are required. The final representation of point dynamics consists of all 19 points (outer and inner contours), as well as width and upper/lower lips heights, for a final feature vector dimension of 41 for each frame.

2.2. Segmentation

The first step of our speech learning system involves the automated video subdivision of a speech. Our segmentation method is based on the classification of moving lips (utterance) from neutral lips (absence of speech). We use stretch dynamics due to its versatile spatial variation properties; we aggregate the 35 vector elements of stretch dynamics to produce a scalar value, $S_{1,}$ as a representation of lip moving degree for each frame. We use a temporal sliding window on each frame of size $2n + 1$. The lip moving degree values S_1 of each frame within a sliding window are then concatenated as the dynamics representation of current frame for neutral/moving classification. Based on our empirical observations, we chose n to be 60 in our system. SVM with linear kernel is used as the classifier.

2.3. Recognition

After video segmentation, the next step is to recognize correct and incorrect utterances between a pair of speakers. To eliminate speech tempo differences among subjects, we perform temporal normalization on the dynamics-based features. We concatenate the stretch, or point dynamics of the normalized frames as the dynamics-based input feature. The framework for recognition includes two inputs (one per speaker). We take the difference between the two input features as the representation for classification. We employ SVM with RBF kernel as the classifier.

3. EXPERIMENTS

3.1. Dataset

A dataset of five pre-recorded native English speakers is collected to assess the effectiveness of our proposed visual speech system. The dataset comprises 220 videos; each video contains five repetitions of a word, and includes 50 different words, which are chosen based on easiness to be understood by a child, and visual utterance distinction.

3.2. Results

As it can be seen in table 1, subject independent and dependent results are very similar. This observation shows the generalization of our proposed lip movement based visual speech segmentation. A contributor factor is the natural flow of the word uttering process; the lips must present an action pattern from open to close to say a word.

Table 1. Segmentation results

	Accuracy	Precision	Recall
Subject Dependent	88.91	85.07	70.84
Subject Independent	89.78	85.59	71.83

The two experiments present similar results for both dynamics (see table 2). Point dynamics outperformed stretched dynamics by a small margin. This is probably due to the spatial nature of speech modulation, and point dynamics has a stronger spatial background than stretch dynamics.

Table 2. Recognition results

Dynamics	Stretch (S.D.)	Point (S.D.)	Stretch (S.I.)	Point (S.I.)
Accuracy	96.03	97.53	96.09	98.18
Precision	93.14	95.30	95.92	97.99
Recall	97.00	99.17	94.00	97.33

4. CONCLUSION

The visual speech segmentation and recognition methods proposed in this paper achieve state-of-the-art performance in both subject dependent and subject independent experiments, which would ultimately provide an aid to assist the blind & visually impaired and deaf & hard of hearing to effectively communicate with others.

This work is supported by NSF under Award #EFRI-1137172