Visual Speech Segmentation and Recognition Using Dynamic Lip Movement

Carol Mazuera, Xiaodong Yang, Shizhi Chen, and YingLi Tian
Dept. of Electrical Engineering at The City College of New York, CUNY

ABSTRACT

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 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. The structure of our proposed system can be divided into two stages: segmentation and recognition. Segmentation performs word fragmentation of a video sequence by detecting lip movements. Recognition determines whether two speakers are saying the same word or not.. With the help of a lip tracking method, which employs landmark points to define the lip shapes, we extract Dynamic features. We utilize these dynamic features along with Space-Time Interest Points (STIP) to capture lip movements. We evaluate our proposed method on a challenging visual speech dataset and achieve the state-of-the-art results.

METHODOLOGY

❖ Low-Level Features

✧ *Stretch Dynamics*

- Requires lip tracking to follow lip movements (see Fig. 2)
- Extraction of 12 landmark points shaping the outer contour of the lips
- Lip shapes are normalized so the that the distances between lip corners equals 1
- Distances between the top and bottom landmark points are calculated (see Fig. 1)
- 35 distances from each frame are concatenated as the feature representation of stretch dynamics
- Features are resistant to rotation, and they do not require alignment of the lip shapes

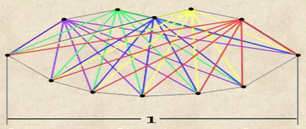

Figure 1: The normalization and computation of stretch dynamics based on distances between selected pairs of landmarks. Each of the five top landmarks has its seven distances that are illustrated in a color for clarification. On each lip shape frame, the width between mouth corners is normalized equal to 1

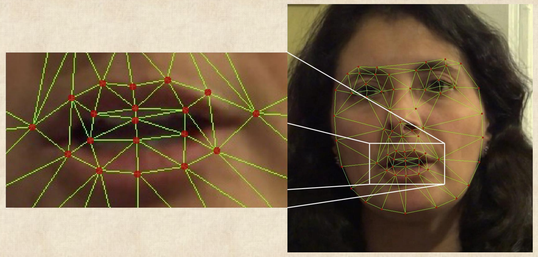

Figure 2: (left) close-up view of the mouth and the 19 landmark points shaping the lips. (right) ASM face tracking exhibiting the 68 landmarks shaping the face of the subject

✧ *Point Dynamics*

- Requires lip tracking to follow lip movements
- Extraction of 19 landmark points shaping the outer and the inner contours of the lips
- Requires rotation and alignment
- Normalization is accomplished by employing a mouth's width and upper/lower heights from a template frame.
- The final feature representation comprises coordinates of 19 landmarks plus the width, and the upper & lower heights of the mouth.

✧ *STIP (Space-Time Interest Points)*

- we employ STIP as the benchmark to model lip movements.

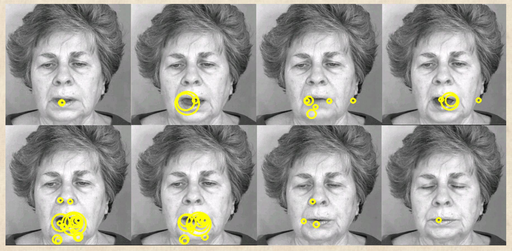

Figure 3: STIP is the product of the change between distinct representative patterns in frames from a video sequence. Here, we show a few video frames of the word "avocado" depicting STIP circle points.

❖ Segmentation

- We use STIP and stretch dynamics features for their versatile spatial variation

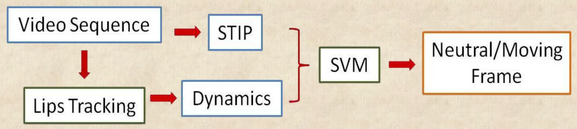

Figure 4: Framework of visual speech segmentation.

❖ Recognition

- Temporal normalization is performed to eliminate tempo variations of the speech among subjects

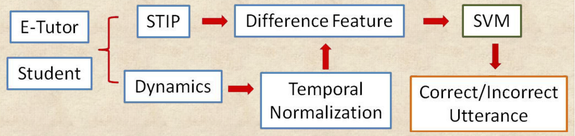

Figure 5: Framework of visual speech segmentation.

EXPERIMENTS AND RESULTS

❖ Dataset

- 220 videos; each video is approximately 500 frames long
- 50 distinct words
- 5 subjects (all are native English speakers)

Table 1: The dataset contains a total of 50 different words, chosen based on easiness to be understood by a child and visual utterance distinction.. There is at least one word beginning with each letter in the alphabet.

Words in our dataset				
Apple	Avocado	Blackberry	Cheese	Cruise
Dishwasher	Dress	Eat	Eggplant	Elbow
Example	Family	Father	Find	Give
Happy	Hello	History	Hospital	Important
Island	Jump	Kangaroo	Kiwi	Laugh
Library	Mother	Music	Notebook	Number
Open	Pineapple	Potato	Present	Question
Respect	Search	Stomach	Together	Tomorrow
Umbrella	Up	Vision	Watermelon	Weather
Window	X-Ray	Yellow	Yesterday	Zebra

Figure 6: Some frames extracted from the video sequence of the word "notebook", Subjects are instructed to start with a closed mouth position (neutral position) before uttering the word shown on the screen, and finalizing with the same neutral position.

❖ Segmentation

- The combination of stretch dynamics and STIP improves the segmentation performance
- Both experiments, independent and dependent, produce similar results

Table 2: Subject independent and subject dependent average results for the 5 subjects in our dataset. 3 different feature modalities are tested: stretch dynamics, STIP, and the combination of the first 2.

	Subject independent			Subject dependent		
Feature	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
Stretch	89.78	85.59	71.83	88.91	85.07	70.84
STIP	91.60	86.75	78.98	91.54	**88.20**	81.05
Combined	**92.77**	**88.64**	**82.01**	**92.85**	87.99	**83.73**

❖ Recognition

- Stretch and point dynamics outperform all feature combinations in both experiments.

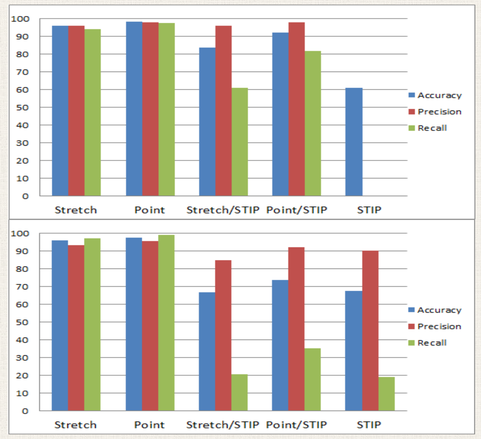

Figure 7 & 8: Subject independent & subject dependent recognition results of different feature combinations.

CONCLUSION

The visual speech segmentation and recognition methods proposed 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.

ACKNOWLEDGEMENTS

This project was supported by the NSF grant IIS-0957016 and EFRI-REM Summer 2012 program, and the City Seeds program