Visual Speech Learning From an E-Tutor via Dynamic Lip Movement-based Video Segmentation and Comparison

Carol Mazuera, Xiaodong Yang, and YingLi Tian
Department of Electrical Engineering
The City College of New York
New York City, NY, USA
cmazuer00@citymail.cuny.edu, {xyang02, ytian}@ccny.cuny.edu

***Abstract*—This paper is motivated by the difficulties that deaf students encounter when learning speechreading and speaking; the skills that enable them to effectively communicate with hearing people. In this paper, we propose a speech learning prototype system based on the analysis and comparison of lip movements of an E-Tutor and those of a deaf student in a video. The main framework of our proposed system can be divided into two stages: lip movement segmentation and speech comparison. Lip movement segmentation fragments the frames of each word from a visual speech video sequence by analyzing the movement and shape of lips. Comparison determines whether a student is producing a correct word utterance or not, this is accomplished by comparing the lip shape and movements according to that of an e-tutor. To model lip movement, we compute two dynamic-based features by using a lip tracking method, which employs landmark points to define lip shapes. We utilize these dynamic features along with Space-Time Interest Points (STIP) to capture lip movements. In order to evaluate the effectiveness of our proposed methods, we collect a visual speech learning dataset consisting of 220 videos and 1100 word utterances. The proposed system achieves promising performances in both visual speech segmentation and visual speech comparison on this dataset.**

Keywords—Visual Speech Learning, Segmentation, Visual Speech Comparison, Dataset Collection, Deaf People

I. INTRODUCTION

According to the National Institute on Deafness and Other Communication Disorders (NIDCD), approximately 2 to 3 out of every 1000 children in the United States are born deaf or hard-of-hearing. Studies indicate that deaf children of deaf parents tend to learn better than deaf children of hearing parents, mainly due to better language communication i.e., using American Sign Language (ASL) when both children and parents are deaf. However, more than 80% of children who are deaf or hearing impaired are born to hearing parents [13]. Compared to deaf children of deaf parents, deaf children of hearing parents face more difficulties in communicating with others and reading because their parents are not likely to be fluent or proficient in sign language [14]. In addition to their parents, deaf children need to communicate with other people with normal hearing, who may know little sign language in daily life. Deaf children who are able to efficiently communicate through speech experience an enhanced inclusion and integration into society. The ability also enhances their future employment opportunities, success in the workplace, independent living, and economic and social self-sufficiency.

Research demonstrates that visual information such as pictures, speech-reading, cued speech, sign language, and other hand gestures are critical for effective literacy education [15, 16]. However, deaf children face the colossal challenge of communicating with hearing people using spoken language. It takes many years of training under the close supervision of a speech/language therapist in order to master the art of speech reading and speaking. Some of the learning techniques make use of the vibratory perception of speech by feeling the movement on throat and face of the tutor as he/she is producing speech [8]. Some of the training techniques involve visual simulation by following the mouth movements of the tutor [8]. These learning techniques are not only costly but also require the constant presence of a tutor. Consequently, a deaf child can only receive a limited training time. Unlike human tutors, a virtual tutor is more accessible due to its absence of time constraints and low acquisition cost.

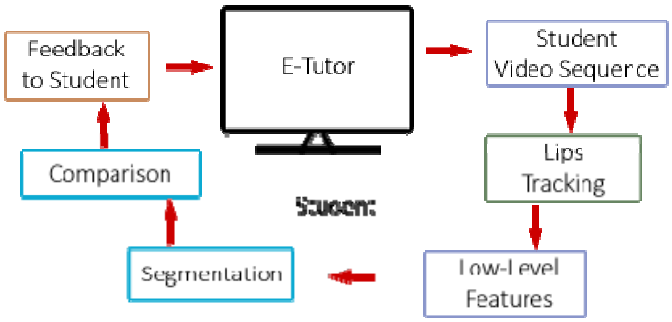


Figure 1. Our proposed visual speech learning framework.

In the absence of hearing, a deaf or hard-of-hearing person heavily relies on his/her vision, i.e., in a sense, hearing through his/her eyes. Visual information plays an important role in the life of these persons when communicating with other people, for this reason, it is not natural for them to make use of common technologies such as "voice recognition" or "voice to text" for the purpose of speech learning; they lack vocal visual cues, and are also very susceptible to background acoustic noise. These individuals can potentially benefit from the development of a visual speech learning system to help them sustain a more dignifying life. There are several such technologies being developed by researchers, which could

potentially assist deaf and hard-of-hearing people with speech learning [1], [6], [7], [10], [11], [12]. Potamianos *et al*, and Matthews *et al*, have used visual information to improve degraded acoustic data for speech recognition with very satisfactory results [6], [7]. Some other research work have proposed visual-based approaches for the speech recognition of single letters, phonemes, and phrases [1], [10], [11], [12]. However, these methods require an extremely extensive database to hold speech words, phonemes, and/or phrases. In addition, video sequence of speech must be segmented in accordance to the format of the database.

In this paper, we propose a more reliable scheme by analyzing the lip movement of a deaf student uttering a word and comparing it, in real time, to those of an on-screen pre-recorded tutor in order to verify the validity of the student's utterance. In other words, we only focus on whether the utterance of a student is correct or not, according to the word spoken by an e-tutor (i.e., a binary classification problem), instead of trying to recognize the word a deaf person is uttering (i.e., a multi-class classification problem). Accordingly, the classifier of this system is trained by correct and incorrect pre-labeled utterance patterns. Our proposed speech learning system requires a computer and a web camera. Fig. 1 depicts our proposed visual speech learning system. A deaf student will follow the pre-recorded e-tutor on the computer screen. The web camera is used to capture the visual utterance of the student. A lip tracking model follows and records the movement of the student's mouth. This data is used to generate low-level features, which are then used to perform segmentation and comparison. Segmentation automatically subdivides the video sequence by identifying the frames which contain the speech being learned by the student. Comparison is employed to verify whether the utterance of the student is correct or not. Lastly, on-screen interactive visuals will provide feedback to the student, based on the student's speech learning performance.

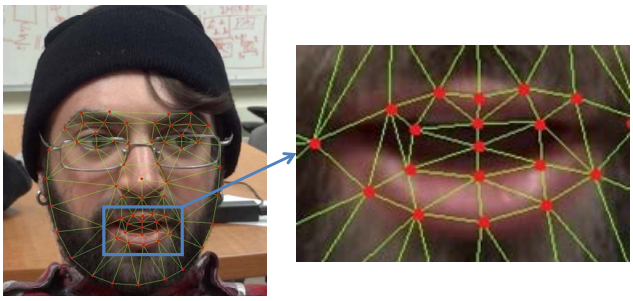
This paper is organized as follows. In Section II, we describe the basic features for the method of dynamic lip movement segmentation and lip reading comparison. Section III and IV explain our visual speech segmentation and comparison structures in detail, respectively. The experimental results and evaluation are presented in Section V. Section VI concludes the paper and discusses our future work.

II. LIP FEATURES EXTRACTION

To implement segmentation and comparison of video sequences, and to find the best feature arrangement performance, we employ various combinations of three different low-level features: stretch dynamics, point dynamics and Space-Time Interest Points (STIP). This section details the structure of each feature.

A. Stretch Dynamics

The stretch dynamics feature requires lip tracking to follow the lip movement when uttering a word. For this task, we employ Active Shape Models (ASM) [2, 9] as the lip tracker. ASM uses prior knowledge of lip shapes in training data, which is simply the concatenation of x and y coordinates of predefined lip landmark points. This model iteratively fits the lip shape, and identifies corresponding lip landmark points in each video frame. In this paper, we use a built-in ASM library [9] trained using 68 landmark points to shape a person's face, as shown in Fig. 2(a). The positions of the 68 landmarks form a shape vector, in which each landmark is represented by its x and y coordinates. As our proposed speech learning system focuses on lip movement, we only keep the 19 landmarks shaping the inner and outer contours of the lips for computing dynamic features, as illustrated in Fig. 2(b). For stretch dynamics, we only employ the 12 landmarks shaping the outer contour of the lips. The distances between selected pairs of top and bottom landmark points are calculated, i.e., 7 distances are computed for each of the 5 top landmarks, as illustrated in Fig. 3. The 35 total distances per frame are concatenated as the feature representation of stretch dynamics.



(a) (b)

Figure 2. ASM based facial landmark point tracking: (a) the 68 landmarks shaping a face and (b) the 19 landmarks shaping the lips (12 outer points and 7 inner points).

We assume that the first five frames in the video sequence are neutral frames. Having this in mind, we then take the average of the five frames, and use this average as our neutral frame for the video sequence. Stretch dynamics feature measures the amount of lips movement deviation from the neutral frame.

In order to eliminate scaling change and variations between lip sizes and shapes of different subjects, as well as to alleviate shape change due to out-of-plane movements (e.g., backward and forward head movements) during speech, a lip shape normalization is conducted so that the width between mouth corners equals 1 (see Fig. 3). The stretch dynamics feature is resistant to mouth shape angular motion, as pair-wise distances are invariant to rotation, which lessens computational cost by not requiring rotation and alignment of the lips.

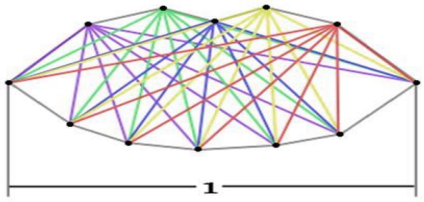


Figure 3. The normalization and computation of stretch dynamics based on distances between selected pairs of landmarks. Each of the five top landmarks has its seven distances; for clarification, they are illustrated in different colors.

B. Point Dynamics

The point dynamics feature also applies ASM to track the lips and capture lip movements. However, unlike stretch dynamics, point dynamics employs all the 19 landmarks, i.e., both outer and inner contours of the lips, as shown in Fig. 2(b). Parallel to stretch dynamics, the average of the first five frames in each video is used as the neutral frame template for the entire video sequence. Point dynamics are highly susceptible to mouth shape variations. In order to handle mouth shape variations among different subjects, we use the width of the mouth, the upper lip height, and the lower lip height from the template frame to normalize other frames from the same video. As for the rotation, we normalize the line connecting two lip corners to a canonical direction (e.g., horizontal). After the two processes, we further align the center of the lip shape to the origin of coordinates. Point dynamics feature represents the lips modulation during speech. The final feature representation of point dynamics consists of coordinates of 19 landmarks, as well as width, upper lip height, and lower lip height, for a final feature vector size of 41.

C. STIP

The Space-Time Interest Points (STIP) [4] is a spatial-temporal feature, which includes two phases: detection (i.e., a feature detector localizes interest points in a spatial-temporal space) and description (i.e., a feature descriptor computes representations of detected points). STIP employs 3D Harris corner detector to identify interest points with large gradient magnitude in both spatial and temporal domains. Histogram of Gradients (HOG) and Histogram of Optical Flow (HOF) are then computed from detected local video volumes as descriptors. Fig. 4 presents detected STIP points on selected frames of a video sequence.

Figure 4. Interest points (yellow circles) detected by STIP in sampled frames of a video with a subject uttering the word "avocado".

A bounding box around the mouth area is set for the extraction of STIP points (not shown in Fig. 4). This ensures that only mouth motion is included in our STIP feature vector.

STIP has been widely used in human action and complex event recognition and detection tasks [5]. In this paper, we employ STIP as the benchmark to model lip movements.

III. Visual Speech Segmentation

The segmentation of each individual word utterance in a video sequence is a prerequisite for further visual speech analysis. The first step of our speech learning system involves the automated video subdivision of a speech. We employ our stretch dynamics feature (described in IIA) for segmentation by summing the difference between the current frame distances and the neutral frame distances as shown in (1); where w represents a distance, i the current frame and N the neutral frame.

$$D_i = \left[\left(w_1^i - w_1^N\right) + \left(w_2^i - w_2^N\right) + \cdots + \left(w_{36}^i - w_{36}^N\right)\right] \qquad (1)$$

The framework of video segmentation is illustrated in Fig. 5. Our segmentation method is based on the classification of moving lips (utterance) from neutral lips (absence of speech). The moving lips indicate that the frames belong to an utterance, while the neutral lips denote absence of speech. A closed-mouth shape may also demonstrate speech; this is due to the versatile spatial variation nature of speech. The first word sequence in Fig. 7 contains an example of a closed-mouth shape during speech. In order to recognize such frames as the moving class as well, we employ a temporal sliding window to incorporate the neighboring frames, which have the open-mouth shapes; for neutral frames (i.e., the frames between separated word utterances), most mouth shapes of their neighboring frames are also closed. Therefore, for each frame, we extend n previous frames and n consequent frames to generate a sliding window with the size of $2n + 1$ frames. The lip moving degree values D_i of each frame within a sliding window are then concatenated as the dynamics representation of a current frame for neutral/moving classification, as shown in Fig. 5.

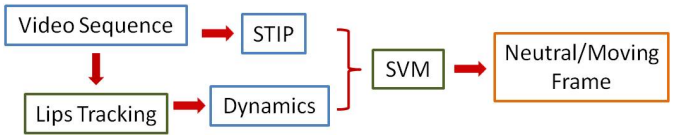


Figure 5. Framework of our visual speech segmentation.

We also employ STIP as an additional feature channel to improve the detection of lip movements around the mouth area. As shown in Fig. 4, the number and appearance of detected interest points surrounding the mouth area demonstrate great difference between moving and neutral frames. Since we only focus on the lip movement, we use the number S_i of detected interest points falling into a bounding box around the mouth area as the second representation of lip moving degree for each frame. Similar to the above case, S_i of each frame in a sliding window are also concatenated as the STIP representation of current frame for neutral/moving classification, as shown in Fig. 5. Based on our empirical observations, we chose n to be 60 in our system.

We select as classifier a Support Vector Machine (SVM). The classifier recognizes each frame into one of two classes: moving (utterance) or neutral (absence of speech). The dynamics and STIP features are concatenated and utilized as input for SVM classification.

SVM utilizes learning algorithms to identify patterns. The algorithm finds an optimal plane which can separate classes with maximum margin. Based on this, SVM can predict the

class of subsequent samples. SVM with linear kernel is used as the classifier for segmentation.

IV. Visual Speech Comparison between Student and E-Tutor

After video segmentation, the next step is to recognize correct and incorrect utterances between the e-tutor and a student. The general framework for comparison is shown in Fig. 6, which includes two inputs: one from the e-tutor and the other from the student. In this paper, we examine five different feature combinations for visual speech comparison: stretch dynamics, point dynamics, STIP, stretch dynamics and STIP in early fusion, as well as point dynamics and STIP in early fusion.

Subjects have their own speaking rhythms: some may articulate speech in a faster fashion than others and some in a slower one. To eliminate these temporal inconsistencies of different word utterances among subjects, we perform a temporal normalization on the dynamics based features to solve this difficulty. In this way, the segmented video clips or word utterances are normalized to the same number of frames. In this paper, we temporally normalize all video clips to 30 frames. We concatenate the stretch or point dynamics of the normalized 30 frames as the dynamics-based input feature. As for the STIP input feature, we use the Bag-of-Words (BOW) model [3] to aggregate HOG/HOF descriptors from a word utterance, we generate our visual dictionary by employing k-means clustering on STIP features. No temporal normalization is performed on STIP, as the pooling process in BOW can naturally handle the temporal inconsistency.

We obtain our final feature vector by finding the difference between the tutor and the student feature vectors, i.e., the pronunciation difference between tutor and student. L2 normalization and scaling to [1 -1] is performed on the final difference feature before sending to the SVM classifier. We employ SVM with RBF kernel as the classifier. The optimal parameters of RBF kernel are obtained by 5-fold cross-validation.

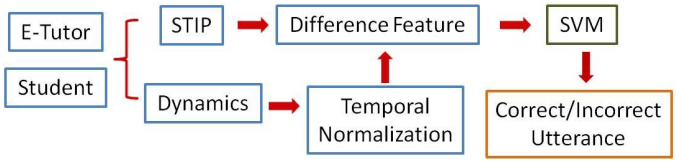


Figure 6. General framework of our visual speech comparison.

V. Experiments and Results

In this section, we talk about the dataset employed in the experiments for our visual speech learning system. We, first explain the evaluation methods used, and then present the results obtained for the two stages of our system: segmentation and comparison. For each stage we perform subject dependent and subject independent experiments. Subject dependent evaluation consists of training and testing based on only one subject, or subject pair (in the case of comparison). Subject independent evaluation employs more than one subject/subject pair for training and testing.

A. Dataset

Figure 7. A sample of frames extracted from the video sequence of the utterance "apple", starting and ending in neutral position.

A dataset of 5 pre-recorded native English speakers is collected to assess the effectiveness of our proposed speech learning system. The dataset is comprised of 50 different words, which are chosen based on easiness to be understood by a child, and by their visual utterance distinctions. The dataset contains at least one word starting with each letter in the alphabet. The words recorded in the dataset are listed in table 1.

These videos are captured at frontal face by an automatic camera with a spatial resolution of 640×480 pixels and a frame rate of 30 frames per second. Subjects are instructed to start with a neutral position (i.e., closed mouth) before uttering the word shown on the screen, and to finalize with the same neutral position. The words are presented to each subject in slideshows of 4 seconds per word. Fig. 7 shows two examples of word utterances starting and ending in neutral position.

TABLE I. The 50 different words in our dataset: 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

This dataset comprises 220 videos, each of which contains 5 repetitions of a word, i.e., 1100 word utterances are included. Each video is approximately 500 frames long, with an average length of 30 frames per word utterance. There are currently 5 subjects in our dataset: two adult females and three adult males. The ground truth marking the beginning and ending of each word utterance in each video is manually labeled. We will make this speech learning dataset public available in the future.

B. Evaluation Metrics

We employ three evaluation metrics to assess the efficacy of our visual speech learning framework. They are: recall, precision and accuracy. Their respective equation is shown in (2), (3), and (4), where *TP* stands for true positive, *TN* for true negative, FP for false positive, and FN for false negative. In our segmentation process, *P* represents a moving frame (utterance) and *N* a neutral frame (no speech). *P* and *N* represent a correct utterance and an incorrect utterance, respectively, in our comparison process.

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

Recall measures the positive frames average that is correctly recognized by the classifier. Precision represents the average number of negative frames classified as positive (a score of "100" signifies that there are zero negative frames classified as positive). Finally, accuracy assesses the total percentage of positive and negative frames correctly identified.

C. Visual Speech Segmentation

We evaluate the performance of our proposed segmentation techniques by implementing subject dependent and subject independent experiments on our collected dataset.

1) Subject Dependent Results: We test our segmentation method by using 30 videos of one subject to train the classifier, and then 20 videos from the same subject to test our segmentation scheme. We perform this same experiment on all five subjects in the dataset. Table 2 shows the average results for all subjects in our dataset.

TABLE II. SUBJECT DEPENDENT SEGMENTATION RESULTS

Feature	Accuracy	Precision	Recall
Stretch Dynamics	88.91	85.07	70.84
STIP	91.54	**88.20**	81.05
Combined	**92.85**	87.99	**83.73**

In general, the results present slightly more moving frames wrongly classified as neutral frames than neutral frames classified as moving frames; this is evident by the lower recall score for all features. Most of the neutral frames incorrectly classified as moving, and vice versa, lie at the beginning and ending of word utterances. The frames in these two areas are particularly difficult to classify due to the mild fluctuation presented by the lip tracking model. This could be possibly improved by enhancing the lip tracking method. As shown in Table 2, the combination of stretch dynamics and STIP is able to improve the segmentation performance.

2) Subject Independent Results: We further perform a subject independent experiment, where 30 videos from one subject are employed to train the classifier, then the trained classifier is tested on 20 videos from each of the other four subjects. We repeat this cycle four times, i.e., each subject is employed once for training. The average results for the three feature combinations are shown in table 3.

As we can see in Table 2 and Table 3, subject independent results are very similar to those of the subject dependent. This observation shows the generalization of our proposed lip movement based visual speech segmentation. This is probably because of the normalization of lip shapes and the duality of motion detection provided by stretch dynamics and STIP. Another contributor is the natural flow of word uttering process: the lips must present an action pattern from open to close to say a word.

TABLE III. SUBJECT INDEPENDENT SEGMENTATION RESULTS

Feature	Accuracy	Precision	Recall
Stretch Dynamics	89.78	85.59	71.83
STIP	91.60	86.75	78.98
Combined	**92.77**	**88.64**	**82.01**

D. Visual Speech Comparison

As in subsection V-C, here we also employ subject dependent and subject independent experiments to examine the performance of our proposed comparison methods.

1) Subject Dependent Results: In the scenario of visual speech comparison, one subject functions as an e-tutor and a second subject as a student. We rotate roles between the 5 subjects, which results in 10 different tutor-student pairs. We employ 90% and 10% of the same tutor-student pair for training and testing, respectevely. Fig. 8 demonstrates the average comparison rates of the 10 tutor-student pairs under five different feature combinations; as discussed in section IV, early fusion is used to combine multiple features.

As shown in this figure, stretch dynamics and point dynamics achieve the best results; the three evaluation metrics score above 90% for both features. The general performance of STIP is significantly inferior to the dynamics-based features. Mainly because the system built upon STIP tends to classify student utterance as incorrect, which result in a very low recall rate. This observation demonstrates that: 1) the lip movement implicitly captured by STIP-based BOW does not discriminate the spatial appearance of an utterance as explicitly as the one modeled by the dynamics-based feature; 2) the temporal order (lost in BOW) is important for comparison; 3) the temporal and spatial normalization helps to reduce intra-class variations. Moreover, the computational cost of STIP is much larger than that of the dynamics-based features, which only involves simple normalization and distance calculation.

The combination of the dynamics-based feature with STIP by early fusion also presents similar results as the STIP feature; the performance of both combinations is deteriorated as well. Several "correct" utterances are classified as "incorrect" utterances, lowering, substantially, the recall percentage.

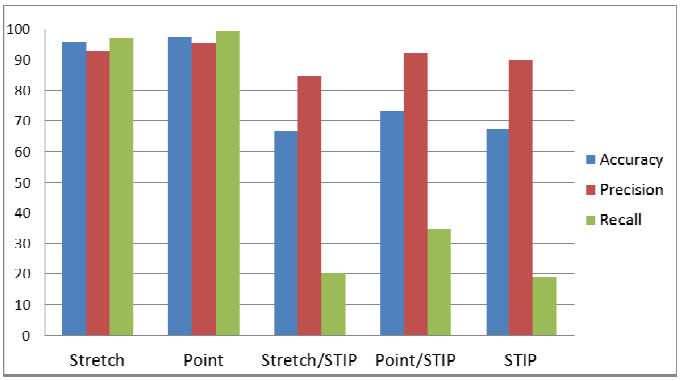


Figure 8. Subject dependent comparison results of different feature combinations.

2) Subject Independent Results: A subject independent experiment is also performed to evaluate our proposed visual speech comparison method. As the case in subject dependent, 90% and 10% of subject pairs are used for training and testing, respectively. Analogous to subject dependent experiments, stretch dynamics and point dynamics achieve the best and comparable performances (see Fig. 9). The features involving STIP present much lower rates than the case in subject dependent experiments, as it is expected for subject independent experimental results. We should note that point/STIP precision scored higher than in the subject dependent experiment, this is the effect of the classifier categorizing most of the utterances as "incorrect", hence lowering the false positive number, and increasing the precision score.

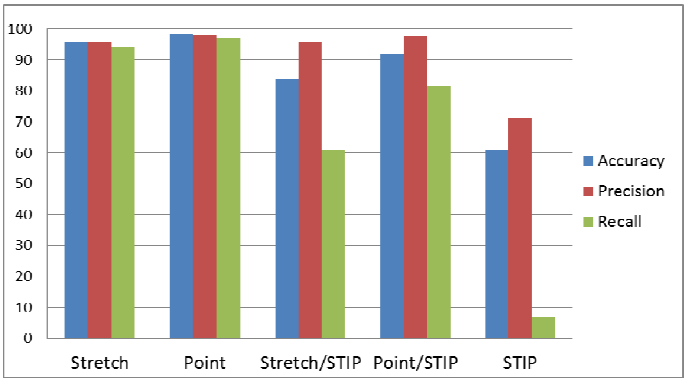


Figure 9. Subject independent comparison results of different feature combinations.

VI. CONCLUSION

This paper presents a visual speech learning system including visual speech segmentation and comparison based on the analysis of lip movement. The first step is to segment each individual word utterance across a video sequence by identifying the frames with moving lips (i.e., uttering a word) and neutral lips (i.e., no speech). The second step is to determine whether a student is correctly uttering the word or speech being taught by an e-tutor. The segmentation and comparison methods proposed in this paper achieve the state-of-the-art performances in both subject dependent and subject independent experiments, which would ultimately provide an aid to assist the deaf and hard-of-hearing persons to effectively communicate with other people through spoken language and speechreading.

In our future work, we will explore surface electromyography (sEMG) for speech comparison and for integration with our visual speech learning system. The final task will be to do a user interface study in order to find the best approach to interact, and present our speech learning system to deaf and hard-of-hearing students via the visual interface of the computer-user environment.

ACKNOWLEDGMENT

This work was supported in part by NSF grants EFRI-1137172, IIP-1343402, and FHWA grant DTFH61-12-H-00002.

REFERENCES

[1] S. Chen, D. Quintian, and Y. Tian. Towards A Visual Speech Learning System for the Deaf by Matching Dynamic Lip Shapes. ICCHP, 2012.

[2] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active Shape Models: Their Training and Application. CVIU, 1995.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual Categorization with Bag of Keypoints. ECCV Workshop on Statistical Learning in Computer Vision.

[4] I. Laptev. On Space-Time Interest Points. IJCV, 2005.

[5] I. Laptev, M. Marszalek, C. Schmid, and B. Rozefeld. Learning Realistic Human Actions from Movies. Realistic Human Actions from Movies. In Proc. CVPR, 2008.

[6] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior. Recent Advances in the Automatic Recognition of Audio- Visual Speech. Proc. of IEEE, 2003.

[7] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey. Extraction of Visual Features for Lipreading. TPAMI, 2002.

[8] D. Stork and M. Hennecke. Speechreading by Humans and Machines: Models, Systems and Applications. Springer, 1996.

[9] Y. Wei, "Research on Facial Expression Recognition and Synthesis", Master Thesis, 2009, software available at: http://code.google.com/p/asmlibrary.

[10] S. Werda, W. Mahdi, and A. B. Hamadou. Colour and Geometric based Model for Lip Localisation: Application for Lip-reading System. ICIAP 2007.

[11] G. Zhao, M. Barnard, and M. Pietikainen. Lipreading with Local Spatial-Temporal Descriptors. TMM, 2009.

[12] Z. Zhou, G. Zhao, and M. Pietikainen. Toward a Practical Lipreading System. CVPR, 2011.

[13] Gallaudet Research Institute. Regional and National Summary Report of Data from the 2007-08 Annual Survey of Deaf and Hard of Hearing Children and Youth. Washington, DC: GRI, Gallaudet University. 2008.

[14] Goldin-Meadow, S. & Mayberry, R. I., How do profoundly deaf children learn to read? Learning Disabilities Research and Practice, 16 (4), 222-229, 2001.

[15] Lee, S., Henderson, V., Hamilton, H., Starner, T. and Brashear, H. A Gesture-Based American Sign Language Game for Deaf Children, CHI2005, 2005

[16] Wang, X., Xue, L., & Yang, D., speech Plot Display for the deaf-mute based on Combined Characters Encoding of speech Signal, IEEE/ICME International Conference on Complex Medical Engineering, 1206 – 1209, 2007.