Towards a Visual Speech Learning System for the Deaf by Matching Dynamic Lip Shapes

Shizhi Chen, D. Michael Quintian and YingLi Tian

Media Lab, Department of Electrical Engineering
The City College, City University of New York, New York, USA
{schen21, dquinti00, ytian}@ccny.cuny.edu

Abstract. In this paper we propose a visual-based speech learning framework to assist deaf persons by comparing the lip movements between a student and an E-tutor in an intelligent tutoring system. The framework utilizes lip reading technologies to determine if a student learns the correct pronunciation. Different from conventional speech recognition systems, which usually recognize a speaker's utterance, our speech learning framework focuses on recognizing whether a student pronounces are correct according to an instructor's utterance by using visual information. We propose a method by extracting dynamic shape difference features (DSDF) based on lip shapes to recognize the pronunciation difference. The preliminary experimental results demonstrate the robustness and effectiveness of our approach on a database we collected, which contains multiple persons speaking a small number of selected words.

Keywords: Lip Reading, Speech Learning, Dynamic Shape Difference Features, Deaf people.

1. Introduction

About 35 million Americans today are deaf or hard of hearing. Approximately 12 out of every 1,000 individuals with hearing impairment are under 18 years of age, based on the most recently available data from the National Center for Health Statistics (NCHS). Recent research has demonstrated that even mild hearing losses can create significant challenges for children as they develop skills to interact with the world [5, 7].

The loss of auditory feedback poses significant difficulties on the speech learning for the deaf people, since it is difficult for them to know immediately if they speak correctly [1, 10, 12]. Some researchers propose to use animations as feedback according to audio signals [6, 11]. The animations can be helpful for the deaf people to know if they speak correctly. However, such animation does not provide feedback on how to correct their speech and how the incorrect speech different from that of the instructor.

On the other hand, visual cue often provides complementary information for speech recognition [8, 9, 15]. Figure 1 shows a lip movement in a video sequence

when speaking word “apple”. It is easier for a deaf person to visualize the difference between the incorrect and the correct utterances by simply looking at the lip movements.

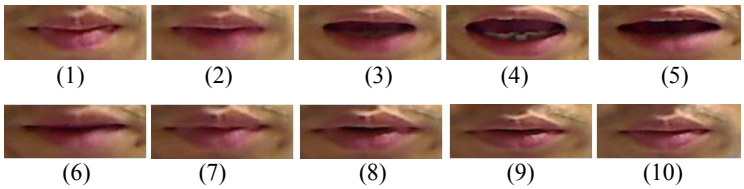

Figure 1: Sample lip movements of a video sequence when speaking word “apple”.

Potamianos *et al*. [9] have shown a significant improvement of speech recognition through both audio and visual modalities as compared to the approach of audio modality only. Matthews *et al*. combined lip contour and lip appearance information to recognize isolated letters A-Z [8]. Then the authors employ Hidden Markov models (HMM) as the classifier to model the temporal dynamics of a speech. The authors demonstrated the effectiveness of the speech recognition based on only visual modality. The visual based speech recognition becomes particularly useful in the noisy environment, in which audio signal is significantly degraded.

Zhou [15] recently captures temporal dynamics of a speech by extending Local Binary Pattern (LBP) to a temporal domain [14], which is also visual based speech recognition. Ten phrases are used for their speech recognition experiments. The experimental results also show a promising performance of visual based speech recognition.

Figure 2: The basic hardware configuration of the proposed interactive intelligent tutoring system, which includes a computer (desktop or laptop), a web camera with auto focus (face to the user), and a microphone.

Inspired by these advances on speech recognition, we propose a visual-based speech learning framework to aid deaf people. As shown in Figure 2, the system configuration is set up as an E-Tutoring system. A deaf student in front of a computer learns speech by following an E-tutor. A web camera is used to capture the student’s face and lip movement. The video of the student is then processed in real-time by comparing the student’s lip movements with those from the pre-recorded tutor. Interactive feedback is provided to students through easily understandable visual displays.

Different from the visual based speech recognition, which usually recognizes a few words, a practical speech learning system usually needs to handle much larger

vocabulary. It would be extremely difficult to design a speech learning system if we have to recognize every single utterance between a student and an instructor. Hence, we propose a new framework by extracting dynamic shape difference features (DSDF) to directly measure the visual difference of lip shapes between two speakers, *i.e.*, the student and the instructor. Therefore, we can reduce a multi-class recognition problem in a speech learning system to a binary class recognition problem, i.e., recognizing whether the student pronounce correctly according to the instructor's utterance.

We have collected a database which consists of 9 words spoken by four people respectively. By pairing up two subjects speaking same or different words, we generate "correct" or "incorrect" samples to evaluate performance of the proposed speech learning framework. The "correct" sample corresponds to the case when both subjects speak the same word, while the "incorrect" sample corresponds to the case when the two subjects speak different words. Our preliminary experiments have shown encouraging results of this approach.

2. Visual Based Speech Learning Method

2.1 Overview

Figure 3 shows an overview of our speech learning framework. First, the lip movements of both student and E-tutor are tracked by an Active Shape Model (ASM) [4, 13]. Then we align the lip shapes to remove the head movements while speaking, *i.e.*, translation, and rotation. In this step we also remove the lip shape variance caused by different subjects. Due to the time resolution difference when speaking a word, we perform temporal normalization over the extracted lip shapes in a video sequence, so that both student and instructor can have same speaking speed. The resulted features are defined as dynamic shape features.

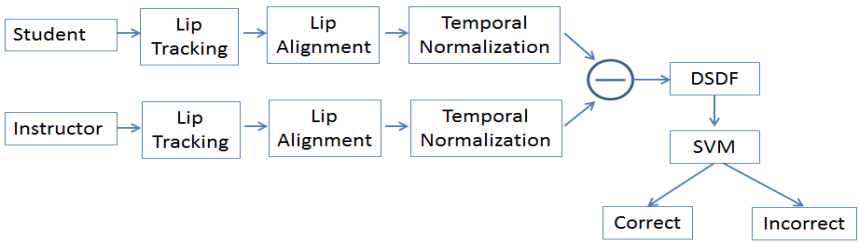

Figure 3: Overview of our proposed visual speech learning framework.

Finally we calculate the difference of the dynamic shape features between the student and the instructor, *i.e.*, dynamic shape difference features (DSDF), as the input to a Support Vector Machine (SVM) based classifier. The SVM classifier then automatically determines if the lip movements of the student correctly follow the lip movements of the instructor based on the visual difference of lip shapes between the student and the instructor.

2.2 Lip Tracking

We employ Active Shape Model (ASM) [4, 13] to track lip movements. ASM is a shape-constrained iteratively fitting method, which utilizes prior knowledge of lip shapes in training images. The shape is simply the x and y coordinates of all landmark points on a lip after appropriate alignments, which is shown in Eq. (1).

$$\mathbf{X}_i = [x_1, y_1, x_2, y_2, \ldots x_j, y_j, \ldots x_n, y_n] \quad , \tag{1}$$

where n is the number of landmark points labeled for a lip. In our experiments, we choose 19 landmark points, including both outer contour and inner contour of a lip. For the simplicity, we use the built-in ASM model, which is trained using the 68 landmark points of the whole face including the 19 lip points [13]. Figure 4 shows a lip tracking example in a video sequence using the ASM model.

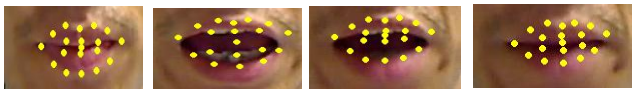

Figure 4: An example of lip tracking in a video sequence by employing Active Shape.

2.3 Lip Alignment

In order to remove the effects of head movements and rotations during the speech, we perform an alignment procedure. The alignment procedure calculates the angle formed by the line connecting both lip corners and x axis. Then we rotate the shape by the calculated angle so that the left lip corner and the right lip corner have the same y coordinate value. The mean x and y values are removed. The entire shape is then adjusted vertically to align the two lip corners on the x axis.

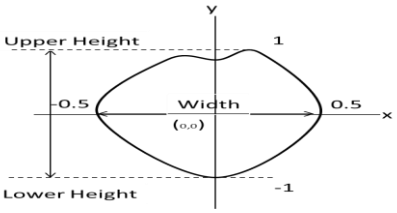

Figure 5: Typical neutral lip shape after the alignment and the normalization with the upper height, lower height, and width of the lip shape on the neutral frame.

Different subjects have different neutral lip shapes. To eliminate these subject dependent shape variations, we perform a normalization using the upper lip height, the lower lip height, and the width of the neutral lip shape for each subject. The neutral frame in our database is simply the first frame in the video sequence. Figure 5 shows a typically aligned and normalized lip shape on the neutral frame without the landmark points. The normalized lip shapes in a video sequence represent how the lip shape deforms from the neutral shape during the speech.

From the experiments, we find that the performance is usually improved by adding the upper height, the lower height, and the width of each frame's lip shape to the

normalized shape vector as described in last paragraph. Finally, we perform the *L2* normalization on the resulted feature vector in each frame.

2.4 Temporal Normalization

The time usually varies for different subjects even when they speak same words. In order to handle this time resolution difference, we temporally normalize the video sequence to a fixed number of frames by linearly interpolating each frame's feature vector along the temporal direction [2, 3]. We choose *30* as the number of temporally normalized frames in a video sequence.

Each frame's shape feature vector has the feature dimension of 41, *i.e.*, *2*19+3*. Therefore, a video sequence is represented by the concatenated frame feature vector with the total dimension of *1230*, *i.e.*, *41*30*. The concatenated feature vector of a video sequence is defined as dynamic shapes.

2.5 Dynamic Shape Difference Features (DSDF)

By taking the difference of the dynamic shapes between the instructor and the student, we form the dynamic shape difference features (DSDF). The DSDF features directly measure the pronunciation difference of the two speakers, regardless the words spoken.

Here, we do not recognize the words spoken by the instructor and the student individually to determine if the student speaks same word as the instructor, since this approach can quickly become too complicated to recognize every word accurately as the number of words increase in the speech learning system. By employing the DSDF feature to recognize the similarity between the utterances directly, our system is not limited to the number of words or utterances spoken, which is desirable for any practical speech learning system.

2.6 Support Vector Machine Classifier

Finally, we employ a support vector machine (SVM) as the classifier with the DSDF feature from the instructor and the student as the input feature vector. The output of the classifier is to determine if the student correctly follows the instructor's utterance regardless words they speak.

SVM is to find an optimal hyper-plane which can separates the opposite classes with the maximum margin. We employ the RBF kernel, which has demonstrated the state of the art performance in many applications, such as object recognition and detection etc.

3. Experimental Results

3.1 Database

We have recorded a database to study the effectiveness and robustness of the proposed speech learning framework. Nine words were chosen such that some words are unique, and some words are similar to each other. The selected words are "apple", "cruise", "find", "hello", "music", "open", "search", "vision", and "window".

In our database, each word is spoken ten times by each subject. There are four subjects in the dataset. The video is captured at frontal face by a web-camera with the entire head of the subject within the image frame, in order to ensure the face has enough resolution. The speaker begins a word with a neutral expression, says the word, and then returns to the neutral expression. Each of the chosen words takes an average of one second to complete. Depending on the speaker, some words take up to two seconds to complete.

All the videos have a spatial resolution of *640x480* pixels, with a frame rate of *30* frames per second. The videos are edited such that the first and last few frames (about *3-5*) contain a neutral expression. The average video sequence is between *20* to *40* frames long. Figure 6 shows a sample video sequence of a subject speaking the word "apple".

Figure 6: A sample video sequence of speaking the word "apple".

3.2 Subject Dependent Results

We evaluate the speech learning framework by pairing up two persons from the database. If the selected two persons speak same word, then we know that one speaker has correctly followed the other speaker. Otherwise, one speaker has incorrectly followed the other speaker. Hence we have the ground truth, whether one speaker correctly follows the other speaker, by simply checking the words they speak. That is, if they speak same word, the ground truth is "correct". Otherwise, the sample consisting of the pair of speech has the label of "incorrect".

Each word is spoken *10* times by each subject, and there are *9* words in our database. Therefore, we have *900* possible pairs of utterances which speak the same word for each selected pair of subjects, i.e., we have *900* "correct" samples. Similarly, for each pair of subjects, there are *7200* possible pairs of utterance which speak different words, i.e., there are *7200* "incorrect" samples. We choose *900* out of the *7200* "incorrect" samples, so that the number of "incorrect" samples from every combination of different words is approximately equal. Then we divide the *900* "correct" samples and the *900* "incorrect" samples to the training and testing sets by the ratio of *9* to *1*.

Our speakers include one American male (M), one Chinese male (S), one Chinese female (X), and one American female (K). The capital letter is an identifier for the person. We adopt precision and recall as our evaluation metrics, which are defined in Eqs. (2) and (3).

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

where *TP* is the number of "correct" samples which are also predicted correctly. *FP* is the number of "incorrect" samples which are misclassified as "correct" samples. *FN* is the number of "correct" samples which are misclassified as "incorrect" samples.

Table 1(a) shows the average precision and the average recall over all words for each selected pair of speakers. That is to train and test the proposed speech learning framework by the same pair of subjects. Table 1(b) shows the detailed recall over the individual words for the corresponding pair of subjects. These results indicate the robustness of the proposed speech learning system.

There are some variations among different pair of speakers on the precision and recall. One explanation for this variation is the fact that different people say the same word differently. When collecting the database, we have observed some speakers open their mouse slightly prior to saying a word.

Table 1: (a) average precision and average recall over all words for each selected pair of speakers; (b) recall over the individual words for each selected pair of speakers.

	Precision	Recall
M-S	95.2%	87.8%
M-K	100.0%	97.8%
M-X	98.9%	96.7%
S-K	96.5%	91.1%
S-X	100.0%	92.2%
K-X	100.0%	92.2%

(a)

	Apple	Cruise	Find	Hello	Music	Open	Search	Vision	Window
M-S	70%	100%	90%	70%	100%	70%	90%	100%	100%
M-K	80%	100%	100%	100%	100%	100%	100%	100%	100%
M-X	80%	100%	90%	100%	100%	100%	100%	100%	100%
S-K	30%	100%	100%	100%	100%	100%	100%	90%	100%
S-X	70%	100%	80%	100%	100%	80%	100%	100%	100%
K-X	80%	100%	100%	90%	60%	100%	100%	100%	100%

(b)

3.3 Subject Independent Results

In order to evaluate the proposed speech learning framework for the subject independent case, we group all "correct" and "incorrect" samples from every pair of speakers as shown in Table 1. Then we just train a single model to recognize if one speaker correctly follows another speaker. The precision and the recall shown in Table 2 demonstrate that the proposed framework is also effective for subject independent case.

Table 2: (a) Average precision and average recall over all words when grouping every pair of speakers in Table 1; (b) detailed recall of (a) over the individual words.

Precision	Recall
98.7%	95.9%

(a)

Apple	Cruise	Find	Hello	Music	Open	Search	Vision	Window
81.7%	96.7%	100.0%	95.0%	95.0%	96.7%	100.0%	98.3%	100.0%

(b)

4. Conclusion

We have proposed a framework to help deaf people learn speech by visually comparing the lip movements of a student and an instructor. The framework utilizes lip reading technologies to determine if the student correctly follows the instructor in pronunciation of a word. Furthermore, our proposed framework is very practical by employing the dynamic shape difference feature (DSDF), which can avoid the large vocabulary problem in traditional speech recognition systems. The preliminary experimental results indicate that our proposed speech learning framework is robust in both subject dependent and subject independent cases. More extensive experiments and user interface study including the system test by deaf people will be conducted in future. A larger database with more subjects and more words will also be collected in order to train a model which can be robust in the practical application.

5. Acknowledgement

This work was supported in part by NSF grant IIS-0957016 and DHS Summer Research Team Program for Minority Serving Institutions Follow-on Award. Shizhi Chen is funded by NOAA CREST Grant NA11SEC4810004.

References

1. S. Awad, "The Application of Digital Speech Processing to Stuttering Therapy", IEEE Instrumentation and Measurement, 1997.
2. S. Chen, Y. Tian, Q. Liu and D. Metaxas. Segment and Recognize Expression Phase by Fusion of Motion Area and Neutral Divergence Features. IEEE Int'l Conf. on Automatic Face and Gesture Recognition (AFGR). 2011.
3. S. Chen, Y. Tian, Q. Liu, D. Metaxas, "Recognizing Expressions from Face and Body Gesture by Temporal Normalized Motion and Appearance Features", IEEE Int'l Conf. Computer Vision and Pattern Recognition workshop for Human Communicative Behavior Analysis (CVPR4HB). 2011.
4. T. Cootes, C. Taylor, D. Cooper and J. Graham, "Active Shape Models – Their Training and Application", Computer Vision and Image Understanding, 1995.
5. J. Hailpern, K. Karahalios, L. DeThorne, & J. Halle, "Encouraging Speech and Vocalization in Children with Autistic Spectrum Disorder", Workshop on Technology in Mental Health, CHI 2008, 2008.
6. Lavagetto, F. (1995); Converting speech into lip movements: a multimedia telephone for hard of hearing people, IEEE Transactions on Rehabilitation Engineering, Volume: 3 Issue:1, 90 – 102.
7. M. Marschark, P. Sapere, C. Convertino, C. Mayer, L. Wauters, & T. Sarchet, "Are deaf students' reading challenges really about reading?", American Annals of the Deaf, 154 (4), 357-176, 2009.
8. I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading", TPAMI, 24(2):198–213, 2002.
9. G. Potamianos, C. Neti, G. Gravier, A. Garg, A. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech", Proceedings of the IEEE, 91(9):1306–1326, 2003.
10. M. Rahman, S. Ferdous, and S. Ahmed, "Increasing Intelligibility in the Speech of the Autistic Children by an Interactive Computer Game", IEEE International Symposium on Multimedia, 2010.
11. R. Riella, A. Linarth, L. Lippmann, P. Nohama, "Computerized System to Aid Deaf Children in Speech Learning", IEEE EMBS International Conference, 2001.
12. O. Schipor, S. Pentiuc, and M. Schipor, "Towards a Multimodal Emotion Recognition Framework to Be Integrated in a Computer Based Speech Therapy System", IEEE Conference on Speech Technology and Human Computer Dialogue (SpeD), 2011.
13. Y. Wei, "Research on Facial Expression Recognition and Synthesis", Master Thesis, 2009, software available at: http://code.google.com/p/asmlibrary.
14. G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatialtemporal descriptors", TMM, 11(7):1254–1265, 2009.
15. Z. Zhou, G. Zhao, M. Pietikainen, "Toward a Practical Lipreading System", CVPR, 2011.