Lip Reading Analysis of English Letters as Pronounced by Filipino Speakers Using Image Analysis

Cruz, Hans Miguel¹, Puente, Jofet Kane T²*, Santos, Christian¹, Vea Larry A.¹, Rajendaran Vairavan²

¹School of Information Technology, Mapua University, 333 Sen. Gil Puyat Ave., Makati City 1200, Philippines
²School of Microelectronic Engineering, University Malaysia Perlis (UniMAP), Kampus Alam, Pauh Putra, 02600 Arau, Perlis, Malaysia.

Abstract. Lip reading or speech reading is a skill of understanding speech by watching the speaker’s actions. Main factors that will lead to a good speech reading, and these are: lip reading experience, good language knowledge, normal vision, good verbal short term memory, and familiarity with the speaker. Hence, this study aims to develop a model that will recognize the spoken English letters through lip reading. This study focuses on the analysis of lip reading through video processing. The video that was used in the experiment was limited to pre-recorded videos of the subjects pronouncing each letter of the English language. The test subjects are limited to a total of thirty Filipino English speakers, composed of 15 men and 15 women, to be specific. Out of the thirty subjects, two of them was an expert in speaking the English language. The test subject does not have a moustache and beard, also that the test subjects does not have lip deformation, and lastly the test subjects are limited to do extra unnecessary movements. The authors created their own image database that was used in the experiments. They pursued the experiment in a controlled environment consisting only of supplemental lighting. The camera that was used to record the video is a DSLR camera using a 640x424 image resolution running at 24fps. Only the movement of the exterior lip region is included in the experiment without regarding to the movement of the tongue and teeth. Overall results showed that the system demonstrated an overall precision accuracy of 45.62%.

1 Introduction

Lip reading or speech reading is a skill of understanding speech by watching the speaker’s actions. The speaker’s actions are said to be the following: speaker’s mouth movements, head, face, eye and torso movements. These are main factors that will lead to a good speech reading, and these are: lip reading experience, good language knowledge, normal vision, good verbal short term memory, and familiarity with the speaker. These factors are the basis that lead to speech reading. It is also stated that the important aspects of visual image capture for speech reading are lightning, view, color, distance, look, and image compression [1, 2 & 3].

There are also conflicts on lip reading. It is stated that these are factors that make lip reading difficult such as; speech is designed to be heard and not viewed, most movements of the tongue and teeth that are not visible, not visible actions of tongue, lips, teeth (m/p, f/v, th), distinctive mouth shapes (’a’, ’i’, ’u’), most consonants that are liable to error, due to lips look alike (’d’, ’t’, ’k’), and lastly the consonants that easily distinguish (m, n, th). The largest part of information is derived from the movements of the mouth region [4]. Lip movements contributes significantly to speech information where the audio signal is absent or of low quality. A major drawback in dealing with visual cues is dealing with the large amount of visual data. This leads to storage problems and additional computational effort; thereby, leading to increase in processing time [5]. Visual speech information through lip reading is very useful for human speech recognition [6]. Much of the research on lip reading uses frontal facial images. A frontal facial image contains much useful information [7].

As Summerfield (1987) has remarked: any comprehensive account of how speech is perceived should encompass audio-visual speech perception. The ability to see as well as to hear has to be integral to the design, not merely a retro-fitted after thought [8]. Study regarding lip reading by [9] was done by utilizing Using 3D on Hidden Markov Model to analyze visual speech information, such as lip movement. A study by [10] demonstrated a novel technique which modelled visual speech movements by employing fuzzy neural network to analyze the behavior of images derived from video film with particular focus on input coding. A work by [11] proposed lip information extraction by employing

* Corresponding author: puentejofet@yahoo.com
fusion of geometric and motion feature analysis. A research by [12] demonstrated application of kohonen algorithm for development of lip recognition software.

Thus, this study aims to develop a model that will recognize the spoken English letters through lip reading. This study demonstrates the Test and validate done on the developed model for the lip reading analysis of English letters pronounced by the Filipino speakers. This study focuses on the analysis of lip reading through video processing. The video that was used in the experiment was limited to pre-recorded videos of the subjects pronouncing each letter of the English language.

2 Methodology

The study applies a controlled-type laboratory experiment consisting of male and female participants. The participants pronounced the English letters that consists from letter A-Z. The set of letters used are stated in the variables. The data gathered are processed under MATLAB to format the video to sequence of images .AVI, segment the lip area and consists of initial eight (8) key feature points. Comparison was done between 8 key feature points, 12 key feature points and 16 key feature points. The data were analysed under certain variables and the authors decided to select thirty (30) students or any person that speaks English letters properly. The test subjects consists of fifteen (15) female and fifteen (15) male and the test was held in an enclosed lighted room that supports the quality of the test video, and the room that was used is the same for all test subjects. The distance of the camera from the speaker/subject is one meter and the height of the camera (from the floor up to the camera lens) is 3 meters. The initial and final state of the test subject’s lips is closed. Figure 1 shows the setup for data gathering. The distance of the camera from the test subject is a meter away, and the camera is focused at the face of the test subject .The authors gathered the test subject’s data by recording a video using a Nikon D3100 camera with the lens specification 18mm-55mm. The subjects pronounced a letter from the English alphabet which is used as a part of the data. The numerical data gathered from the subjects was used to train the model.

![Fig. 1. Data gathering Setup](image)

2.1 Converting Video to image sequence using MATLAB integrated in Java

In this phase, the authors converted the pre-recorded video to image sequences for the image analysis.MATLAB was utilized. The authors developed a process in MATLAB that will automatically create the folder and convert the video into images. Twelve (12) frames of images for processing were taken. The authors converted the images in .jpg format for the pictures.

2.2 Lip detection using Viola-Jones Algorithm and point plotting using Point distribution model tracking KLT Algorithm

In Viola-Jones technique is used based on exploring the input image by means of sub window capable of detecting features [13]. Viola Jones algorithm has 4 stages: Haar Feature Selection, Creating an Integral Image, Adaboost Training and Cascading classifiers. Haar features are like simple square waveforms which detects “darker” blocks adjacent to “lighter” blocks also known as “features”. The feature value is calculated by the sum of the pixel intensities in the light rectangles subtracted from the sum of the pixels in the dark rectangle. In order to improve the summing of the intensities of the pixels in a given rectangle, Integral image was calculated at every point. Integral image is computed by the sum of the pixels of the rectangle from upper then to the left corner of the given point; after calculating the integral image of every point the area of any rectangle can be now calculated. The formula in solving for the area of the rectangle is $D+A-(B+C)$. Adaboost Training combines the “weak” classifiers to come up with a “strong” classifier. The last step is to cascade classifiers, each sub window of the original image will now be tested against a series of classifiers and if the sub window fails at any stage of the testing, the sub window is rejected and if it passes all through the classifiers it is classified as a face.
After the images are pre-processed, the next step is the process on how the lip segmentation works. The original mask image is the result of inverted output during the detection of the flesh. The image is converted into CIELAB color space and the flesh is detected, after the flesh is detected the output is being inverted so that the red color (lips) is detected. After wards, the image goes through the Sobel mask detection which helps to enhance the mask that will be created. After all that process was done, the last part is to combine the output of the previous phases.

Point Distribution Model (PDM) and Kanade Lucas Tomasi (KLT) tracking algorithm template is created using sixteen point. This model describes the lip contour by a set of points. Figure 2 is the representation of Point Distribution model for Lip point plotting in MATLAB works. Comparison was done between 8 key feature points, 12 key feature points and 16 key feature points. The comparison was done by clustering the several alphabets into 8 clusters and 5 frames (F2 to F6). This was then repeated again with second set of 8 clusters and 5 frames (F7 to F11). The difference between these two major clusters is the arrangement of several alphabets in each respective frames. Figure 3 shows the key features for the lips. The max number of points consist of 16 feature in relation of the horizontal and vertical distance of the lips. The 16 points are labelled as: left corner, right corner, top left, top right, top left left, top left right, top right left, top right right, bottom left, bottom right, bottom left left, bottom left right, bottom right left, bottom right right, bottom left and bottom right feature set. This is done speaker-wise. The normalized values are used for feature selection and classification.

Fig. 2. Image sequence point plotting

Fig. 3. Key features of the lips (a) 8 key features (b) 12 key features (c) 16 key features.

2.3 Lip Feature Extraction And Selection Using Matlab Works

This is done by defining new important lip features from selecting twelve (12) relevant images that will be used for image processing and will produce raw data sets. The feature extraction phases are done under MATLAB works by means of the extracting data through point plotting to produce the raw data sets of the distance of each feature points. After that process, the raw data are converted in to decimal points that pertains to the distance of the lips while moving according to image sequence; all of the data was automatically saved in an excel file, and it is converted to .mod extension file by using the prototype integrated program in java .mod converter that is done by the authors and the .mod file is used as the model data in rapid miner. From the feature extracted, the authors selected twelve relevant frames for each letter. The selected twelve relevant frames are used for point plotting that is consisted of sixteen feature points that are used for Feature Extraction. For the feature selection phase, it is done by analysing the decision tree of the model in the rapid miner and they analyse all the extracted data and focus on the selected frames which consists of frames 2, 4, 6, 8, and 10.

2.4 Lip Feature Extraction And Selection Using Matlab Works

After the extraction and selection of data, a model was created based on the results of the classification. Decision tree is a structure where it mutually partitions the inputted dataset and labels it to identify the data points. The authors used a total of thirty test subjects with two tries per English letter that was used for establishing the trained data sets, and for
the test sets they used different test subjects that consists only of one try per letter. The model that is created in Rapid Miner using decision tree J48 algorithm was embedded in the prototype, which was also the same program that was used for pre-classification.

2.5 Model Testing and Analysis

For the model testing, the authors used another 10 test person to establish data for the test sets per letters that will consist of 1 try only. After the authors established the data sets, it is saved as CSV file and it is loaded to the prototype. Once the test data is loaded to the prototype, the software will start to recognize the test sets to the trained model to see how accurate is the model. The authors used Java to create a prototype, with the displayed process, and used the model built by Rapid Miner and tested it by developing a prototype, by using decision tree J48 algorithm [14].

3 Result and Discussion

Table 1 shows the results for Key features of lips according to its respective key features. Based on Table 1 percentage by having a sixteen key features and it is used to establish the model training data. From Frame 2, it has a 75.09% average because it’s where the lips is starting to open and most of the letters has the same starting phase of the lips. Then in Frame 3, it increases to 76.46% because in this Frame the lips is starting to move from its starting phase. In frame 4 it continue to rise to 78.21%, Frame 5 has a percentage of 76.92. Compare from frame 4 frame 5 was lower than Frame 4 because some of the letters depends on how it was pronounced by the speaker. Frame 6 has the value of 77.15%, Frame 7 has a percentage of 76.99. Frame 8 and has a same value of percentage which is 77.45%. Frame 10 has 77.60% and lastly Frame 11 which is the closing state of the lips has a percentage of 76.31%. Based on Table 1, it is observed that 16 key features point showed best percentage per frame for the letter clustering compared to the 12 key feature points. These results are based from the experiment that is done by the authors to see which key features has the higher percentage per frame that is used to establish the model for this study. Those percentage per frame is based on the distance of the key feature points. Thus 16 key feature points is established for the prototype testing as it shows better evaluation points on the lip movement. The results of the prototype testing is shown in Table 2. The results is based on the precision of the prototype to distinguish each letter as it is being pronounced.

| Frame | 8 key features (%) | 12 key features (%) | 16 key features (%) |
|-------|--------------------|---------------------|---------------------|
| F2    | 71.53              | 73.50               | 75.09               |
| F3    | 73.73              | 75.25               | 76.16               |
| F4    | 72.44              | 76.23               | 78.21               |
| F5    | 74.03              | 75.17               | 76.92               |
| F6    | 72.67              | 76.17               | 77.15               |
| F7    | 74.11              | 75.40               | 76.99               |
| F8    | 75.63              | 75.63               | 77.45               |
| F9    | 73.80              | 75.40               | 77.75               |
| F10   | 72.51              | 76.23               | 77.60               |
| F11   | 71.53              | 76.77               | 76.31               |

| Letter | Precision | Letter | Precision |
|--------|-----------|--------|-----------|
| A      | 25%       | N      | 41%       |
| B      | 46%       | O      | 32%       |
| C      | 72%       | P      | 23%       |
| D      | 81%       | Q      | 48%       |
| E      | 58%       | R      | 44%       |
| F      | 61%       | S      | 31%       |
| G      | 32%       | T      | 42%       |
| H      | 50%       | U      | 47%       |
| I      | 42%       | V      | 31%       |
| J      | 35%       | W      | 28%       |
| K      | 51%       | X      | 34%       |
| L      | 77%       | Y      | 49%       |
| M      | 51%       | Z      | 55%       |
| Total Accuracy | 45.26%   |

Table 1. Results for Key features of lips

Table 2. Prototype testing results
The overall results will be discussed in terms of per letter analysis. Based on the results, it indicates that letter A has a precision rate of 25%; upon reviewing the results the authors found out that mostly of the frames has a precision rate below 50% and the only frame that has a higher precision rate is Frame 1. Letter B has a precision rate of 46%. By reviewing the results of each frames, the authors observe that it is difficult to predict in the middle stages of the frames (4,5,6), because those frames are the peak opening of the lips that causes some of letters clustered with B is similar with each other.

Letter C has a high precision rate of 72% and Letter D has a high precision rate of 81% and letter E showed a precision rate of 58%. Letter F showed a good precision rating with 61%. Upon reviewing the results it shows that almost all the time that this letter is correctly predicted. Letter G has a precision rate of 32%. The authors noticed that each letters has a precision rate that is ranging from 20-50% and the highest value of precision rate is 50%. Letter X, Letter Y, Letter Z have the precision value of forty percent (40%). Letter J had a precision rate of 41%. Based from the assessment, Frame 8 has a precision rate of zero percent (0%) while the rest of the frames had a precision rate value below fifty percent (50%).

Letter K gave a precision rate of fifty one percent (51%) during the testing phase. Letter L showed a seventy seven percent (77%) precision rate during the testing phase. Letter M gave a precision rate of seventy seven percent (77%). Upon reviewing the result it showed that almost all of the time, this letter is correctly predicted and if not, usually it is predicted when it’s clustered with letter L during the Frames 4 and 8; this two (2) letters has almost identical movements. Letter N has a precision rate of forty one (41%), upon reviewing the results, the authors found out that most of the frames have a precision rate below fifty percent (50%) and the only frame that has a higher precision rate is Frame 3. Letter O had a precision rate of 25% upon reviewing the results the authors found out that the value of the frames is ranging from 10-50%. Letter P has a low precision rate of 23%. Letter R had a precision rate of forty four (44%), upon observing the result, the authors found out that some of the frames has a precision value below 50% and frame 8 has the higher precision value which is 70%. Letter S had a precision rate of thirty one (31%).

Based from the assessment, Frame 8 and 9 had a lowest precision rate while the rest of the Frames has a precision rate value below 50%. The highest precision rate is Frame 10. Letter T had a precision rate of thirty one percent (31%). Based from the reviewing of the result, Frame 10 and 11 has a lowest precision rate while the rest of the Frames has a precision rate value below fifty percent (50%). Letter U had a precision rate of forty seven (47%) during the testing phase. Letter V had a precision rate of thirty one percent (31%). Based from the reviewing of the result, Frame 10 and 11 has a lowest precision rate while the rest of the Frames has a precision rate value below fifty percent (50%). Letter X had a precision rate of forty nine (49%) during the testing phase. Letter Y had a precision rate of forty seven (47%) during the testing phase. Letter Z had a precision rate of fifty five percent (55%). Based from the reviewing of the result, Frame 8 and 9 had a lowest precision rate while the rest of the Frames has a precision rate value below 50%. Letter A had a precision rate of eighty percent (80%), upon reviewing the results it shows that the most useful and distinguishable feature movements are the vertical distance of features 4(TOPLR-BOTLR), 5(TOP-BOT), 6(TOPRL-BOTRL), F7(TOPR-BOTR), F9(TOPLL-BOTLL), F10(TOPL-R0P), F11(TOPLR-L), F12(BOTLL-BOTRR), F13(BOTL-BOTR), F14(BOTLR-BOTRL). Based on this, it was found out that the most useful and distinguishable feature movements are the vertical distance of features 4(TOPLR-BOTLR), 5(TOP-BOT), 6(TOPRL-BOTRL) and the horizontal distance of feature 1.

The model is correctly recognized by means of conducting a test that will have a good performance accuracy rate that established a good model data sets. If the test data sets are tested to the model data it brings a good recognition rate from the model. By using some of the algorithms in RapidMiner it established a model in batch cross fold validation and has a relevant recognition rate between the model and test sets. Using the KNN algorithm, this is applied for making the model of lip reading analysis because it recognizes a random data and does not give a better result.

From the results taken from Decision tree J48 batch algorithm method in RapidMiner, using this algorithm for letter recognition this gives the most efficient way compared from the other algorithm. It consists of sixty two percent (62%) correctly classified instances and has a percentage of error thirty eight percent (38%). Most of the letters have an error in classification because of phonemes, because some letters has the same way on how to pronounce but different letter. As the results some of the testing has a different result because of that error.

As the authors could not find concrete evidence to show on how to improve and on how to prevent detecting words that have the same pronunciation but different letters. The authors recommend future research for its future study. The Authors concluded that the key features of the lips are the vertical distances especially Feature 4, 5, 6 which is the middle vertical feature in the lips and the horizontal Feature which is Feature 1.

They also recommend the following to the future authors of this study. In order to achieve a better and more accurate lip segmentation, they can use lipsticks on the subjects to easily discriminate the skin from the lips. Another benefit from this is that it can lower the processing time of the method because the detection of the flesh and inversing its result to detect the red/lips can be removed from the process. They showed a deeper and thorough studies to fully understand and differentiate the phoneme letters with each other; restrict the train subjects to professional English
speakers such as call center agents, English linguist, and latter to have a more accurate model. It is better to use a controlled light such as face lights and reflectors aside from the room lights to acquire better video quality.

4 Conclusion

The objective of the study was to detect letter that has been said by the speaker thru video processing to determine the highest recognition rate in lip reading analysis by using the 16 key points for point plotting, and also to develop a model that recognize the English letters from A-Z by taking video of college students that are Filipino English speakers. Thus, the objective set were achieved. In order to achieve a better and more accurate lip segmentation in the future, lipsticks on the subjects is suggested to easily discriminate the skin from the lips. This research is deemed significant to serve as a stepping stone for the future researches related to the topic. It will also help in translating the conversation of the people in a video to a subtitle which will be helping the authorities in case of reviewing CCTV footage from a crime scene.

References

1. 1.WenJuan, Y.; YaLing, L.; MingHui, D. In Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on; IEEE, 2010; Vol. 6, pp V6-363.
2. M.Z. Ibrahim, D.J. Mulvaney, Geometrical-based lip-reading using template probabilistic multi-dimension dynamic time warping; In Journal of Visual Communication and Image Representation, Volume 30, 2015, Pages 219-233.
3. Sunil S. Morade, Suprava Patnaik, Comparison of classifiers for lip reading with CUAVE and TULIPS database, In Optik - International Journal for Light and Electron Optics, Volume 126, Issue 24, 2015, Pages 5753-5761.
4. Saitoh, T.; Konishi, R. In Pattern Recognition (ICPR), 2010 20th International Conference on; IEEE, 2010; pp 1356–1359.
5. Huang, Y.; Liang, J.; Pan, B.; Fan, X. In Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on; IEEE, 2010; pp 2402–2405.
6. Okubo, M.; Watanabe, T. In Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on; IEEE, 1998; pp 187–192.
7. Mahmoud, H. A.; Muhaya, F. Bin; Hafer, A. In Future Information Technology (FutureTech), 2010 5th International Conference on; IEEE, 2010; pp 1–4.
8. Summerfield, Q. 1987.
9. Min, K. Y.; Zuo, L. H. In Electronics and Optoelectronics (ICEOE), 2011 International Conference on; IEEE, 2011; Vol. 1, pp V1-115.
10. Bothe, H.-H. In Proceedings of the 1995 ACM symposium on Applied computing; ACM, 1995; pp 450–454.
11. Yingjie, M.; Haiyan, Z.; Yingjie, H.; Jinyang, L. In Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on; IEEE, 2011; Vol. 4, pp 2186–2190.
12. Demetz Clement (2003). Lip recognition Software using a Kohonen Algorithm for image Compression. Computer-Aided Engineering, University of Wisconsin-Madison.
13. Khan, I.; Abdullah, H.; Zainal, M. S. Bin. Int. J. Eng 2013, 3 (4), 8269.
14. Kaur, G.; Chhabra, A. (2014). Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications, 98(22).