Speech Analysis Based On Image Information from Lip Movement

2013 IOP Conf. Ser.: Mater. Sci. Eng. 53 012016

(http://iopscience.iop.org/1757-899X/53/1/012016)

Speech Analysis Based On Image Information from Lip Movement

Kamil S. TALHA[1], **Khairunizam WAN, S.K.Za'ba, Zuradzman Mohamad Razlan and Shahriman A.B**
Advanced Intelligent Computing and Sustainability Research Group
School Of Mechatronics Engineering, Universiti Malaysia Perlis (UniMAP),
02600 Arau, Malaysia

Email: kamilsyahid89@gmail.com, khairunizam@unimap.edu.my

Abstract. Deaf and hard of hearing people often have problems being able to understand and lip read other people. Usually deaf and hard of hearing people feel left out of conversation and sometimes they are actually ignored by other people. There are a variety of ways hearing-impaired person can communicate and gain accsss to the information. Communication support includes both technical and human aids. Human aids include interpreters, lip-readers and note-takers. Interpreters translate the Sign Language and must therefore be qualified. In this paper, vision system is used to track movements of the lip. In the experiment, the proposed system succesfully can differentiate 11 type of phonemes and then classified it to the respective viseme group. By using the proposed system the hearing-impaired persons could practise pronaunciations by themselve without support from the instructor.

1. Introduction

Speech recognition is not purely auditory. When a listener can see the speaker, visual information is used in the speech recognition process. The contribution of this visual information to overall speech recognition was first reported by McGurck and MacDonald [1]. Speech command based systems are useful as a natural interface for users to interact and control computers. Such systems provide more flexibility as compared to the conventional interfaces such as keyboard and mouse. However, most of these systems are based on audio signals and are sensitive to signal strength, ambient noise and acoustic conditions [3]. To overcome this limitation, speech data that is orthogonal to the audio signals such as visual speech information can be used. The systems that combine the audio and visual modalities to identify utterances are known as audio-visual speech recognition (AVSR) system. Visual speech recognition (VSR) system refers to the systems which utilizes the visual information of the movement of the speech articulators such as the lips, teeth and somehow tongue of the speaker. The advantages are that such a system is not sensitive to ambient noise and change in acoustic conditions, does not require the user to make a sound, and provides the user with a natural feel of speech and dexterity of the mouth. This paper is structured as follows: Section 2 addresses the related researches to the approaches, applications and problems of recognizing the human gesture. Section 3 describes the configuration of the system and describes the proposed algorithm for the classification of motion patterns. Section 4 presents the results of the classification and the article is concluded with the summary in section 5.

[1] To whom any correspondence should be address

2. Literature Review

Lip-Reading has been practised over centuries for teaching deaf and dumb to speak and communicate effectively with the other people. The automatic lipreading (or multimodal speech processing in general) quickly becomes a mainstream part of the speech related research. In the recent years some prototype systems have already been presented or announced. The need for lipreading in human computer interaction is no longer questioned as the speech recognition based only on audio signal hits its limits.

2.1. Lip reading

Lip reading, also known as lipreading or speechreading, is a technique of understanding speech by visually interpreting the movements of the lips, face and tongue with information provided by the context, language, and any residual hearing. Each speech sound (phoneme) has a particular facial and mouth position (viseme), although many phonemes share the same viseme and thus are impossible to distinguish from visual information alone. Thus a speechreader must use cues from the environment and a knowledge of what is likely to be said.

Several experiments show that lip-reading can be efficiently applied in limited-vocabulary speech recognition [4,5], recognition of speech uttered by speech impaired [6] and also in case of continuous speech signal [7]. Techniques developed for automatic lip-reading find their way also in the world of computer generated facial animation and multimodal speech synthesis [8]. The lip movements and other visually distinguishable changes in articulatory system are represented by different researchers in multitude of possible geometric and non-geometric models [9,10]. According to Conrad (1979)[11], the capacity for lipreading seems to be determined by the person's degree of hearing and the levels of intelligence and of speaking.

However, the studies that have focused on establishing the relation between the degree of hearing and the level of lipreading have not reached unanimous conclusions. Some have observed that hearing people are usually better lipreaders than deaf people and, therefore, they have concluded that the more the loss of hearing, the more difficult lipreading will be. However, there are 43 phonemes in the English language, while there exist only 28 different mouth shapes that separate them [2]. For example, 'd' and 't', or 'f' and 'v' produce the same mouth shape. Therefore, the art of lip reading for humans is context sensitive: it consists not only in visually recognising mouth shapes, but also mentally recognising key elements to predict the word, as well as further recognising key words to predict the sentence.

Figure 1. Lipreading researched by Leon J. M. Rothkrantz (2006)

Table 1. Fourteen Visemes

Visemes Number	Corresponding phonemes	Vowel/ Consonant
1	/p/, /b/, **/m/**	Consonant
2	/f/, **/v/**	Consonant
3	**/th/**, /d/	Consonant
4	**/t/**, /d/	Consonant
5	**/g/**, /k/	Consonant
6	**/ch/**, /j/, /sh/	Consonant
7	**/s/**, /z/	Consonant
8	**/n/**, /l/	Consonant
9	**/r/**	Consonant
10	**/A/**	Vowel
11	**/E/**	Vowel
12	**/I/**	Vowel
13	**/O/**	Vowel
14	**/U/**	Vowel

3. Methodologies

The system has three stages: preprocessing, tracking and developing database. Figure 2 shows the block diagram of the proposed method. Videos for different movement of lips are captured and Figure 5 shows some of the sample images explained. During the preprocessing, the movie frames are converted into indexed image format. Median filter is applied on these images and the unwanted noises are removed. For each image frame, the images are converted into binary.

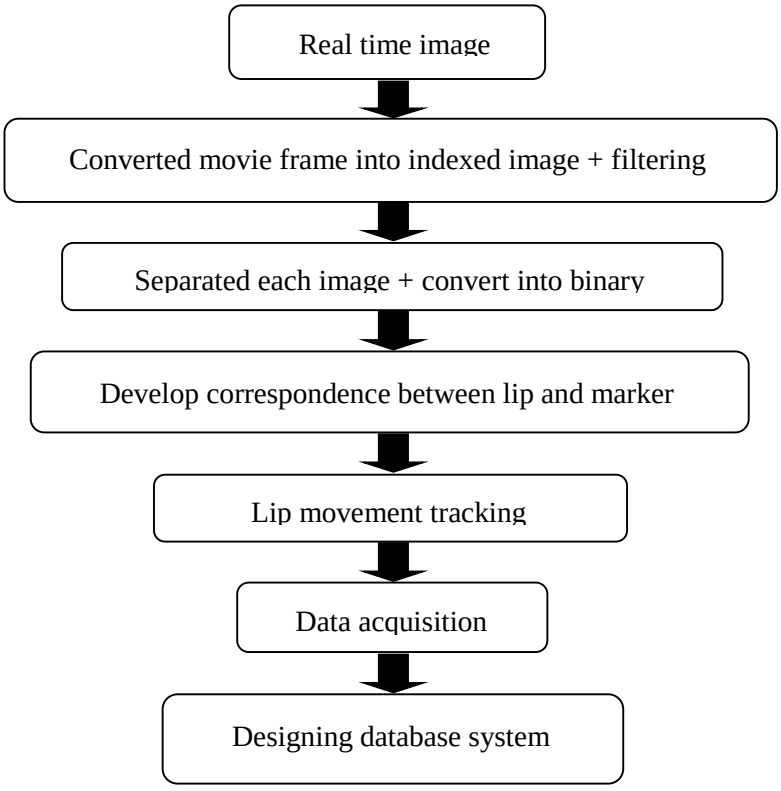

Figure 2. The overall process of the system

3.1. Pre-processing

3.1.1. RGB colour on image
A pixel is a single point in a graphic image. Each pixels of an image contains 3 dimension of colour which is Red, Green and Blue (RGB). The RGB colour is added together in various ways to reproduce a broad array of colours. A RGB decimal value is range from 0 to 255 where will be represented by (0, 0, 0) to (255, 255, 255).
In this project, red marker will be use and will be extracting from other color on an image. The marker will be placed on lip which is 3 on the upper lip and another 3 on the lower lip. The movement of the lip will represent the word that will be spoken. Extracting must be done precisely to eliminate any noise and also to eliminate background colour that can be same colour as the marker.

3.1.2. Intensity image (grayscale image)
This is the equivalent to a "gray scale image" and this is the image we will mostly work with in this course. It represents an image as a matrix where every element has a value corresponding to how bright/dark the pixel at the corresponding position should be colour. There are two ways to represent the number that represents the brightness of the pixel: The double class (or data type). This assigns a floating number ("a number with decimals") between 0 and 1 to each pixel. The value 0 corresponds to black and the value 1 corresponds to white. The other class is called uint8 which assigns an integer between 0 and 255 to represent the brightness of a pixel. The value 0 corresponds to black and 255 to white. The class uint8 only requires roughly 1/8 of the storage compared to the class double. On the other hand, many mathematical functions can only be applied to the double class.

3.1.3. Thresholding
Thresholding is the simplest method of image segmentation. From a grayscale image, thresholding can be used to create binary images. During the thresholding process, individual pixels in an image are marked as "object" pixels if their value is greater than some threshold value (assuming an object to be brighter than the background) and as "background" pixels otherwise. This convention is known as *threshold above*. Variants include *threshold below*, which is opposite of threshold above; *threshold inside*, where a pixel is labeled "object" if its value is between two thresholds; and *threshold outside*, which is the opposite of threshold inside (Shapiro, et al. 2001:83). Typically, an object pixel is given a value of "1" while a background pixel is given a value of "0." Finally, a binary image is created by coloring each pixel white or black, depending on a pixel's labels.

3.1.4. Binary image
A binary image is a digital image that has only two possible values for each pixel. Typically the two colors used for a binary image are black and white though any two colors can be used. The color used for the object(s) in the image is the foreground color while the rest of the image is the background color. In the document scanning industry this is often referred to as bi-tonal.

Binary images are also called *bi-level* or *two-level*. This means that each pixel is stored as a single bit (0 or 1). The names *black-and-white*, *B&W*, monochrome or monochromatic are often used for this concept, but may also designate any images that have only one sample per pixel, such as grayscale images.

3.2. Lip tracking
On this project, the system must track a lip in real-time condition where a real-world process is simulated at a rate that matched that of the real process. In other word, processing the image must be done as the image been captured by camera. considering that camera can captured images at a rate of 30 frames per

second (fps), several image from this frames will be grab to be process. This can be done by using MatLab command below;

```
data=getsnapshot(vid);  where
vid=source of
```

After frame has been grabbed and save, image processing can be done by extracting red colour. To extract the desired colour, each pixels on the image will be examined whether it achieve desired RGB value. For instant, if red colours want to be extract, every pixel that contain RGB decimal value from (0, 0, 0) to (255, 0, 0) will be selected. These selected pixels than will be convert into grayscale colour according to their intensity of red colour. Grayscale is also known as black and white, which are composed exclusively of shades of gray, varying from black at the weakest intensity represented by 0, to white at the strongest represented by 255. This method can be explained through this mathematical equation;

$B_{xy} = A_{xy}$ where; B = gray intensity
x,y = coordinate of pixel

This mathematical equation can be simplify by using MatLab command below:

```
diff_im = imsubtract(data(:,:,1),
rgb2gray(data));
```

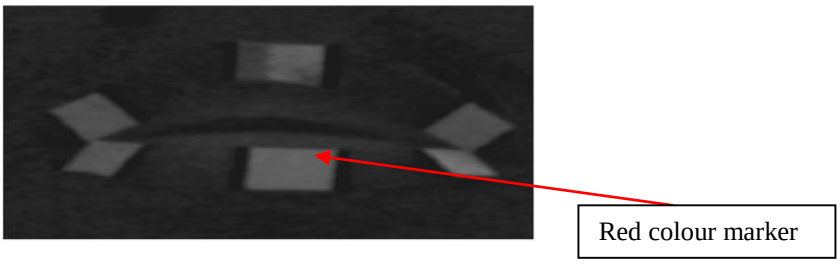

Figure 3. Gray scale image

3.2.1. Red colour extraction

Once the grayscale has been determined, the grayscale image is converted into binary image. By setting factor of binary conversion into certain value, it will remove noise and also background image. Below is the mathematical equation for this method;

$$C_{ij} = \begin{cases} 1, & B_{ij} \geq 200 \\ 0, & \text{else} \end{cases}$$ where C = binary conversion

This equation can be simplified by using the MatLab command below;

```
diff_im = im2bw(diff_im,0.18);
```

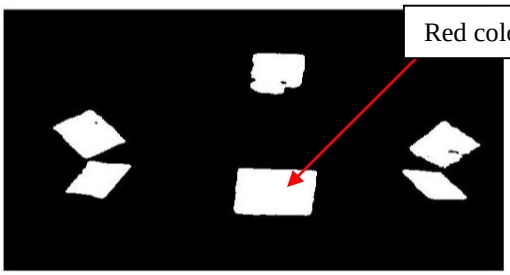

Figure 4. Binary image

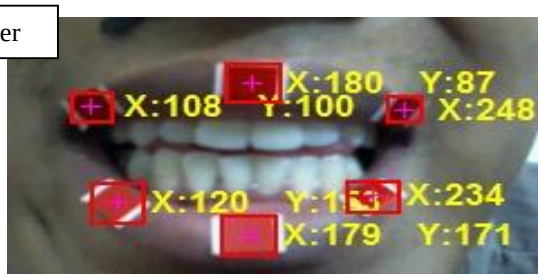

Figure 5. Labeling the image with red rectangular and centroid

3.2.2. Blob analysis

A blob (binary large object) is an area of touching pixels with the same logical state. All pixels in an image that belong to a blob are in a foreground state. All other pixels are in a background state. In a binary image, pixels in the background have values equal to zero while every nonzero pixel is part of a binary object. After the system able to track the red marker, it will now labeling or draw a virtual graphic on the image for further extraction. On this project, red rectangle will be draw on perimeter of red marker. This method also will indicate that the system able to detect red marker when red rectangle is present. MatLab command for this method is;

```
bw = bwlabel(diff_im, 8);
stats = regionprops(bw,
'BoundingBox', 'Centroid');
```

In MatLab command above, there are two image properties analysis that been use which are BoundingBox and Centroid. BoundingBox is the bounding boxes around the blobs in a binary image. This function extracts the coordinates and dimension values of each blob from stats.BoundingBox structure one by one and draws the rectangle around them. Centroid is function return X and Y coordinates.

4. Experiment

4.1. Subject

In this project, only one subject had been tested to doing the corresponding phoneme. Before start the test, on the subject lip need to put a red color marker. After that, subject need to sit in front of the laptop and the lips must be approximately in front of the webcam. The webcam need to focusing to on the lips so that the system may determine the color marker. There are 11 phoneme had been tested in this experiment. The result from the phoneme then we determine it pattern to show which phoneme is in the same group and the difference pattern to the other group.

Table 2 - Five visemes tested

VISEMES	PHONEMES
1	p /b / m
2	f / v
3	t / d
4	g / k
5	s / z

The above table is the five visemes that been tested in this project. From this five viseme tested it may determine the corresponding phoneme group.

4.2. Testing

When performing the test, 5 to 10 time of test for every phoneme conducted. The value from the phoneme was calculated to get it average value. First a comparison among the test of same phoneme will be conducted. After had satisfied, the average value of the phoneme then been compare to the others phoneme.

4.3. Experimental result

The following result is from the experiment. The result divided to 4 type of viseme and every viseme group contain it corresponding phoneme. In this result, it shows the X-coordinate and Y-coordinate graph. For more detail and to show it pattern more clearly, for every graph had been include with staked line graph. Stacked line graph is the type of graph that categorizes the entire phoneme by alphabetical sequence.

4.3.1. Viseme 1

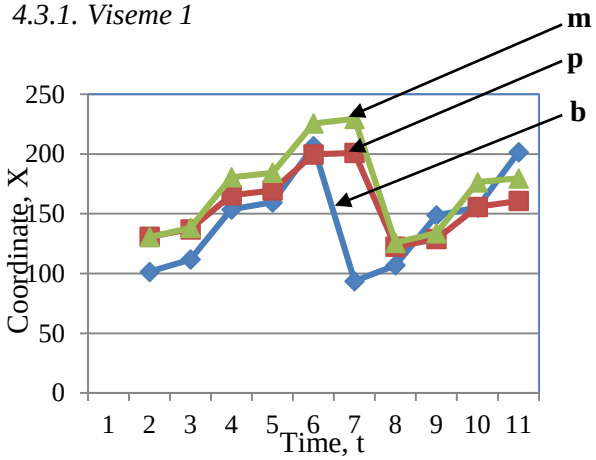

Figure 6. Trajectories of lip for p/b/m

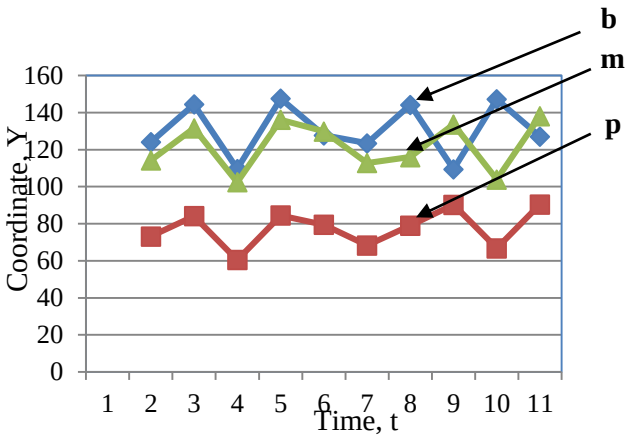

Figure 7. Trajectories of lip for p/b/m

4.3.2. *Viseme 2*

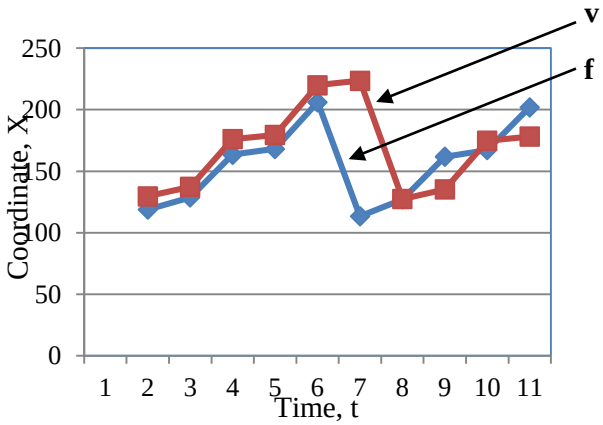

Figure 8. Trajectories of lip for f/v

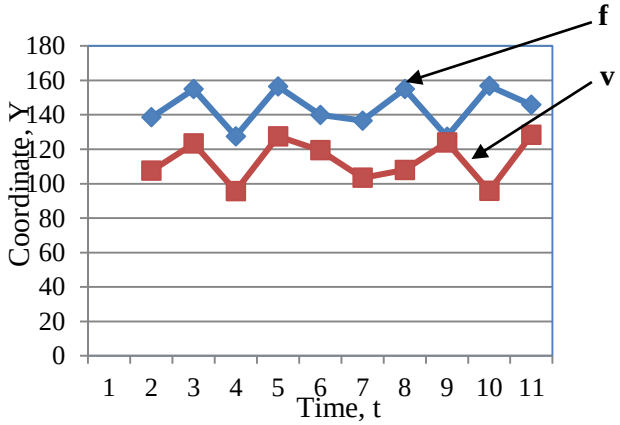

Figure 9. Trajectories of lip for f/v

5. Conclusion

Overall for this project, the concept of image or video analysis had been taken seriously in determining the success of this project. From starting of this project, many concepts to analyze of image had been done. Several techniques such as colour marker detection, optical flow analysis and snake tracking been taken look. After making a research, the colour marker detection is much easier and much more reliable to determine the movement of the lip. The other technique more reliable to be use for others application and for different situation.

From the research and experiment that had been done, the result may be use to make a more reliable lip reading system. Factor like intensity of light in the test environment need to be consider. Besides that, when operating with the system that need more space of memory, our device of the laptop been use in this experiment need to be consider more. From the test had been done, after making several of test, the laptop may have a problem to processing the image. It may come out the result with an error or sometime the laptop itself will automatically shut down because of the problem of memory dump.

For determining the successful of the project, a lot of researches need to be done and a lot of tests need to be working on. When facing with the signal processing concept, it needs a lot of understanding of the concept and need to analyze a lot of mathematical problem to determine our flow of the project.

6. References

[1] D. McGurck and J. MacDonald. Hearing lips and seeing voices. *Nature,* 264, December 1976.
[2] A. Greenwald. *Lipreading made easy.* Alexander Graham Bell Association for the Deaf, 1984.
[3] T. Chen, "Audiovisual speech processing. Lip reading and lip synchronization," *IEEE Signal Processing Mag*., vol. **18**, January 2001, pp. 9-21.
[4] S. Lucey, S. Srindharan and V. Chandran (2001): "An investigation of HMM classifier combination strategies for improved audio-visual speech recognition," In: [*Dalsgaard et al. (2001) Dastard, Lindberg and Benner*], pp. 1185-1188.
[5] J. C. Wojdel and L. J. M. Rothkrantz (2001): "Using aerial and geometric features in automatic lipreading," In: *[Dalsgaard et al.(2001)Dalsgaard, Lindberg and Benner],* pp. 2463-2466.
[6] G.Potamianos and C.Neti (2001): "Automatic speechreading of impaired speech," In: *[Massaro et al.(2001)Massaro, Light and Geraci]*, pp. 177-182.
[7] P. Wiggers J. C.Wojdel and L. J. M. Rothkrantz (2002): "Medium vocabulary continuous audiovisual speech recognition, " In: *[Hansen and Pellom(2002)]*, pp. 1921-1924.
[8] T.Chen, (January 2001): "Audiovisual speech processing," In *IEEE Signal Processing Magazine*, pp.9-21.
[9] M. E.Hennecke, D. G. Stork and K. V.Prasad(1996): "Speechreading by humans and machines," In: *[Stork and Hennecke(1996)]*, pp. 331-349.
[10] G.Potamianos, C. Neti, G. Gravier et. al(2003): "Recent advances in the automatic recognition of audiovisual speech," In *the Proceedings of IEEE*, Volume **91**, No. 9, pp. 1306-1388.
[11] R.Conrad(1979): "The deaf school child: Language and cognitive functions," London*: Harper & Row*.