An Extraction Method of Lip Movement Images from Successive Image Frames in the Speech Activity Extraction Process

Eung-Kyeu Kim, Soo-Jong Lee, Nohpill Park

An Extraction Method of Lip Movement Images from Successive Image Frames in the Speech Activity Extraction Process

Eung-Kyeu Kim*, Soo-Jong Lee**, Nohpill Park***

*Dept. of Infor. & Commu. Engineering, Hanbat Nat'l Univ., Korea
kimeung@gmail.com

**Automatic Speech Translation Research Team, ETRI, Korea
sjleetri@etri.re.kr

***Dept. of Computer Science, Oklahoma State University, USA
npark@cs.okstate.edu

Abstract. In this paper, we propose an extraction method of lip movement images from successive image frames and present the possibility to utilize lip movement images in the speech activity extraction process of speech recognition phase. The image frames are acquired from the PC image camera with the assumption that facial movement is limited during talking. First of all, one new lip movement image frame is generated with comparing two successive image frames each other. Second, the fine image noises are removed. Each fitness rate is calculated by comparing the lip feature data as objectly separated images. It is analyzed whether or not there is the lip movement image through verification to the objects and three images which have higher rates in their fitnesses. As a result of linking the speech & image processing system, the interworking rate shows 99.3% even in the various illumination environments. It was visually confirmed that lip movement images are tracked and can be utilized in speech activity extraction process.

Keywords: Lip Movement image, Image Frames, Acoustic Noises, Speech & Image processing system.

1 Introduction

In recent years, information technologies have spread rapidly due to the miniaturization and increased mobility of these information appliances. These changes have made it necessary to develop an speech interface technology that can effectively control these information appliances. The most difficult obstacle in the speech recognition phase, through technology which converts speech into text, is acoustic noises. The service environment of speech recognition is full of acoustic noises. It is no exaggeration to say so. The process of speech recognition is plagued by a large variety of acoustic noises including noises from its own drives of the

appliances, network noises, and other environmental noises. Channel noises or stationary noises are almost completely eliminated, because their sizes and frequencies are easily identified by their consistency. The real challenge lies in the elimination of more dynamic noises, which are more difficult to be identified due to their irregular sizes and frequencies. Once the acoustic noises entered into the speech recognition phase, the noises are not removed and are the main cause for the low speech recognition rate. By the way, regardless of the ambient acoustic noises, images will be acquired and processed continually.

This paper is a part of the scheme to utilize lip movement image in order to prevent acoustic noises from being recognized as speech[1]-[2]. There is certainly lip movement whenever someone speaks, and any sound energy is extracted. Also, the sound energy which is not correspondingly associated with lip movement would be identified as noises. If lip movement is confirmed[3]-[4] in the speech activity extraction process of speech recognition phase, lip movement can efficiently prevent the acoustic noises from being classified as speech.

In this paper, we propose an extraction method of lip movement images from successive image frames in the speech activity extraction process[5] which is preprocessing phase of speech recognition. The image frames are acquired from the PC image camera. We also present the possibility to utilize lip movement images in the speech activity extraction process of speech recognition phase. Ultimately, we try to apply this extraction method to speech recognition of Robert surrounded with any outside environment including the dynamic acoustic noises.

This paper is organized as follows. Section 2 describes image frame acquisition environments. In section 3, the procedure of lip movement image extraction is described including the noise image deletion and the extraction of lip movement image features. In section 4, we show the verification of lip movement images and it's experimental environment, where the template matching method is showed to verify lip movement images accurately. In section 5, we bring to conclusions.

2 Image frame acquisition environments

As the multimedia environment has progressed rapidly, cameras are equipped with all kinds of information appliances, which can easily, acquire movement images, and process them. Especially image cameras for PCs have come into use widely in the recent years and those allowed the general public to generate images directly, and to edit them. The PC image camera is mainly used for image communication with a remote site or image chatting, but it can also be utilized continuously in acquiring, comparing and analyzing the image frames. Nowadays, the PC image camera generates 320*240 sized color images at the minimum speed of more than 15 frames per second. Thus it is sufficient to study the process of acquiring and processing image frames.

Fig. 1 shows speech and image processing system in the speech recognition phase. The processing system is composed of three parts, that is, image processing part, and speech processing recognition part, and shared memory part. Lip movement images can be acquired under the lighting conditions of the general home and office

environment. It is also assumed that there was no excessive movement of the head or face during the conversation. Therefore, excessive movements of a face were excluded from this study.

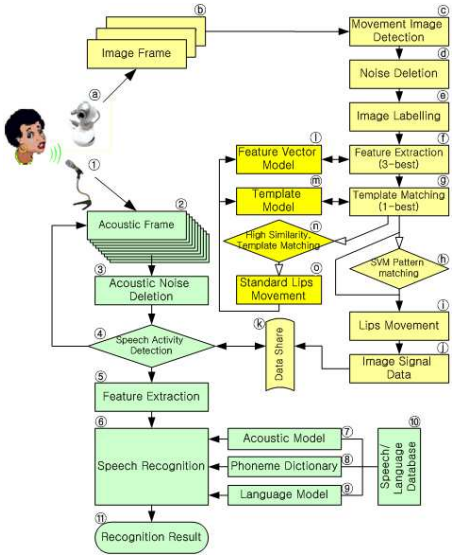


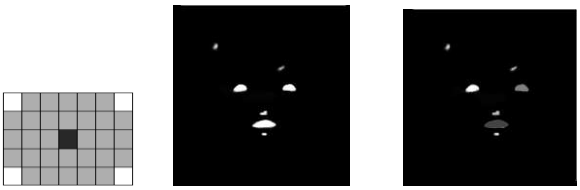
Fig. 1. Speech and image processing system in the speech recognition phase

3 Procedure of lip movement image extraction

The extraction of lip movement images uses a five phase processes, which is as follows: acquiring movement images, removing fine image noises, separating the partial images as different areas, comparing the area images to lip features, and verifying whether it is the lip image or not, etc. This process is repeated for each image frame unit.

Figure 2 shows five steps of the process for extracting lip movement image. Image frames “a” and “b” are the shape of the facial components when they are changing in the speaking situation.

(a) (b) (c)



(d) (e) (f)

Fig. 2. Process for extracting lip movement image

"c" is composed of the results which are changed between "a" and "b". As a structuring element "d" is used to remove minute image noises which appear in "c". "e" is the result of the application of the structuring element. "f" is the result of the different image areas, which have different pixel value. The following sections address the specific processes involved in the completion of component "f".

3.1 Movement distinction and it's image acquisition

In order to distinguish whether or not there is a lip movement, a comparison between one image frame and the other image frame must be done in a pixel unit. A fine difference of pixel value in each frame is constantly occurring phenomenon because of the image processes and illumination. Therefore, after setting up a threshold for a pixel value difference, it is desirable to distinguish those values which are in excess of the threshold. The setting of a threshold value is necessary to consider environmental illumination and PC computing performance. Here the threshold was set to 30. Therefore, the decrease of the computational scale was processed after the color images were converted to monochrome

In order to acquire and compare the successive image frames, a minimum of two image buffers are necessary. While continuously updating the new images, they are simultaneously compared to the buffer images by pixel unit each other. Whether or not there are image frames or their components change is interpreted according to the differences in the related pixel value. The results which are compared with a pixel unit are acquired, and those are composed of new movement image frames.

3.2 Image noise removal and movement image separation

The fine image noises are removed by the "opening"[6] technique as one of the morphological image processing method. The "opening" technique is composed of a two step technique, "erosion" and "dilation". By applying the structuring element of Figure 2(d), to images smaller than the structuring element, the smaller images are removed. The structuring element used in "opening" can be set up with several directions, that is, up and down, right and left, diagonal, and so on. This process is expressed by their symbols.

$$e = c \circ d = (c \ominus d) \oplus d$$

The structuring element ("d") is set up as 7*5 pixel size with a length that is longer than the width that takes into consideration the lip movement image shape. The lip movement image width is shorter than the length.

Nevertheless after removing the detailed image noises, many other large sized movement image parts will remain. These image parts include eye blink unrelated to eye movements, jaw movements, face shake, and many extra movements in the background and the light. In addition, there are lots of extra images according to the background and the light.

In order to isolate the lip movement image, it is necessary to separate those residual image parts into different areas each other. The separation of the image part is performed by the "grassfire transform"[7] technique. Through all the image pixels, it makes groupings of the neighborhood pixels with the same value by means of giving a distinct value to each.

3.3 Application of lip movement image features

After the objects are separated into different areas, a variety of data can be extracted, such as width, lengthwise, pixel numbers, position, etc. If a comparative analysis of the data is done, the characteristics of lip image can be determined.

Several features of lip movement images can be determined mainly through it's relation to the eyes. First, it is the extent of the rectangular form. Second, it is done by finding the center of the image in relation to the eyes and the lip. Third, it is the number of pixels. Although the pixel numbers vary largely according to the lip movement scales, the maximum and/or minimum counts are important to distinguish each image area.

Table 1 below is the results of the data for the feature elements of lip movement images which are collected through a PC image camera at the interval of 50cm.

The dimension rate is calculated as the actual pixel numbers divided by the rectangular dimension's maximum size for width and length.

The fitness rate of the lip movement feature is calculated by combing all the feature data above. The more the fitness rate of the image is high, the more the possibility of the lip movement image part is great. The next section will show the method of verification by identifying the three images which have the highest fitness rates.

Table 1. Feature elements of lip movement images

	width	legnth	width/leng.	dimension rate	pixels	length location	width location
collected data (distance:50cm)	6	24	3.630	0.436	330	0.733	0.812
	5	28	4.140	0.426	329	0.873	1.869
	3	23	4.800	0.322	214	0.707	0.932
	4	23	5.400	0.335	275	0.727	0.935
	6	28	4.667	0.438	384	0.669	0.947
	5	24	4.800	0.360	333	0.736	0.994
	8	31	3.875	0.492	504	0.707	1.424
	4	18	4.500	0.346	208	0.747	1.194
	7	19	2.714	0.652	204	0.480	1.011
	3	4	3.667	0.327	101	0.600	2.404
	4	17	4.500	0.393	183	1.013	0.980
	2	20	3.000	0.324	37	0.667	1.014
	6	18	4.667	0.403	417	1.136	0.938
	3	19	3.333	0.337	89	0.693	1.392
	.....	.....	.....	.....	.....	.....	.....
average	4.6	18.1	4.284	0.395	251	0.781	1.282
standard dev.	1.453	7.053	0.776	0.070	121	0.180	0.399
maximum	8	31	6.250	0.652	504	1.175	2.404
minimum	2	4	2.714	0.286	37	0.480	0.812

4 Verification of lip movement images and experimental environments

The preliminary step in the verification process should be the establishment of threshold values for the fitness rates of a lip movement image. However, the fitness levels for eyes and lip movement images are changeable on occasion.

The fitness rates are made as the result of the correlation rate from the template matching, which is added to the fitness rate by feature elements. The fitness rates are used as a final score to decide the lip movement image. The following Figure 3 shows the partial results that were collected from the largest verification rate per every frame.

The horizontal axis represents the fitness rates and vertical axis represents the frequencies. The highest convex curve is related to the lip image areas, and the second convex curve is a related to the eyes. The valley point between the first and the second convex curve is a threshold classifying the lip image area and the others. The threshold value is formed and extracted automatically. The threshold value is situated at the point between two convex curve changes. where, valley point is converged into a unified value, namely 0.57.

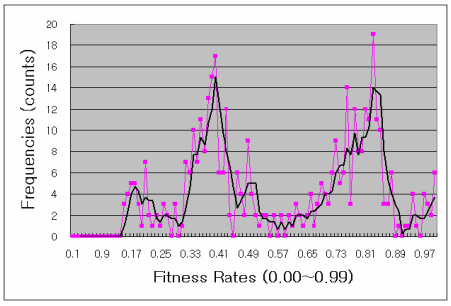
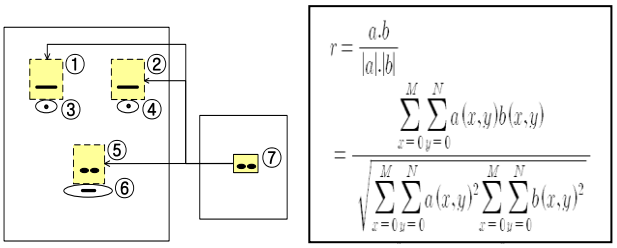


Fig. 3. Fitness rates and frequencies

The next step is considering the template matching technique which can be used for accurate verification. Template matching is a suitable method for measuring the correlation between two images in a pixel unit. The correlation can compute the differences between the two images per related pixels. Fig. 4 shows the model (a) and the calculating formula (b) for measuring the template matching rate.



(a) (b)

Fig. 4. Template matching rate measuring model (a), calculating formula(b)

In Fig. 4(b), a(x,y) shows the brightness value considering the average value of E(g) obtained from the input image g(x,y). Additionally, b(x,y) shows the brightness value considering the average value of E(t) obtained from the template image t(x,y). During the implementation of the template matching technique, only three movement images which are selected in order of their fitness levels are actually verified.

The lip movement itself should not be included as an element in the template during talking. Unlike the lips, the nose is unchangeable even during talking, so the difference between the light and dark shade is distinct.

Figure 5(a) shows tracking results of three lip movement image candidates and 5(b) does tracking of a final lip movement image which is selected from three lip movement image candidates.

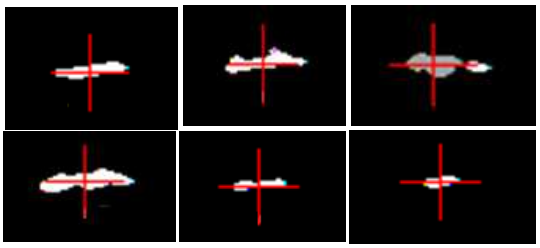


Fig. 5(a). Trackings of three lip movement image candidates

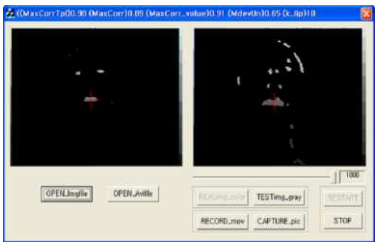


Fig. 5(b). Tracking of a final lip movement image

Prior to the combined speech and image processing system, it was tested whether the speech recognition phase is influenced by external acoustic noise which is not a part of the speech recognition object. The confirmation of illumination environment was investigated while the brightness of the lighting application was changed. Additionally, it was determined whether any acoustic noise was blocked during the speech recognition phase through the combination of the speech and image processing system. These experimental results are summarized in table 2, and table 3. This system was implemented by linking the existing speech recognition engine with the image processing test bed for lip movement image tracking. The computer utilized for the experiment was a PC Pentium IV with a 3.6GHz processor under the general office environment.

Table 2. The image processing and illumination adaptation

Tracking object	Illumination level (lx)	Success rate (%)
Lip movement	Office (500~300) Home (300~100) Laboratory (100~)	About 95

Table 3. Combined speech and image system

Input object	Content	Success rate (%)
Acoustic noise	Blocking of the speech recognition progress	100
Speech utterance	Speech recognition execution	99.3
	Speech recognition non-execution	0.7

5 Conclusion

In this paper, we proposed a method to extract lip movement images from successive image frames and presented the possibility to utilize lip movement images in the speech activity extraction process of speech recognition phase. The image frames are acquired from the PC image camera with the assumption that facial movement is limited during talking. First of all, one new lip movement image frame is generated with comparing two successive image frames each other. Second, the fine image noises are removed. Each fitness rate is calculated by comparing the lip feature data as objectly separated images. It is analyzed whether or not there is the lip movement image through the verification to the objects and three image candidates which have higher rates in their fitnesses. As a result of linking the speech and image processing system, the interworking rate showed 99.3% even in the various illumination environments. It was visually confirmed that lip movement images are tracked and can be utilized in speech activity extraction process. Ultimately, we try to apply this

extraction method to speech recognition of Robert surrounded with any outside environments including the dynamic acoustic noises. In the future study, the confirmation of the real-time processing for extracting further robust lip movement images remains to be solved.

References

1. M.T. Chan, Y. Zhang, and T.S. Huang, "Real-Time Lip Tracking and Bimodal Continuous Speech Recognition", IEEE Second Workshop on Multimedia Signal Proceeding, 65-70, 7-9 Dec. 1998.
2. G. Potamianos, C. Neti, J. Luettin and I. Matthews, "Audio-visual automatic speech recognition: An overview, " in issues in Visual Speech Processing, MIT Press, 2004.
3. A.W. Liew, S.H. Leung, and W.H. Lau, "Lip contour extraction from color images using a deformable model, Pattern Recognition", Vol.35, No.12, pp.2949-2962, 2002.
4. Visual Speech Recognition: Lip Segmentation and Mapping, A. Liew and S. Wang, editors, IGI Global, 2009.
5. V. Libal, J. Connell, G. Potamianos, and E. Marcheret, "An embedded system for in-vehicle visual speech activity extraction, " In Proceedings of the International Workshop on Multimedia Signal Processing(MMSP 2007), Chania, Greece, pp.255-258, 2007.
6. Z.Q. Wu, J.A. Ware, W.R. Stewart, and J. Jiang, "The Removal of Activitying Effects Caused by Partially Overlapped Sub-activity Contrast Enhancement", Journal of Electronic Imaging, Vol.14, Issue 3, 033006(8 pages), July-Sept. 2005.
7. F. Leymarie and M.D. Levine, "Simulating the Grassfire Transform Using and Active Contour Model". Trans. IEEE Pattern Analysis and Machine Intelligence, Vol.14, No.1, 56-75, 1992.