



# Preventing an External Acoustic Noise from being Misrecognized as a Speech Recognition Object by Confirming the Lip Movement Image Signal

Soo-jong Lee<sup>1</sup>, Jun Park<sup>2</sup>, Eung-kyeu Kim<sup>3</sup>

<sup>1,2</sup>Automatic Speech Translation Research Team, Speech/Language Information Research Center, ETRI, Korea

<sup>3</sup>Division of Information Communication & Computer Engineering, Hanbat National University, Korea

{sjleetri, junpark}@etri.re.kr, kimeung@hanbat.ac.kr

## Abstract

This paper describes an attempt to prevent an external acoustic noise from being misrecognized as a speech recognition object by confirming the lip movement image signal of a speaker as well as the analysis of the acoustic energy in the speech activity detection procedure, which is the preprocess phase of the speech recognition. An image camera for a PC is added to the existing speech recognition environment, and the collected image is analyzed to capture the movement of lips and classify whether it is acoustic speech made by a human or not. It is possible to determine to continue the recognition process based on the confirmation result of image signal data stored in the shared memory.

We combined a speech recognition processor and an image recognizer, and the interworking function successfully operated at the rate of 99.3%. In the case of a subject facing the image camera and speaking, processing normally progressed to the output of the speech recognition result. However, the speech recognition result was not obtained without facing the camera, since the acoustic energy is regarded as noise if any lip movement is not confirmed.

**Index Terms:** multimodal, external noise, lip movement

## 1. Introduction

Basically, a speech recognition function analyzes acoustic energy, but various acoustic noises exist in practical circumstances, especially dynamic acoustic noise, which can appear suddenly. Thus, an effective way to process these dynamic noises is becoming indispensable for actual service. In particular, for continuous speech recognition such as Non-PTT (Push To Talk), which cannot artificially regulate recognition time, acoustic noise prevention countermeasures are essential for accurate speech recognition.

We always make our lips move when talking. A method of speaking without moving lips practically does not exist. Moreover, an image signal can be obtained and processed without being concerned with the acoustic noise. Therefore, in the acoustic energy analysis process for speech recognition, the introduction of the lip movement image signal can be an effective countermeasure for acoustic noise processing.

Thus far, lip-reading to utilize these advantages has been continuously studied. However, it remains only in the case of recognizing speech signals based on the lip shape in extremely noisy surroundings and when vowels and some parts of words whose mouth shapes are easily distinguished [1][2][3]. It is not just because the attributes of image signals are different from

those of speech, but also because the throughput of images is abundant in comparison with speech.

This paper is mainly about the strategies to avoid recognizing the external dynamic acoustic noise as speech recognition objects. The function confirming lip movement in the image signal was added to the process of speech activity detection, and it was implemented by Visual C++. In order to minimize the additional computation load, the speech recognizer and the image processor were indirectly connected and executed independently. In the speech recognition procedure, the go-stop is determined according to the existence of lip movement. Any type of acoustic energy without a movement is considered to be noise, and even if it has a movement, the lip movement should be distinguished from other movements.

## 2. System overview

Figure 1 is an illustration showing how the image processor (yellow part) is combined to the existing speech recognition processor (light green part) in order for the lip movement image signal to be confirmed in the speech activity detection process. In addition to the current mode in which acoustic energy is obtained via a microphone and analyzed in the recognition process, the image of the frontal speaker or object through the PC image camera is an input. The data are analyzed at the same time. Once a movement is captured, it is determined whether it is a motion of lips or not, and the result is utilized to decide whether the acoustic energy inputted to a microphone is a noise. The major procedure is outlined as follows.

At first, the image frames (b) are continuously taken through the camera (a). The pixel values are compared in the taken image frames [4]. The fresh movement image frame (c) consisting of the compared values is generated. The minute noise images included in the movement image frame are removed (d) and the residual movement images are separated from each other (e) [5]. Each movement image feature value is extracted (f) from these separated images. Also, there is a stage to distinguish (g) lip motion among them. It is comprised of two important steps. The first step compares the feature value (f) produced in the above step and the lip movement image feature model value (h) obtained in advance through off-line and selects 3 which have high similarity. The second step is computing the template matching rate (i) of these 3 movement images, and then the image with the highest matching rate is chosen. If the matching rate is greater than or equal to the critical value, it is classified as a lip movement image. The template image used here is prepared off-line in advance. And the critical value for the discrimination is automatically generated according to the

distribution data of the template matching rates which are accumulated while passing through the matching procedure. The feature value similarity and template matching rate of the movement images which have the highest matching rate are combined (j) and stored in the shared memory (k).

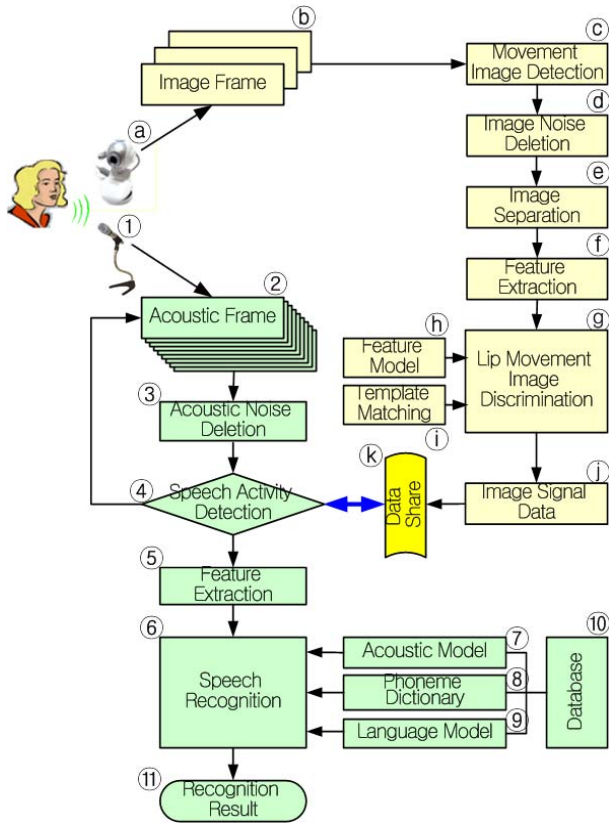


Figure 1. The interworking construction diagram of the speech recognition processor and the image processor.

Meanwhile, what enters the microphone is not only the speech of a man for speech recognition but also various types of acoustic energy (1). By dividing acoustic energy data into frames of appropriate sizes (2) and passing it through the filtering process, static noise which has a constant size and high frequency (3) is removed. Subsequently, if the magnitude of the acoustic energy, persistence and lip movement image signal data (k) are confirmed so that it is determined to be the acoustic energy made by speech, then it is labeled as a speech activity interval (4). In the speech activity interval, features are extracted (5) and the speech recognition (6) is performed. The acoustic model (7), the phoneme dictionary (8) and the language model (9) are constructed from the audio/text data base (10) in advance. In the speech recognition process (6), the speech feature values (5) and these parameters are compared, searched and combined so that the recognition results (11) are obtained.

The newly added image processing and some functions attached to the speech recognition processor are described in detail in the next chapter. In the image processing part, the lip movement feature value calculation, template matching and critical value extraction are mainly described. Regarding the

speech recognition processor, the function which is added to confirm the image signal data is mentioned in the chapter on the system interworking.

### 3. Lip movement image signal extraction procedures

In order for the lip movement image signal to be utilized effectively in the speech activity detection procedure of the speech recognition, we should be able to discriminate the lip movement exactly.

Figure 2 visually shows the process of extracting the lip movement image. The image frame “a” and “b” illustrate changes of the facial component during speaking. The “c” frame is that which is generated by extracting all differences between “a” and “b”. Frame “d” enlarges the image noise removing filter. The minute image smaller than 5x7 size among the images generated in “c” is removed by “d”, the image noise deletion. “e” is the outcome of the noise elimination. “f” is the result of separating and labeling images in “e”, and classifying visually by giving different pixel values. Then, 3 movement images which have the highest similarity with the lip movement image feature are chosen among the movement images in “f”, and “g” is the result. “h” is to show that we process the mapping from the locations of the movement image in “g” into “a”. As a partial image of a nose, as shown in “h”, “i” is used for template matching with the upper side of each movement image at “g”. Because of having the highest rate in the template matching result at “h” and showing the matching rate higher than the specified level (critical value), “j” is selected as the lip movement image.

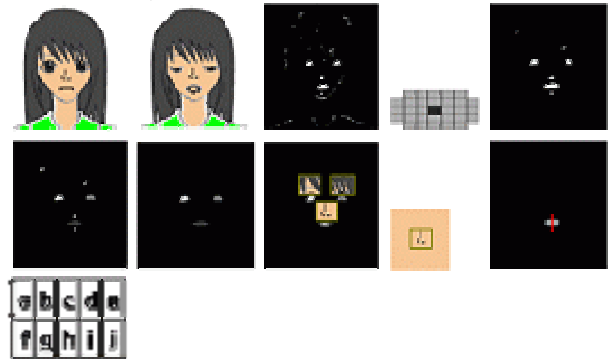


Figure 2. Lip movement image extraction procedure

#### 3.1. Successive image frames acquisition, and movement image detection

The PC camera obtains an image at least to the speed of 15 frames per second. The size of a screen is 320x240 (width x vertical, pixel number). In the situation where the image frame is successively received, the movement images are captured by analyzing value change of brightness in each pixel between the adjacent frames. The relation can be shown by the following formula (1).

$$d_{ij}(x, y) = \begin{cases} 255 & \text{if } |f(x, y, t_i) - f(x, y, t_j)| > T \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

In implementation, the threshold was set up as 10. In order that the brightness difference is utilized as the feature value of the lip movement image, the absolute value  $|f(x, y, t_i) - f(x, y, t_j)|$  was used instead of 255.

### 3.2. Movement image separation and lip movement image feature extraction

Movement images can appear in various areas of our body, i.e. body movement, face movement, eye blinks, chin movement, etc. In particular, blinks of the eye cannot be controlled artificially. Therefore, such movements must be singled out separately in order to distinguish them from lip movement. Image separation was applied by the grassfire technique [5]. After extracting the feature values from each separated movement image, they are compared with the lip movement image feature value set up off-line in advance and their similarity is calculated. We used the width and height, the rate of these, dimension, average pixel value, relative location on the space coordinates, etc. as the movement image features. In off-line, lip movement image features set up in advance were collected at 50cm distance from a camera. The several feature vectors of the lip movement image are shown in the following table 1.

Table 1. Lip movement image features

Feature vector	(1) Length	(2) Width	(3) Width/ Length	(4) Dimension Rate	(5) Pixel Value	(6) Length Location	(7) Width Location
Average	5	20	3.95	0.37	16	1.03	0.68
Standard Deviation	2.18	6.88	0.87	0.07	3.92	0.18	0.1

As mentioned above, each movement image feature is compared with the lip movement image feature and the similarity is measured. Formula (2) summarizes the method of how similarity is measured.

$$(M_i) sim = \sum_{j=1}^k \left( -0.1 * \frac{|M_{i(j)} - L_i avg_{(j)}|}{L_i std_{(j)}} + 1 \right) * w_{(j)} \quad (2)$$

In Formula (2),  $M_i$  and  $L_i$  present the feature value of movement image and lip movement image respectively. The “j” is an index of feature element. The “avg<sub>(j)</sub>” and “std<sub>(j)</sub>” abbreviate the average and standard deviation of a feature value of each element. The “sim” stands for similarity. The feature weighted value “w<sub>(j)</sub>” was not set up, just made equal. Table 2 is an example of similarities calculated between each movement image, which can be generated as the various image frames appear, and the lip movement image feature vector.

Table 2. Similarities of movement images for each movement image frame.

	M <sub>1</sub>	M <sub>2</sub>	M <sub>3</sub>	M <sub>4</sub>	M <sub>5</sub>	M <sub>6</sub>	...
F <sub>1</sub>	0.38	0.06	0	0	0	0	
F <sub>2</sub>	0.54	0.40	0.38	0.28	0.18	0	...
F <sub>3</sub>	0.93	0.75	0.66	0.56	0.31	0.29	...
F <sub>4</sub>	0.84	0.76	0.72	0.56	0.29	0.23	...
...	...	...	...	...	...	...	...

In Table 2, F<sub>i</sub> shows the movement image frames. M<sub>i</sub> shows the movement images in each movement image frame. Its range is from the domain (M<sub>1</sub>), in which a similarity with the lip movement image feature is the highest, to (M<sub>6</sub>) where the similarity is the lowest. Sometimes the similarity is altogether low (F<sub>1</sub>), so it is determined that the lip movement image is not included. Occasionally, a frame is seen to have a high similarity (F<sub>3</sub>) more than 0.90, and it is also a difficult case (F<sub>2</sub>) to judge the inclusion of the lip movement image. Therefore, in order to judge whether the lip movement image is included more accurately, an additional analysis is needed besides the similarity measurement.

### 3.3. Template matching and lip movement image detection

Among the facial components, a part image of the nose which does not have much change in shape and size, and shows clear contrast of brightness, was used as a template image. Figure 3 illustrates the distribution of fitness rates between the movement image and the template image and their frequencies.

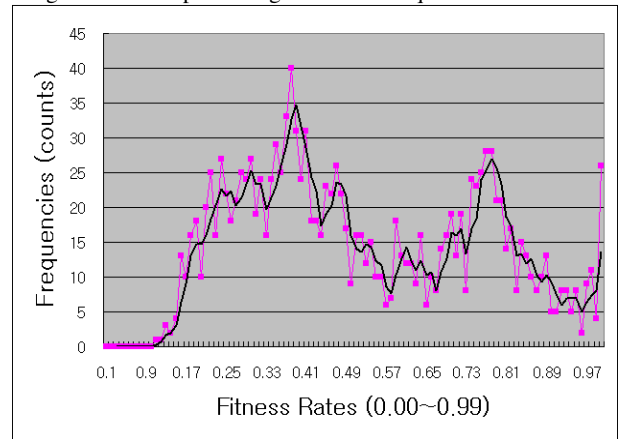


Figure 3. Template matching rate distribution

In the above picture, it can be seen that the matching rate of the lip movement image and the matching rate of the other facial components movement image are distinguished. The matching rate (here, 0.57) showing the lowest frequency between the convex curve of the right side and that of the left side is the critical value which classifies the movement of lip and the other facial components. The critical value is automatically drawn on a real time basis.

### 3.4. Two stage detection

Even in the case of an image actually being a lip movement, the similarity with the lip movement image feature may be a bit decreased due to the fact that there are various ways how the size and shape of the lip movement change. As many as possible of these images have to be searched for. Therefore, all moving areas (M<sub>1</sub>~M<sub>6</sub>) including the movement image which was not included in the top 3 of the similarity rank in Section 3.2 were reaffirmed around the location showing the high conformity.

### 3.5. Brightness compensation

In a specific illumination surrounding, the discrimination ability is quite poor since the brightness distribution is biased. To prevent this problem, the brightness distribution was expanded

to the range of 0~255[6]. The mapping function for stretching is the next formula (3).

$$V_{new}(x, y) = \frac{V_{old}(x, y) - V_{min}}{V_{max} - V_{min}} * 255 \quad (3)$$

$V_{old}(x,y)$ ,  $V_{min}$ , and  $V_{max}$  show the original pixel value, minimum value and maximum value respectively.  $V_{new}(x,y)$  contains the pixel value after the conversion.

## 4. Image and speech processor interworking

### 4.1. Speech activity detection function expansion

In the speech activity detection module of the speech recognition processor, the function of evaluating the image signal value was added. If the lip movement image is detected so as to minimize the effect of the image detection error on speech recognition, the result is carried through to several frequencies.

### 4.2. Interworking function execution

The image processor and speech recognition processor are independently executed, being indirectly connected. In the environment in which two processors are performed simultaneously, the interworking function is activated. The image processor records the movement image signal value in the shared memory. The speech recognition processor performs the speech activity detection while checking the shared memory.

## 5. Experiments

Data which can be extracted from the execution procedure of the image processing such as lip movement image, feature similarity, template matching rate, the critical value, etc. were visually confirmed. Under everyday illumination surroundings, it was checked whether or not it progressed properly until it reached the output of the speech recognition result. The PC Pentium IV, 3.6GHz computer environment conducted this experiment.

### 5.1. Confirming the lip movement image detection

The success rate of the lip movement image detection reached 95%. Among the facial components, the element inducing the most detection errors were the eye blinks. Table 3 summarizes the results described above.

Table 3. Lip movement image detection experiment

Illumination surrounding	General home / office environment (100~500 lx)
Distance (speaker and camera)	About 50cm
Frequent Error elements	Eye blink, face/Background boundary
Error compensation	Brightness compensation, two stage detection
Image detection success rate	95%

### 5.2. Image and speech linked experiment

Determining the progress of the speech recognition according to the existence of an image signal, the interworking success rate

of 99.3% resulted under the experiment environment connecting an image and speech processor. It did not decrease the speech recognition function of existing almost. Moreover, without the image processing function, and in the case of stopping the image function, it was confirmed that the speech recognition function performed as normal.



Figure 4. Linked test environment

## 6. Conclusions

In this paper, a method and experimental result were presented for preventing the interruption of the speech recognition process caused by dynamic acoustic noise, by means of the lip movement image signal. The process from the image acquisition to the image signal extraction was examined. The proper operation mode of the speech recognition function was confirmed under linkage with the image processor. This result is expected to be actively used for the dynamic noise processing in the continuous speech recognition process.

## 7. References

- [1] G. Potaminanos, H.P. Graf, and E. Cosatto, "An Image Transform Approach for HMM Based Automatic Lipreading, Image Processing, 1988. ICIP 98, Proceeding, pp.173-177, Oct. 1998.
- [2] 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, pp.65-70, 7-9 Dec. 1998.
- [3] Shogo Nishida, "Speech Recognition Enhancement by Lip-Information", Media Laboratory, MIT Cambridge, MA 02139, pp.198-204, April 1986.
- [4] Rafael C. Gonzalez, Richard E. Woods, "Digital Image Processing, Second Edition", 2002. pp.567-642.
- [5] F. Leymarie and M.D. Levine, "Simulating the Grassfire Transform Using an Active Contour Model", Trans. IEEE Pattern Analysis and Machine Intelligence, 14(1):56-75, 1992.
- [6] Z.Q.Wu, J.A.Ware, W.R.Stewart, and J.Jiang, "The Removal of Blocking Effects Caused by Partially Overlapped Sub-Block Contrast Enhancement", Journal of Electronic Imaging -- July - September 2005 -- Volume 14, Issue 3, 033006(8 pages).