嘉南藥理科技大學專題研究計畫成果報告

聽障者與不能發音者的電腦輔助溝通系統

計畫類別：□個別型計畫　　　　　　□整合型計畫

計畫編號：CNMI92-04

執行期間：92 年 1 月 1 日至 92 年 12 月 31 日

計畫主持人：黃文楨

共同主持人：

計畫參與人員：

執行單位：資訊管理系

中華民國　　93　年　　1　　月　7　　日

A Visual Communication System for Hear-impaired and Talk-impaired Patients

Wen-Chen Huang (黃文楨)

Keyword: lip-reading, spatial-temporal image difference, pattern recognition, 3D reconstruction

Abstract

The most common communication disorders clinically are speech and hearing impairment. For the lengthening of life expectancy, hearing impairment becomes one of the most important issues in our society. After medical or surgical therapeutic processes, rehabilitation is the last resort to restore the capability of communication. A key component for such restoration depends on an alternative media through which the impaired function can be processed as well as possible. For speech and hearing disorders, communication procedure can usually be augmented by visual stimulation, for example, by signs or text. Owing to the development of world-wide-web system, not only our traditional oro-aural communication mode can be replaced by cyber-text and far-distance communication but also the communication mode of the speech or hearing impaired people can be. For the promising technique of data analysis, we can also build a visual lip-reading system for these patients, either for their post-operation communication or for their later speech rehabilitation period.

The purpose of this research is to build a visual lip-reading system which recognizes the sentence from the image sequences of speakers. At the same time, many face images are taken from different angles through digital cameras to build a vivid 3D human head model. The recognized text drives the 3D human head model to talk as the real person. There are three parts in our framework: lip-reading recognizer, 3D head model generator, and talking face animation. The integrated system combines the web-based visual communication interface. Experimental results show that the recognition rate is about 97 percent of ten sentences for a specific person.

1. Introduction

Cancer of the larynx is the most common head and neck cancer, representing approximately 2~5% of all cancers. More and more patients are vulnerable to such malignant disorder and succumb to loss their voice for the surgical treatment of laryngeal cancer. Alternative modes of communication after laryngectomy include pneumatic devices, electrolaryngeal devices, esophayngeal speech and tracheoesophageal fistulization [1]. All these speech rehabilitation have their own indications and limitations. For many years of evaluation, electrolaryngeal devices and trachoesophageal fistulization remain the most popular alternative communication methods [2]. In consideration of its “robot-like” tone and one-hand- operation, traceoesophageal fistulization seems to be more convenient for patients. But the following wound infection, scar stenosis and pulmonary hygiene are still manifesting these laryngectomee patients. For the promising development of data analysis, we can build a visual lip-reading system for these patients, either for their post-operation communication or for their later speech rehabilitation period. Such system can also restore patient’s voice data before the operation and reconstruct their voice by the lip-reading recognizing and visual analysis.

The development of visual communication tool between hear-impaired or talk-impaired patients becomes possible through the researches in computer graphics and computer vision. Many successful results can be found in different applications such as speech training, e-learning, tele-conference, and long-distance medical care, et al. The purpose of this research is to build a visual lip-reading system which recognizes the sentence from the image sequences of speakers. At the same time, many face images are taken from different angles through digital cameras to build a life-like 3D human head model. The recognized text drives the 3D human head model to talk as the real normal person without hear-impaired or talk-impaired symptom. With this technology, talk-impaired or hear-impaired patients will keep in touch no matter where they are through the internet. Figure 1.1 shows the framework of the proposed method. There are three parts in our framework: lip-reading recognizer, 3D head model generator, and talking face animation. The assumption of this system is sentence-based recognition on limited vocabulary for specific patient. Although more sentences could be added into our system to increase the vocabulary database, however, it’s the topic out of this research scope.

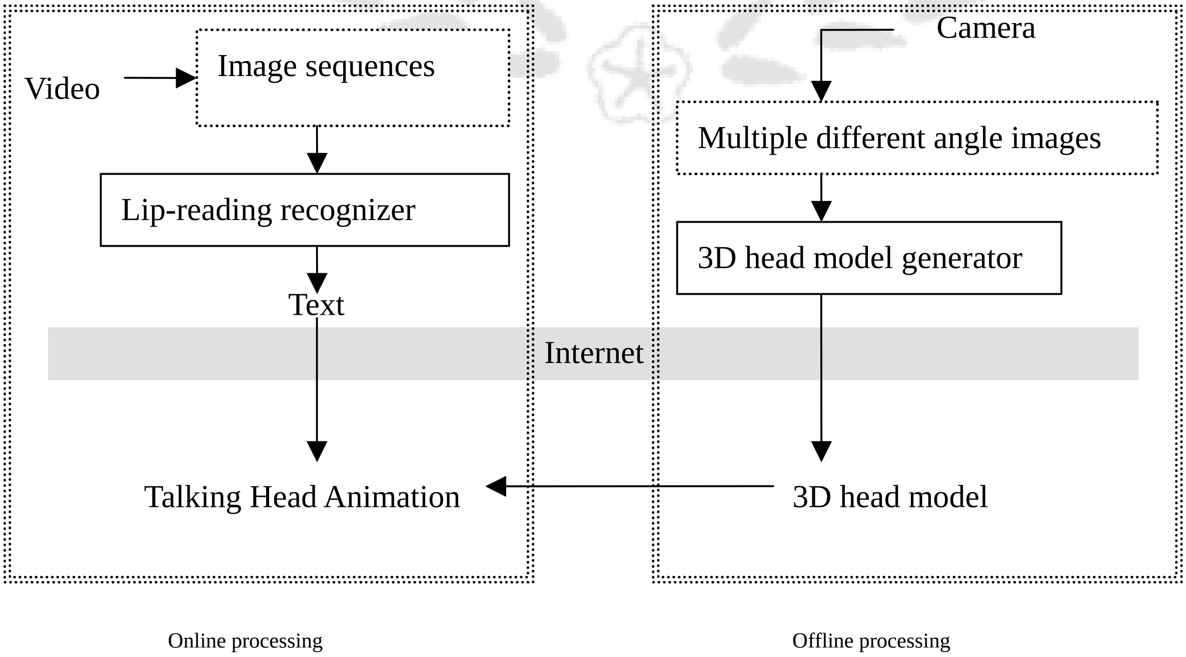

Figure 1.1: Framework Overview

For the lip-reading recognizer, a spatial-tempo image differences (STID) method is used to recognize limited sentences from the same speaker without voice. The image sequences of multiple sentences are the training and testing data. Each sequence of images of a sentence is a basic unit for the training or testing data. The sequences of images include the entire face region with complex background (not limited only on the lip region as most lip-reading systems nowadays) under normal illumination. The registration and crop process are executed through all the image sequences. For removing the complex background, the skin-tone color statistics is used to detect the face region. The lip region is located by using the Gaussian mixture model. After finding the lip region, the center of lip region is easily calculated by averaging the x-y coordinate of all the pixels upon lip region. For feature extraction, we propose to use the spatial and temporal features from image sequences. The temporal difference image makes the recognizing scheme to be more robust. The Euclidean Difference method is used for recognition.

The contribution of this research is to propose an integral visual communication system for those patients who cannot speak out with a voice or have the difficulties of hearing the sound. The proposed system combines the recognition and 3D animation functions with the remote communication capability through the internet environment. The capability of correcting the sentence spoken by the talking-impaired or hear-impaired patients with more precise sentence is also demonstrated. Unlike other avatar systems, a new web-based visual communication interface is proposed. The advantage of our recognition method is simple and robust. With this system, hear-impaired or talk-impaired patients would be able to have visual communication as they meet the real persons in remote position.

2. Lip-reading recognizer

The proposed system as shown in Figure 2.1 consists of six components: (1) Registration of image sequences, (2) Face region detection, (3)Lip region detection, (4) Spatial-temporal sampling of image sequences, (5) Training the resulted image sequences, and (6)Recognition .

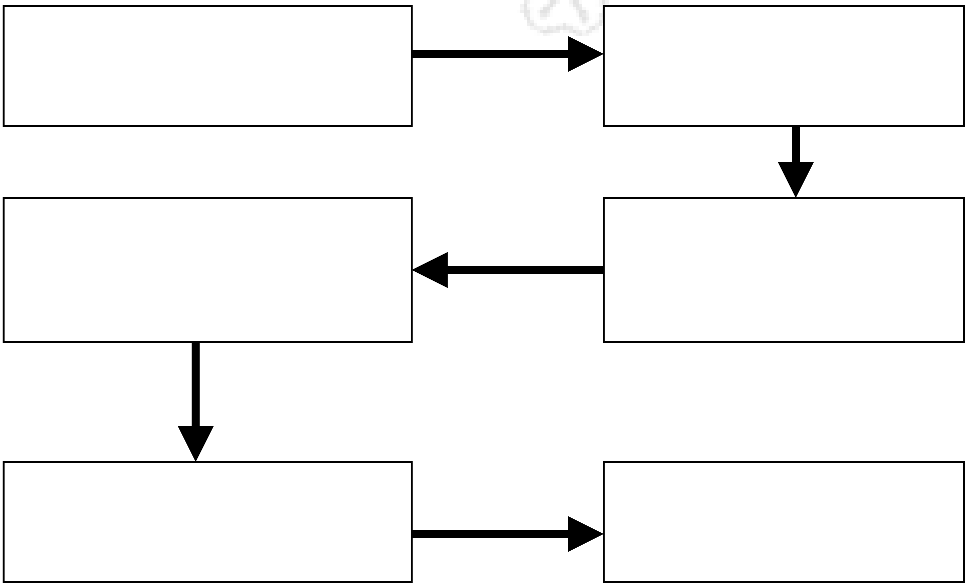

Figure 2.1: Information flow of the proposed lip-reading system

2.1 Spatial-temporal sampling of image sequences

The overall image sequences are denoted as I_{rpq}, where the sentence: r=1,..,M, training times: p=1,..,N, and the decomposed image sequences: q=1,..,K_p. In general, M and N are fixed, but K_p is different in each case (please see Table 1). The purpose of spatial-temporal sampling of image sequences is to adjust K_p into a fixed number K. For every r and p, there are image sequences I_{rp1}, I_{rp2} ,.., I_{rpt} where t is the number of image sequences, say K_p. The difference of image sequences

$$d_i = \| I_{rpi} - I_{rp(i+1)} \|, \quad i=1, .., t-1. \qquad (9)$$

Find the minimal index of d_i , say b. In other word,

$$b = \arg(\min_i (d_i)) \qquad (10)$$

where i=1, .., t-1. Let

$$I_{temp} = (I_{rpb} + I_{rp(b+1)})/2. \qquad (11)$$

Then

$$I_{rpb}=I_{temp} \text{ and } I_{rp(b+1)}=I_{temp}. \qquad (12)$$

The algorithm is described as below:

```
Input : $I_{rpq}$, r=1,...,M, p=1,…,n, q=1,..,$K_p$ ($K_p$ is varied)
Output :$I_{rpq}$, r=1,..,M, p=1,..,N, q=1,..,K (K is fixed, and K < $K_p$)
Step1: For every r and p, and those unmerged image sequences, compute $d_i$, i=1,…,t-1, t is initialized as $k_p$
Step2:   find the minimum of $d_i$
Step3:   $b = \arg\left(\min_i d_i\right), i = 1,\dots t-1$
Step4:   Merge $I_{rpb}$ and $I_{rp(b+1)}$
Step5:   Tag $I_{rpb}$ as merged, t=t-1
Step6:   if t<k then go to step1
```

The diagram of describing the above algorithm is shown in figure 2.2.

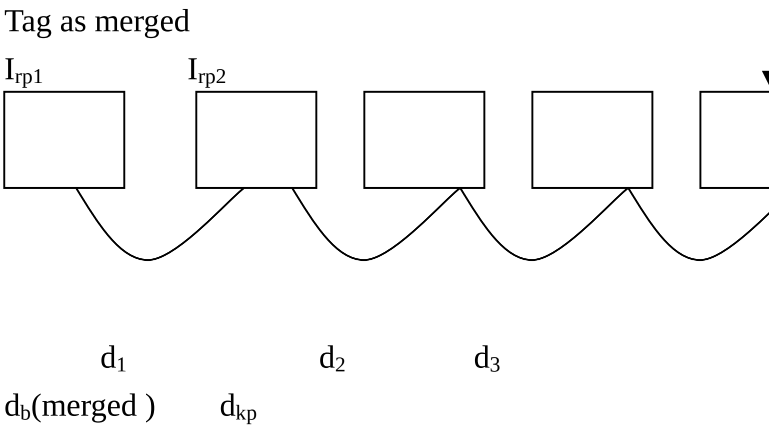

Figure 2.2: Spatial-temporal algorithms, where t =k_p

After Spatial-temporal Image difference algorithm on all the image sequences, the resultant number of image sequence is fixed, say K. For every sentence r, the image sequences are listed as follows:

1st trained image sequences:I_{r11},I_{r12},-----,i_{r1k}

2nd trained image sequences:I_{r21},I_{r22},----I_{r2k}

………

Nth trained image sequences :I_{rn1},I_{rn2},----I_{rnk}

Average trained image sequences: $\hat{I}_{r1}, \hat{I}_{r2}, \cdots, \hat{I}_{rk}$

Where

$$\hat{I}_{rc} = \left(\sum_{y=1}^{N} I_{ryc} \right) \Big/ N, c = 1,\dots k \qquad (13)$$

Now the trained image sequences for sentences are as follows:

The 1st trained sentence: $\hat{I}_{11}, \hat{I}_{12}, \cdots, \hat{I}_{1k}$

The 2nd trained sentence: $\hat{I}_{21}, \hat{I}_{22}, \cdots, \hat{I}_{2k}$

....

The Mth trained sentence: $\hat{I}_{M1}, \hat{I}_{M2}, \cdots, \hat{I}_{Mk}$

2.2 Recognition phase

Given a test image sequence T$_1$, T$_2$,….,T$_L$, we need to perform ***Spatial-Temporal Image difference (STID)*** algorithm before recognition phase. After processing of STID, the test image sequences become $\hat{T}_1, \hat{T}_2, ..., \hat{T}_k$. The Euclidean distance between the test image sequence and trained image sequences is used as a measurement of similarity. The Euclidean distance between the ith trained image sequences is defined as

$$E_i = \left\| \left(\hat{T}_1 - \hat{I}_{i1} \right) \right\| + \left\| \hat{T}_2 - \hat{I}_{i2} \right\| + ... + \left\| \hat{T}_k - \hat{I}_{ik} \right\|, i = 1, ..., M, \quad (14)$$

where M is the number of sentences.

Then $\arg\left(\underset{i}{Min}\left(E_i \right) \right)$ is the index of desired sentence to be recognized.

3. Experimental Results

For the lip-reading recognizer phase, the image sequences are taken from regular desktop CCD camera in the windows XP operating system. The image formats are RGB-based and are saved as a sequential of compressed JPEG images. The image sequences of multiple sentences are the training and testing data. There are ten sentences used as the trained sample data. Each sentence was spoken ten times by the same person. The test data are also recorded by speaking ten sentences and each sentence ten times. Table 3 shows the Euclidean distance between the first test image sequence and the trained image sequences as an example, where S$_{i\_j}$ means the tested image sequences speaking the ith sentence of the jth time, and Q$_p$ means the pth sentences of trained image sequence. In Table 5.1, the smallest Euclidean distance for each test image sequences is marked as bold. Note that there are nine smallest value located in the first row of the table, which means the recognizing ratio for the first sentence is 90%. In the column of S$_{1\_6}$, the smallest distance is located in the Q$_4$th row, which means this sentence is recognized as the 4th sentence.

Table 5.2 shows the confusion matrix of the test image sequences. From this table, the 1st sentence is recognized as the 1st sentence 9 times and is recognized as the 4th sentence 1 time. The sentence 3, 4, 7, 9 are recognized perfectly. The recognition ratio for sentence 1,2,5,8,10 is 90%, and for sentence 6 is only 60%. The overall recognition ratio is 91% (for k=10). Table 5.3 shows the recognition rate versus k. Note that the recognition rate is proportional to the k value. Figure 5.1 shows the recognition rate versus the fixed number of frames after spatial-tempo algorithm. Compared with the most popular lip-reading recognition method, HMM, which we performed the same experiment only with the recognition rate is about 25%. The reason maybe is that HMM cannot handle so many number of image sequences since most HMM recognize one articulation each time in most lip-reading systems.

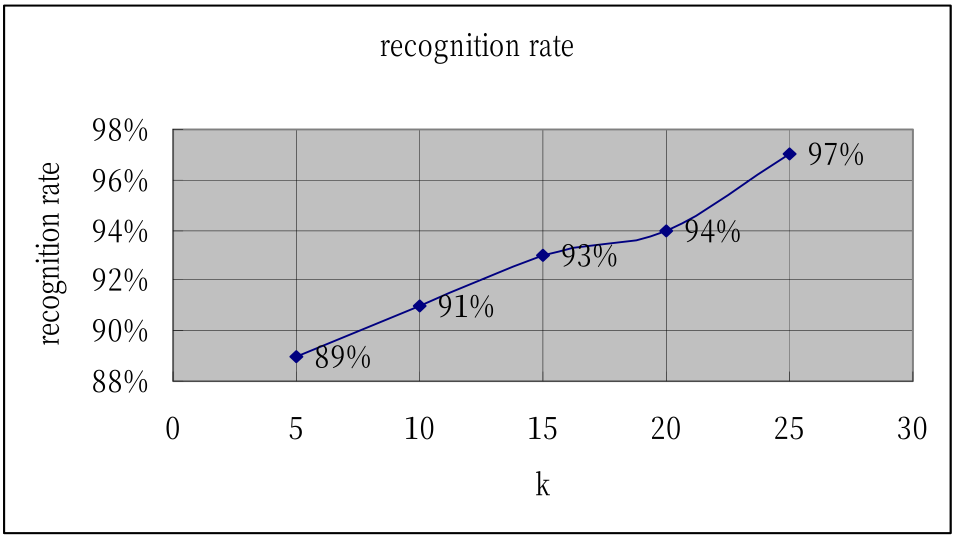

Figure 3.1: The recognition rate versus the fixed number of frames after spatial-tempo algorithm.

4. Conclusions and further research

After medical or surgical therapeutic processes, rehabilitation is the last resort to restore the capability of communication. A key component for such restoration depends on an alternative media through which the impaired function can be processed as well as possible. For speech and hearing disorders, communication procedure can usually be augmented by visual stimulation, for example, by signs or text. Owing to the development of world-wide-web system, not only our traditional oro-aural communication mode can be replaced by cyber-text and far-distance communication but also the communication mode of the speech or hearing impaired people can be. For the promising technique of data analysis, we can also build a visual lip-reading system for these patients, either for their post-operation communication or for their later speech rehabilitation period. Such system can also restore patient's voice data before the operation and reconstruct their voice by the lip-reading recognizing and visual analysis.

The purpose of this research is to build a visual communication system, which recognizes the sentence from the image sequences of hear-impaired and talk-impaired patient for communication. The proposed system uses Spatial-Temporal Image Difference technique to recognize limited sentence from the same speaker. Several research results are demonstrated as followings:

- Recognizes the limited basic sentences by the sequences of images taken from the same speaker.
- Provides a new user interface to the computer systems.
- Helps the hear-impaired patient for communication.

Eventually, the proposed system is capable of incorporating audio-features for improving speaking recognition system, especially in the noisy background environment. The talking education of the talk-impaired or hear-impaired patients by showing the movement of lip to correct their pronunciation is another potential application. On-line or Off-line increase the vocabulary database so that the system can handle a general case of visual communication is also interesting. The text-to-speech mechanism should be included in the near feature so that can help the talk-impaired or hear-impaired patients to communicate with normal person more naturally. The other potential application is to incorporate audio-features for improving speech recognition system, especially in the noisy-sound background environment.