Voice Coding Using Only the Features of the Face Image

*Youn-Soo Cho, *Jong Whan Jang

※ This paper was supported by 1997 Non Directed Research Fund, Korea Research Foundation.

Abstract

In this paper, we propose a new voice coding using only the features of the face image such as mouth height(H), width(W), rate(R=W/H), area(S), and ellipse's feature(P). It provides high security and is not affected by acoustic noise because we use only the features of face image for speech. In the proposed algorithm, the mean recognition rate for the vowels approximately rises between 70% and 96% after many tests.

I. Introduction

Recently, the demand for communication systems that permit speech input within various environments has grown; for example in the areas of man/machine interfaces and mobile communications. However, speech input has problems: 1) degradation caused by surrounding acoustic noise; 2) generation of acoustic noise to the environments; 3) low security because speech can be overhead; and 4) in case of hearing-impaired not to communicate. On the other hand, lip reading using computer vision to understand speech, is an interesting alternative because actual utterance is not necessary to input information. It has been reported that information from articulators can be obtained by lip reading. Thus, lip reading is capable of overcoming the problems of speech input.

Several lip reading methods have been studied for speech coding using images [1,4,7] and for auxiliary means of speech recognition [8]. Speech communication with these methods is limited in ability, however, because coding depends on a limited number of prepared reference patterns. Because of limited memory capacity and computation time, it is hard to increase the number of reference patterns to improve coding. Moreover, even a few coding errors seriously degrade word coding or sentence understanding when the system is based on phoneme coding.

In this paper, we propose a new voice coding method using the edge detection and matching algorithm based on the features of the face image [2,3,5,6]. Because an actual utterance is not necessary to input into this method, it provides high security and is not affected by acoustic noise. In addition, it shows great promise as a speaking-aid system for the people whose vocal cords are injured. We show the speech communication capability of the image coding for five vowels in this investigation.

In Section Ⅱ, we present the structure of a proposed system that allows speech synthesis from oral motion images. Next, in Section Ⅲ, we discuss the proposed system's performance in handling five vowels and show that the proposed system is capable of speech communication. Finally, in Section Ⅳ, we present our conclusions and areas for future research.

II. The structure of a proposed system

One of the methods to distinguish the speech is to divide it into a round speech or an unround and closed speech. For example in vowels, o and u represented by an open mouth shape belong to a round speech. Similarly, a, e, and i belong to a closed speech that represents a closed mouth. The important elements to make a distinction of vowels are tenseness, rounding, lengthening, nasalization and tone, etc. But in the visual voice coding research, rounding and lengthening are basic and visibly distinguishable units of vowel speech.

Figure 1 shows a block diagram of the proposed system. The system consists of five parts; face image capture, image edge detection, lip image selection, feature extraction and speech signal estimation.

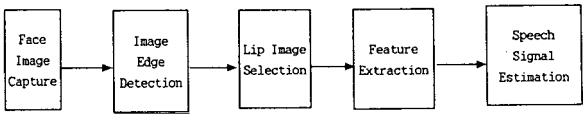

Figure 1. A block diagram of the proposed system.

* Division of Computer, Informations & Communications, and Electronics Engineering, PaiChai University.

Manuscript Received : May 20, 1999

2.1 Face image capture

Images are taken to be used as input signal with a CCD camera with 30 Hz sampling in sufficient lighting keeping the distance fixed from the camera. The resolution of the sampled image is 128×128 and the sampled images were quantized as four bits per pixel for the faster computation processing, so they were 16-level gray images.

2.2 Image edge detection

We use four 3×3 Sobel filters to obtain face edge image and find out lip image from it. Edge points can be thought of as pixel locations of abrupt gray-level change. Four 3×3 Sobel filters with directional characters are as follows.

$$sobel\ 0 = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

$$sobel\ 45 = \begin{bmatrix} 0 & -1 & -2 \\ 1 & 0 & -1 \\ 2 & 1 & 0 \end{bmatrix}$$

$$sobel\ 90 = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

$$sobel\ 135 = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & -2 \end{bmatrix}$$

2.3 Lip image selection

We use lip patterns to extract the lip edge image with 64×64 from the face edge image.

2.4 Feature Extraction

The features in the lip image such as mouth height(H), width(W), ratio(R=W/H), area(S), and ellipse's feature(P) are given in Figure 2. Mouth height is composed of the number of pixels. Other features are determined in the same way.

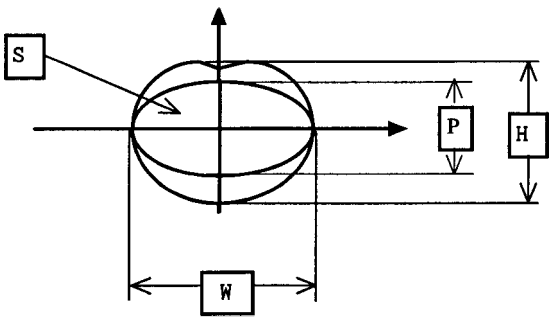

Figure 2. The features in the lip image. H, W, R, S, and P are height, width, ratio, area, and ellipse's feature respectively.

2.5 Speech signal estimation

In the proposed system, the vowels are estimated from the features obtained in Figure 2. To analyze the vowels, we first define Z_m by Eq. (1)

$$Z_m = [\sum_{n=1}^{5} f_{mn}^{2}]^T = [z_a\ z_e\ z_i\ z_o\ z_u]^T \tag{1}$$

where $m = 'a,\ e,\ i,\ o, u'$ and $n = 'W, H, S, R, P'$. Z_m represents values of the typical features of five vowels stored in DB from the many tests in the best environments. To extract one of vowels from the oral features, Y is defined as follows

$$Y = FX \tag{2}$$

$$\text{where } F = \begin{bmatrix} f_{aW} & f_{aH} & f_{aR} & f_{aS} & f_{aP} \\ f_{eW} & .. & .. & .. & f_{eP} \\ f_{iW} & .. & .. & .. & f_{iW} \\ f_{oW} & .. & .. & .. & f_{oW} \\ f_{uW} & .. & .. & .. & f_{uW} \end{bmatrix} \tag{3}$$

$$X = [x_W\ x_H\ x_R\ x_S\ x_P]^T \tag{4}$$

$$x_n = \acute{x}_n \cdot T_n \tag{5}$$

$$f_{mn} = \acute{f}_{mn} \cdot T_n \tag{6}$$

$(T_{W=0.1}, T_{H=0.1}, T_{R=0.5}, T_{R=0.2}, T_{P=0.1})$ $\acute{x}_n$ and $\acute{f}_{mn}$ are the feature values measured in the oral image. x_n and f_{mn} are the values that $\acute{x}_n$ and $\acute{f}_{mn}$ are multiplied by the weighting factor T_n respectively. F is a value of the features stored in dB. The weighting factor T_n is used to make a distinction of vowels easily. The factor ranges from 0.1 to 0.5 according to the significant correlations. The most important element to make a distinction of vowels is rounding. Therefore, T_R set the greatest value to 0.5. From Eq. (2), we get

$$Y = [\sum_n (f_{an}X_n) \sum_n (f_{en}X_n) \ldots. \sum_n (f_{un}X_n)]^T \tag{7}$$

To choose the actual speech of five vowels, $|Z_m - Y|$ is defined as follows

$$|Z_m - Y| = [|Z_a - \sum_n (f_{an}X_n)| \ldots. |Z_u - \sum_n (f_{un}X_n)|]^T \tag{8}$$

The speech signal is obtained by choosing the lowest value in the $|Z_m - Y|$.

III. The experiment results

We carried out simulations from the images with 128×128 pixel with 16 gray levels for five vowels. Figure 3 shows test image of a face. Using the Sobel function, the edge image is obtained in Figure 4. We use lip patterns to detect the lip image in the edge image. Lip edge image with 64×64 pixels obtained from the edge image using the lip edge selector is given in Figure 5.

Figure 3. Test image of a face with 128×128 pixels with 16 gray levels.

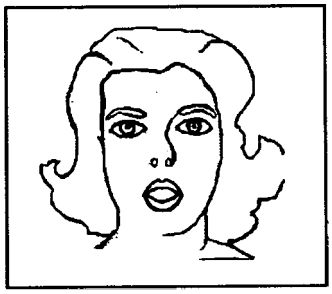

Figure 4. Edge image obtained from the test image.

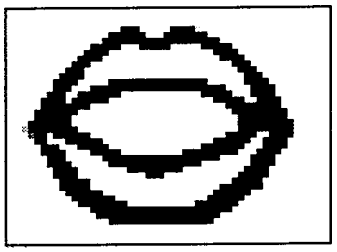

Figure 5. Lip edge image with 64×64 obtained from the edge image using the lip edge selector.

Using the proposed algorithm, we present good results. The coincidence rate of each vowel is as follows. a, e, i, o, and u are 97%, 93%, 88%, 84%, and 89% respectively. From the experiment with sample o, the outcome was obtained as follows; a, e, i, o, and u are 0, 2, 1, 38, 4 respectively. To increase the coincidence rate, we multiply the features by the weighting factor. We focused on the still image this time and it could be extended to the moving pictures easily. This method is expecially promising as a speaking-aid system for people whose vocal cords are injured. Since this is a basic investigation of media conversion from image to speech, we focus on lip shapes, and conduct experiments on media conversion of vowels. It provides high security and is not affected by acoustic noise because it is not necessary to input the actual utterance. These features are extracted from oral images in a learning data set. We conclude the proposed system has potential as a method of nonacoustic communication. To increase the coincidence rate, we may use others features not including this experiments.

References

1. Y. Fukuda and S. Hiki, "Characteristics of the mouth shape in the production of Japanese stroboscopic observation," *J. Acoustics Soc. Japan*, (E), Vol. 3, No. 2, pp. 75-91, 1982.
2. T. Hasegawa and K. Otani, "Oral image to voice converter-image input microphone," in *Proc. ICSS/ISITA*, '92, Vol. 20, No. 1, pp 617-620, 1992.
3. F. Lavagetto, "Speech-assisted motion compensation in videophone communications," Symp. on Multimedia Communication and Video Coding, New York, Oct. 1994.
4. K. Mase and A. Pentland, "Lipreading by optical-flow analysis," *Trans. IEICE Japan, Vil. J73-D-II*, No. 6, pp. 796-803, 1990.
5. K. Otani and T. Hasegawa, "Speech synthesis from oral motion image," In *Proc. NOLTA '93*, Vol. 4, pp. 1355-1358, 1993.
6. R. Rao and T. Chen, "Cross-modal predictive coding," Sym. on Multimedia Communication and Video Coding, New York, Oct. 1994.
7. K. Uchimura, J. Michida, M. Tokou, and T. Aida, "Discrimination of Japanese vowels by image analysis," *Trans. IEICE Japan*, Vol. J71-D, No. 12, pp. 2700-2702, 1988.
8. B. P. Yuhas, M. H. Golodsein, Jr., T. J. Sejinowski, and R. E. Jenkins, "Neural network models of sensory integration for improved vowel recognition," *Proc. IEEE*, Vol. 78, No. 10, Oct. 1990.

▲Youn-Soo Cho

Youn Soo Cho received the B.S degree in Physics from PaiChai University in 1995. She is currently working forward the M.S. degree in Dept. of Inf. & Com. Eng., PaiChai University. Her research interests include image processing, multimedia communications.

▲Jong-Whan Jang

Jong Whan Jang received the B.S.E.E degree in Electronic Communication Engineering from HanYang University, in 1979. He also received the M.S.E.E. and Ph.D. degrees in Electrical and Computer Engineering from North Carolina State University, Raleigh, NC, USA in 1986 and 1990 respectively. Since 1990, he has been with PaiChai University, where he is currently an associative professor of the Division of Computer, Informations & Communications, and Electronics Engineering. His research interests include image processing, multimedia communications, and multimedia database. He served the Director of Computer Center in PaiChai University from 1992 to 1996. He is Chief Information Officer at Taejon Metropolitan City and is also the Director of PaiChai Information and Communications Venture Business Center supported by MIC since 1998.