The Image Input Microphone—A New Nonacoustic Speech Communication System by Media Conversion from Oral Motion Images to Speech

Keiichi Otani and Takaaki Hasegawa, *Member, IEEE*

Abstract—**In this paper, we propose a new speech communication system to convert oral motion images into speech. We call this system "The Image Input Microphone." It provides high security and is not affected by acoustic noise because it is not necessary to input the actual utterance. This system is especially 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 vowels, and conduct experiments on media conversion of vowels. The vocal-tract transfer function and the source signal for driving this filter are estimated from features of the lips. These features are extracted from oral images in a learning data set, then speech is synthesized by this filter inputted with an appropriate driving signal. The performance of this system is evaluated by hearing tests of synthesized speech. The mean recognition rate for the test data set was 76.8%. We also investigate the effects of practice by iterative listening. The mean recognition rate rises from 69.4% to over 90% after four tests over four days. Consequently, we conclude the proposed system has potential as a method of nonacoustic communication.**

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 environs; and 3) low security because speech can be overheard. On the other hand, lip reading, using 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 [1]. Thus, lip reading is capable of overcoming the problems of speech input. Several lip reading methods have been studied for speech recognition using images [2], [3] and for auxiliary means of speech recognition [4], [5]. The purpose of this work has been speech recognition. Speech communication with these methods are limited in ability, however, because recognition 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 recognition. Moreover, even a few recognition errors seriously degrade word recognition or sentence understanding when the system is based on phoneme recognition.

Manuscript received September 16, 1993; revised June 21, 1994.

The authors are with the Department of Electrical and Electronic Engineering, Saitama University, Urawa-shi, Saitama 338, Japan.

IEEE Log Number 9406090.

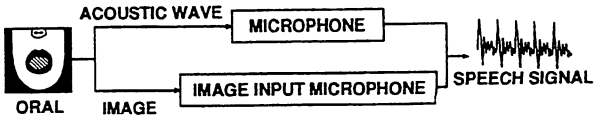


Fig. 1. The comparison between the conventional microphone and the image input microphone.

On the other hand, media conversion techniques that allow conversion between communication media such as text, image, and speech, have become more significant for human/machine interfacing and intelligent communication. For example, speech recognition and synthesis using some set of rules is media conversion between speech and text. This makes speech coding possible at a very low bit rate. A facial motion synthesis system, which converts speech or text to images for a more natural man/machine interface has been reported [6]. Such a method can be considered media conversion from image to text. Nevertheless, media conversion from image to speech which would help overcome the problems of speech input has received little attention.

In this paper, we propose a new speech communication system, "The Image Input Microphone" [7]–[9]. This system converts oral motion images to speech without recognition, so speech communication is not limited by the languages used [7]–[9]. Because an actual utterance is not necessary to input into this system, the system 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. In this paper, we show the speech communication capability of the image input microphone. We used five Japanese vowels in this basic investigation.

In Section II, we explain the principle of the image input microphone. Then in Section III, we present the structure of a proposed system that will allow speech synthesis from oral motion images. Next, in Section IV, we discuss the proposed system's performance in handling five Japanese vowels, and show that the proposed system is capable of speech communication. Finally, in Section V, we present our conclusions and areas for future research.

II. Principle of the Image Input Microphone

Fig. 1 compares the conventional microphone and the image input microphone. The conventional microphone converts acoustic waves to speech. However, the image input microphone converts images of oral motion to speech.

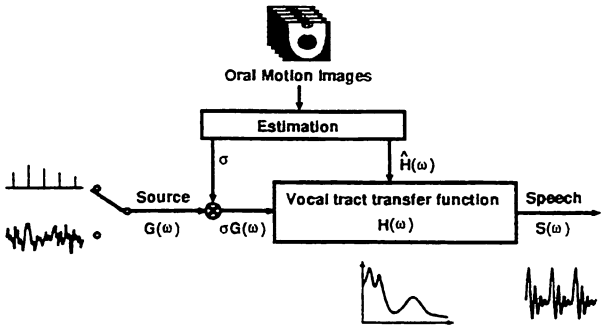


Fig. 2. The principle of the image input microphone system.

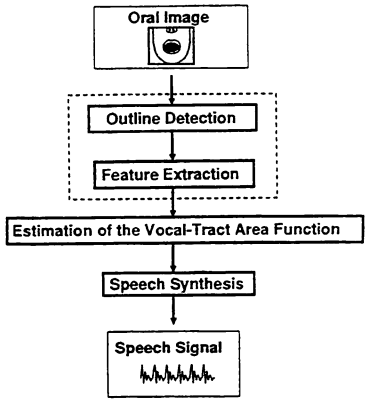


Fig. 3. A block diagram of the proposed system.

Fig. 2 shows the principle behind the image input microphone. The proposed system is based on a model of the speech production process. In speech processing, this model, which separates speech into the source $G(\omega)$ and articulation $H(\omega)$, is generally used. Speech waveform $S(\omega)$ is obtained from the equation

$$S(\omega) = G(\omega)H(\omega). \tag{1}$$

Usually, the source is approximated by a pulse train or a white noise, and articulation is approximated by an all-pole filter or a pole-zero filter. In the actual utterance, various forms of articulation are performed by changing the conditions of the vocal tract. That is, the vocal tract can be considered an articulation filter. If the transfer function of this filter is estimated from an oral image, we can synthesize speech by driving this filter. The condition of the vocal tract is controlled by movable articulators, i.e., the lips, tongue, velum, and so on. In particular, the movable and visible parts, i.e., the lips, tongue, and jaw, decide the condition of the vocal tract directly because they have such a wide range of movement. They also make the other less movable and invisible parts change dependently. It is thought that the transfer function of this articulation filter is estimated from features of the lips, tongue, and jaw, which are extracted from oral images. On the other hand, there are some utterances made with similar mouth shapes, for example /b/, /p/ and /m/. However, people have a superior information processing ability, so they should be able to recognize synthesized speech in context.

In this paper, we focus on vowels in a basic investigation of media conversion from oral motion images to speech. We used a vocal-tract area function as the transfer function of the articulation filter. The vocal-tract area function consists of a cross-sectional area of each section of the vocal-tract and can be obtained from PARCOR analysis of speech. The speech can also be obtained from PARCOR synthesis [10]. In the proposed system, each area of a vocal-tract area function is estimated from features of the lips, tongue, and jaw. We can then obtain the vocal tract filter from the estimated function. Finally, we can synthesize speech by driving the estimated filter.

III. System Structure

Fig. 3 shows a block diagram of the proposed system. The system consists of three parts: a features extractor, an estimator of the vocal-tract area function, and a speech synthesizer.

We use both oral motion images and uttered speech as inputs during the learning period. That is, during learning, both inputs are used in constructing the estimation system. After that, only the images are used for speech synthesis.

A. Features Extraction

Outline Detection Recently, outline detection methods based on an energy minimizing model have been studied. In these methods, the energy function is defined as the energy amounts to minimum value when the estimated outline becomes an appropriate one. In particular, "Snakes" [11] is used to try to track the motion of an outline. A lip outline detection method that uses knowledge of lips has also been reported [12]. We carried out the outline detection using these energy-minimizing models.

Estimated outline points $\mathbf{v}_i = (x_i, y_i)(i = 1, 2, \ldots, n)$ are on a closed spline curve connected by N discrete points $\mathbf{V}_j = (X_j, Y_j)(j = 1, 2, \ldots, N)$, where $n \gg N$. The outline energy function E_{out} is defined as follows

$$E_{\text{out}} = E_{\text{int}} + E_{\text{image}} + E_{\text{lips}} \tag{2}$$

where E_{int} is internal energy which represents the smoothness of the outline; E_{image} is image energy which represents the fit between the outline and the features in the image such as lines, edges, and so on; E_{lips} is knowledge of the lips energy which represents the sufficiency of the outline obtained by using knowledge of the lips. Each energy is defined as follows:

1) *Internal energy*: this energy determines the property of a closed spline curve.

$$E_{\text{int}} = \frac{1}{2}\sum_{j=1}^{N}\{\alpha|\mathbf{V}_j - \mathbf{V}_{j-1}|^2 + \beta|\mathbf{V}_{j-1} - 2\mathbf{V}_j + \mathbf{V}_{j+1}|^2\} \tag{3}$$

where $\mathbf{V}_0 = \mathbf{V}_N$. The coefficients of the first and second terms, α and β, can control the smoothness of the outline.

2) *Image energy*: this energy makes the outline change to adapt to features in the image. In particular, this energy causes the outline to be attracted to lines and edges.

E_{image} is defined as follows

$$E_{\text{image}} = w_{\text{line}}E_{\text{line}} + w_{\text{edge}}E_{\text{edge}} + w_{\text{dir}}E_{\text{dir}} \quad (4)$$

where w_{line}, w_{edge}, w_{dir} represent the weights of these energies; they are adjusted for convergence with an adequate outline.

The line energy E_{line} is defined as follows

$$E_{\text{line}} = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{v}_i) \quad (5)$$

where I represents the luminance level of the image. Depending on the sign of w_{line}, the outline is attracted to a line whose luminance level is high or low.

The edge energy E_{edge} is defined as follows

$$E_{\text{edge}} = -\frac{1}{n}\sum_{i=1}^{n} |\nabla I(\mathbf{v}_i)| \quad (6)$$

where ∇ represents the spatial differential.

If the outline converges to an accurate edge, the outline and the edge are mutually orthogonal. Then the edge direction energy E_{dir} is defined as follows

$$E_{\text{dir}} = \frac{1}{n}\sum_{i=1}^{n} |\cos\theta_i| \quad (7)$$

where θ_i represents the angle between the outline at $\mathbf{v}_i$ and the edge as shown in Fig. 4. This energy prevents the outline from converging with a wrong edge, for example a wrinkle or a furrow.

3) *Knowledge of lips energy*: E_{lips} is defined as follows

$$E_{\text{lips}} = w_{\text{sym}}E_{\text{sym}} + w_{\text{oral}}E_{\text{oral}} \quad (8)$$

where w_{sym} and w_{oral} represent the weights of E_{sym} and E_{oral}; they are adjusted for convergence with an the adequate outline.

The symmetrical energy E_{sym} represents the symmetry of the outline. Since the outline of the lips is considered symmetrical, this energy is used to detect the outline. E_{sym} is defined as follows

$$E_{\text{sym}} = \frac{1}{w}\sum_{x=0}^{\frac{w}{2}} \left[\{f_u(x) - f_u(-x)\}^2 - \{f_l(x) - f_l(-x)\}^2\right] \quad (9)$$

where w represents the width of the outline of the lips, and $f_u(x)$ and $f_l(x)$ respectively represent the outlines of the upper and lower lips. (see Fig. 5).

It is thought that the luminance level of the oral region, except for the tongue and teeth, is lower than that of the region of the skin. For this reason, we performed binarization with an appropriate threshold to detect the oral region. Oral region energy E_{oral} is defined to include this region as follows

$$E_{\text{oral}} = -\sum_{x,y\in O} g_0(x,y) \quad (10)$$

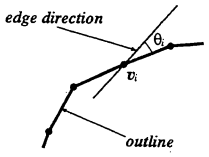


Fig. 4. Edge direction energy.

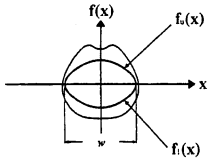


Fig. 5. Symmetrical energy.

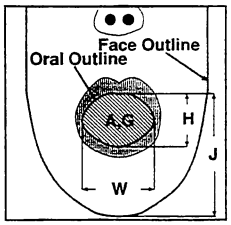


Fig. 6. The oral features.

where O represents the region within the outline, and g_0 is as follows:

$$g_0(x,y) = \begin{cases} 1 & I(x,y) \le T_{\text{oral}} \\ 0 & I(x,y) > T_{\text{oral}} \end{cases} \quad (11)$$

where T_{oral} represents a threshold obtained by DTSM [13].

Using these energy functions, oral and face outlines are determined from oral motion images. The outline points $\mathbf{V}_j$ are initialized by manual operation in the first frame. To minimize the energy function makes each $\mathbf{V}_j$ move to the point from among eight neighbors that requires the minimum energy until it converges with the energy function. The converged outline is obtained by spline interpolation of $\mathbf{V}_j$. The converged outline is then used as the initial outline in the next frame. Values for $\mathbf{V}_j$ are obtained to divide the outline into N parts equally. In this system, N (we used the number of N from three to six in the experiments that we will discuss later) depends on the size of the outline.

Oral Features After outline detection, the oral features were extracted from the outlines as shown in Fig. 6. The oral outline's area A, height H, width W, aspect ratio $\lambda = W/H$, jaw opening J, and the mean luminance level of its pixels G, were used as the oral features.

B. Estimation of the Vocal-Tract Area Function

In the proposed system, the vocal-tract area function, which is equivalent to the vocal-tract transfer function, is estimated from the oral features.

PARCOR analysis, which is a method of analyzing and synthesizing of speech, is often used for speech coding [14]. With a transmitter, some PARCOR coefficients and a gain

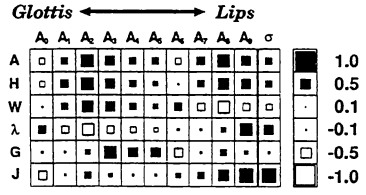


Fig. 7. Correlation coefficients between the areas of each section of a vocal-tract area function obtained from speech (columns) and the oral features extracted from oral motion images (rows). The size of each box represents the value of the correlation coefficient, and black and white represent positive and negative correlation, respectively.

parameter obtained from speech are sent to a receiver. At the receiver, a synthesis filter and a driving signal respectively are obtained from the received PARCOR coefficients and gain parameter. After that, speech is synthesized by driving the filter with the signal. On the other hand, it is well known that we can compute a vocal-tract area function from PARCOR coefficients [10]. The area of each section is computed from PARCOR coefficients recursively. Similarly, we can compute PARCOR coefficients from the area of each section.

To estimate the vocal-tract area function, we investigated the relation between vocal-tract area functions obtained from speech and the oral features extracted from oral motion images. We set the number of sections of the vocal-tract area function at ten. Then A_0 represents the area of the glottis and A_9 represents the area of the lips. Fig. 7 shows an example of correlation coefficients between the area of each section of such a function and the oral features. In this figure, absolute values of correlation coefficients range from about 0.1 to 0.8. We can see that there are significant correlations between the area of each section of this function and these features.

Therefore, we can estimate the area of each section and the gain parameter by multiple regression equations of the oral features. These equations were computed from the learning data set.

$$\mathbf{A} = \mathbf{RF} \tag{12}$$

where

$$\mathbf{A} = [A_0 \quad A_1 \quad \cdots \quad A_9 \quad \sigma]^T \tag{13}$$

$$\mathbf{R} = \begin{bmatrix} r_0 & r_{A0} & r_{H0} & r_{W0} & r_{\lambda 0} & r_{J0} & r_{G0} \\ r_1 & r_{A1} & r_{H1} & \cdots & \cdots & \cdots & r_{G1} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ r_9 & r_{A9} & r_{H9} & \cdots & \cdots & \cdots & r_{G9} \end{bmatrix} \tag{14}$$

$$\mathbf{F} = [1 \quad A \quad H \quad W \quad \lambda \quad J \quad G]^T. \tag{15}$$

Since the area of each section must be greater than zero, we prepared each appropriate threshold, then each threshold was set to the minimum value of each area.

C. Speech Synthesis

The proposed system synthesizes speech by the PARCOR vocoder [14]. We can compute PARCOR coefficients from the area of each section

$$k_m = \frac{A_{m-1} - A_m}{A_{m-1} + A_m} \tag{16}$$

where A_m represents the area of the mth section and k_m represents the PARCOR coefficient of the mth order. Speech is synthesized by driving the synthesis filter, which is obtained from the PARCOR coefficients, using the driving signal. For the PARCOR vocoder, we use the pulse train as the driving signal of voiced speech [14]. We set the pitch period of the pulse train at the mean value previously calculated for the speaker's pitch periods. Thus the pulse train $e(n)$ is obtained as follows [14]

$$e(n) = \begin{cases} \sigma(I/N)^{1/2} & n = 0, I, 2I, \ldots \\ -\frac{\sigma(I/N)^{1/2}}{I-1} & n \neq 0, I, 2I, \ldots \end{cases} \tag{17}$$

where σ represents the gain parameter, I represents the pitch period of the pulse train, N represents the length of the analysis window. We performed De-emphasis on the output signal of the synthesis filter.

IV. Performance Evaluation

We carried out simulations of speech synthesis from oral motion images of five Japanese vowels. We took oral motion images, to be used as input signals with a CCD camera with 30 Hz sampling in sufficient lighting keeping the distance fixed at about 20 cm from the camera. The resolution of the sampled images were 64 × 60, and the sampled images were quantized as eight bits per pixel, so they were 256-level gray images. At the same time, we took the uttered speech samples by 10 kHz sampling and they were quantized as eight bits. We performed pre-emphasis on the speech to get rid of effects on the glottal wave and on the radiation impedance. Each sample was uttered by a male Japanese speaker clearly and slowly. The length of each pattern was 3 s and the data set consisted of eleven different patterns. We took two kinds of data sets, i.e., a learning set and a test set. We computed the multiple regression equations from the learning set. The test set was used for speech synthesis. The frame length of speech was 33.3 ms due to the sampling rate of oral motion images. We updated the speech parameters, PARCOR coefficients, and gain parameter every pitch period by linear interpolation. The pitch period was 7.9 ms, which was the mean value of that of the speaker.

A. Evaluation of Spectral Envelopes

We show some examples of vocal-tract area functions in Fig. 8. These examples show that the estimated functions are in general agreement with the computed ones. In particular, /a/ and /o/ show reasonable agreement between the estimated functions and the computed ones. Estimation of the vocal-tract area function by multiple regression equations of features of the lips is clearly a useful method.

In addition, we show examples of spectral envelopes in Fig. 9. The synthesized speech is in general agreement with the uttered speech in terms of formant frequencies. It is well-known that formant frequencies are significant in the perception of vowels. We show an example of a sound spectrogram /aeiou/ in Fig. 10. The upper illustration represents uttered speech and the lower represents synthesized speech. We can see that the upper and the lower have similar formant frequencies.

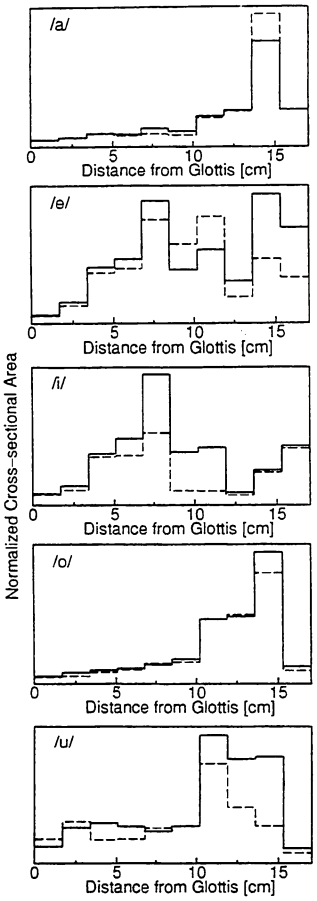


Fig. 8. Vocal-tract area functions of five Japanese vowels. Solid lines represent the estimated vocal-tract area functions and broken lines represent area functions computed form uttered speech.

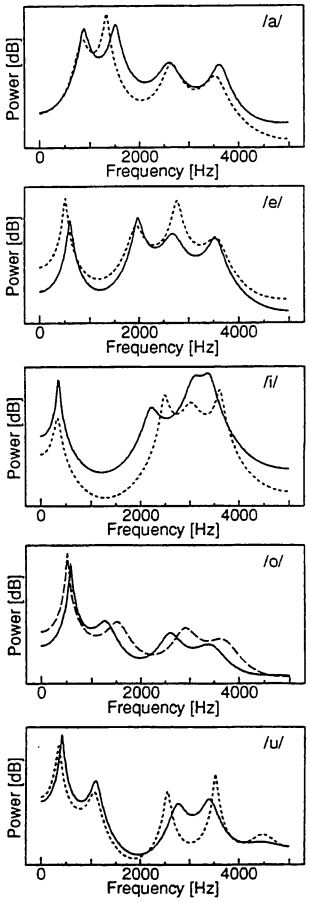


Fig. 9. Spectral envelopes of five Japanese vowels. Solid lines represent synthesized spectral envelopes and broken lines represent uttered ones.

Fig. 11 shows an example of the estimated gain parameter of the uttered /aeiou/. The tendency of the gain parameter of synthesized speech is similar to that of uttered speech. For an objective evaluation of this system, we investigated the mean LPC cepstral distortion. It was 4.7 dB and its standard deviation was 0.71. Therefore, it is likely that the proposed system can synthesize audible speech with respect to vowels.

B. Communications Possibility

Our main interest in this paper is to determine the ability of the proposed system to communicate by speech. For an subjective evaluation of this system, we carried out hearing tests to evaluate the performance of its speech communication. We presented random sequences of eight patterns to ten subjects, and investigated the recognition rate. Each pattern consisted of five Japanese vowels for a total of forty synthesized vowels per subject. The subjects were not familiar with synthesized speech.

Table I shows the recognition rates of the five Japanese vowels. The mean recognition rate of the test set was 76.8%. Thus, the speech synthesized from oral motion images of five Japanese vowels was audible. Hence the proposed system can communicate by speech with respect to vowels.

To further investigate recognition tendencies, we created a confusion matrix of the experimental results as shown in Table

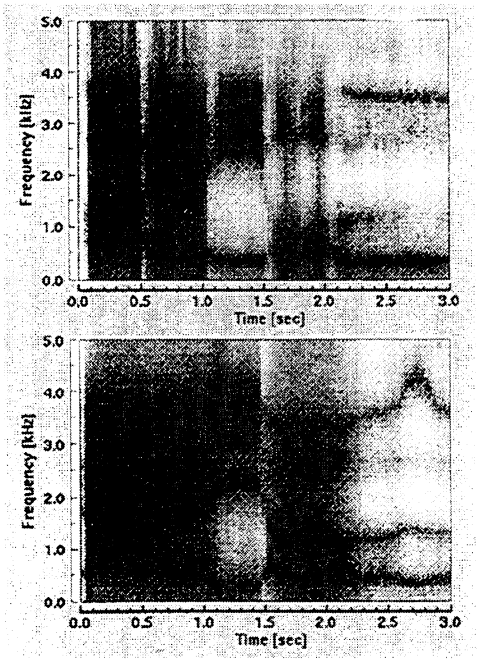


Fig. 10. An example of a sound spectrogram /aeiou/. The upper illustration represents uttered speech and the lower represents synthesized speech.

II. This table shows that /o/ and /e/ tend to be occasionally confused with /a/. We explain these tendencies as follows: although we do not show it in Table II, 68% of the confusion of

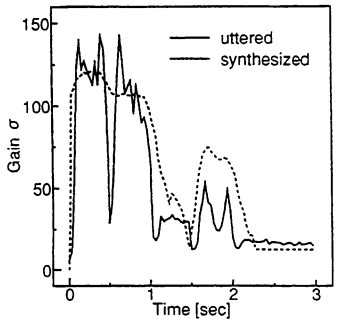


Fig. 11. An example of the estimated gain parameter of the uttered /aeiou/.

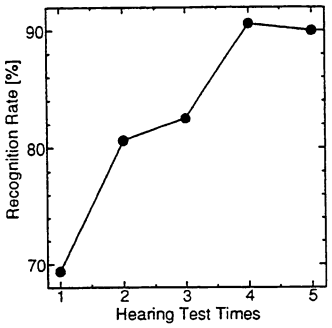


Fig. 12. The effects of practice by iterative listening.

TABLE I
RECOGNITION RATES OF FIVE JAPANESE VOWELS

	/a/	/e/	/i/	/o/	/u/
Recognition Rate[%]	98	68	76	59	84

TABLE II
CONFUSION MATRIX OF FIVE JAPANESE VOWELS

Input Image \ Recognition	/a/	/e/	/i/	/o/	/u/
/a/	78	2	0	0	0
/e/	12	54	2	6	6
/i/	1	8	61	1	9
/o/	19	6	1	47	7
/u/	0	0	2	11	67

/o/ with /a/ is due to /o/ being included in two specific patterns. We found that the spectral envelopes of the confused /o/ were similar to those of /a/ with respect to formant frequencies. We think that the visible vocal-tract area is large when /a/ is uttered so synthesized speech can be estimated accurately. However /o/, which has a small visible area, is not precisely estimated because of insufficient resolution. In our experiments, the tendency was to estimate /o/ as /a/ in this case. On the other hand, when /e/ was recognized as /a/, we found that the recognition error was most common in three specific subjects. Their errors accounted for 92% of the errors recognizing /e/ as /a/. Because the subjects were not familiar with synthesized speech, we think that the subjects who recognized /e/ as /a/ at the first hearing tend to recognize /e/ as /a/. While the subjects who recognized correctly at the first hearing tend to recognize correctly. Consequently, it is most probable that the proposed system will achieve nonacoustic communication if there is practice before use of the proposed system.

C. The Effects of Practice

In Section IV-B, we discussed recognition by people who are not familiar with synthesized speech. However, if the proposed system is applied as a speaking-aid system, we must investigate the effects of practice by iterative listening. We carried out the hearing test described in Section IV-B on four subjects each day for five days. We varied the sequence of patterns presented every day. Fig. 12 shows the effects of practice by iterative listening. Needless to say, we did not give the correct answers to the subjects after each experiment. At the first hearing, the mean recognition rate was 69.4%, however, it steadily increased as the tests were repeated. As a result, it was over 90% after four days, and it seemed to level off at that point.

Thus, prior practice with the synthesized speech increases the recognition rate. Consequently, the proposed system will be able to communicate by speech more effectively with people familiar with the system. So this system can be applied to speaking-aid communication systems.

V. CONCLUSIONS

We proposed a new speech communication system that converts oral motion images into speech. We call this system "The Image Input Microphone." The system provides high security and is not affected by acoustic noise because it is not necessary to input the actual utterance.

In the proposed system, the vocal-tract area function, which is equivalent to the transfer function of the vocal tract, is estimated from the features of oral images. The gain of source signal is also estimated. The synthesis filter is obtained from the estimated vocal-tract area function, and speech is synthesized by driving this filter with the driving signal whose gain is estimated. Since we found that there are correlations between the oral features and the areas of each section of the vocal-tract area function, the vocal-tract area function is estimated from multiple regression equations of these features. Similarly, the gain parameter was estimated in the same way.

We created the proposed system with five Japanese vowels on a computer, and carried out the following experiments. First, speech was synthesized using this system, and we investigated the spectral envelopes of the synthesized speech, which were in good agreement with the spectral envelopes of uttered speech in terms of formant frequencies. Based on this result, it seemed likely that audible speech could be synthesized by the proposed system with respect to the vowels. Next, we carried out hearing tests to investigate the system's communication ability. The mean recognition rate was 76.8% for the five Japanese vowels. We also investigated the effects of practice by iterative listening. When the test subjects heard the samples for the first time the mean recognition rate was 69.4%. After four repetitions over four days, however, it had risen to over 90%, at which it seemed to level off. Therefore, the proposed system seems capable of nonacoustic communication. In particular, it was shown that the system will be able to communicate with people who have had prior

practice with the synthesized speech. The proposed system seems applicable to a speaking-aid system.

Our future research will focus on applying the proposed system to speech which includes consonants, investigating how the proposed system applies to other speakers, and estimating the vocal-tract transfer function in the spectral domain.

ACKNOWLEDGMENT

We thank Professor M. Haneishi of Saitama University for his helpful guidance.

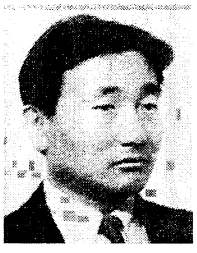
REFERENCES

[1] Y. Fukuda and S. Hiki, "Characteristics of the mouth shape in the production of Japanese—Stroboscopic observation," *J. Acoust. Soc. Japan*, (E) vol. 3, no. 2, pp. 75–91, 1982.
[2] K. Mase and A. Pentland, "Lipreading by optical-flow analysis," *Trans. IEICE Japan*, vol. J73-D-II, no. 6, pp. 796–803, 1990.
[3] 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.
[4] J.-T. Wu *et al.*, "Neural network vowel-recognition jointly using voice features and mouth shape image," *Trans. IEICE Japan*, vol. J73-D-II, no. 8, pp. 1309–1314, 1990.
[5] B. P. Yuhas, M. H. Goldstein, Jr., T. J. Sejnowski, and R. E. Jenkins, "Neural network models of sensory integration for improved vowel recognition," *Proc. IEEE*, vol. 78, no. 10, Oct. 1990.
[6] S. Morishima and H. Harashima, "A media conversion from speech to facial image for intelligent man-machine interface," *IEEE J. Select. Areas Commun.*, vol. 9, no. 4, May 1991.
[7] T. Hasegawa and K. Otani, "Oral image to voice converter—Image input microphone," in *Proc. ICSS/ISITA '92*, vol. 20.1, pp. 617–620.
[8] K. Otani and T. Hasegawa, "On the image input microphone," Tech. Rep. HC92-63, IEICE, 1993.
[9] K. Otani and T. Hasegawa, "Speech synthesis from oral motion images," in *Proc. NOLTA '93*, vol. 4, 1993, pp. 1355–1358.
[10] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.* vol. AU-21, pp. 417–427, Oct. 1973.
[11] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vision*, pp. 321–331, 1988.
[12] K. Mitsumoto *et al.*, "Lip contour extraction, complement, and tracing by using energy function and optical flow," in *Trans. of IPSJ*, vol. 31, no. 3, Mar. 1990, pp. 444–453.
[13] N. Otsu, "An automatic threshold selection method based on discriminant and least square criteria," *Trans. IEICE Japan,* vol. J63-D, no. 4, pp. 349–356, Apr. 1980.
[14] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech.* New York: Springer-Verlag, 1976.

Keiichi Otani was born in Sendai, Japan, in 1969. He received the B.E. and M.E. degrees in electrical engineering from Saitama University in 1992 and 1994, respectively.

He joined Fujitsu Limited in 1994 and has been engaged in the development of communication systems. His research interests are in multimedia communications and human communications.

Mr. Otani is a member of the Institute of Electronics, Information and Communication Engineers of Japan.



Takaaki Hasegawa (S'82–M'86) was born in Kamakura, Japan, in 1957. He received the B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University in 1981, 1983, and 1986, respectively.

He joined the Faculty of Engineering, Saitama University, in 1986 and has been an Associate Professor since 1991. His research interests are in human communications, human machine communications, and spread spectrum communications.

Dr. Hasegawa is a member of IEICE (The Institute of Electronics Information and Communication Engineers, Japan) and SITA (The Society of Information Theory and Its Applications, Japan).