Voice Recognition Method with Mouth Movement Videos Based on Forward and Backward Optical Flow

Kohei Arai [1]
Graduate School of Science and Engineering
Saga University
Saga City, Japan

***Abstract*—Lip reading method with mouth movement videos based on backward optical flow is proposed. Through experiments with 10 of mouth movement videos, it is found that the proposed lip reading method is superior to the conventional optical flow based method.**

Keywords- lip reading; optical flow; Hiddeen Markov Model; mouth movement

I. INTRODUCTION

Although voice recognition is now world widely available, recognition performance is not good enough for normal conversations. For instance, voice recognition performance of the typical Hidden Markov Model: HMM based method [1] (this is referred to the conventional voice recognition hereafter) with the feature of Formant is less than 50 % when the signal to noise ratio is below 5dB. In other words, voice recognition performance is totally affected by noise. In normal conversation among us, not only voice but also mouth movement is used for recognitions. Mouth movement video analysis makes voice recognition much better performance. The proposed lip reading method is for improvement of voice recognition performance.

Usually, Hidden Markov Model based method or neural network based method is used for voice recognitions as well as optical flow [2]-[9] based analysis of the mouth movement videos. Forward direction (from the past to the future) of optical flow is usually used for mouth movement analysis. Voice recognition performance can be improved by adding backward direction (from the future to the past) of optical flow for correction of voice recognition errors through a confirmation of recognized results. In this process, two voice elements are treated as a unit for the proposed backward optical flow. The conventional forward direction of optical flow recognizes by voice element by voice element, though. In order to make sure the recognized results, two voice elements are much easier and efficient manner. This is because transient between voice element and voice element is so important for voice recognitions. This is the basic idea of the proposed lip reading method.

Experiments are conducted with 10 of mouth movement videos which are acquired by different peoples. Voice recognition performance, then is evaluated and is compared to the conventional forward direction of optical flow based method. The experimental results show that the proposed backward optical flow is superior to the conventional method.

The following section describes the proposed method followed by some experiments. Then conclusion is described together with some discussions.

II. PROPOSED METHOD

A. Overview of the Proposed Voice Recognitions

Process flow of the proposed voice recognition method is shown in Figure 1.

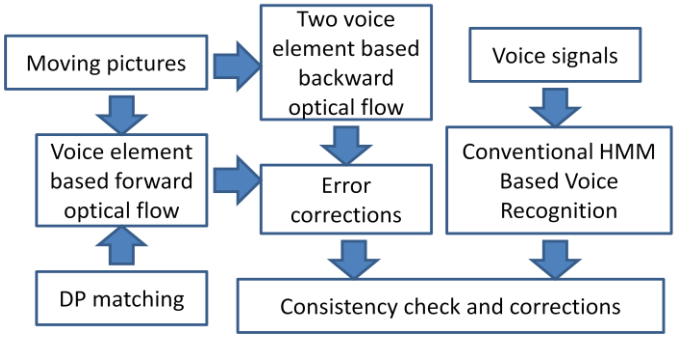

Fig. 1. Process Flow Of The Proposed Voice Recognition

Time series of moving pictures and voice signals are acquired first. Using the conventional HMM based voice recognition method, time series of voices are recognized. This is referred to voice based recognition, hereafter.

On the other hands, lip reading is performed based on forward optical flow with time series of moving pictures of mouth movement which are acquired at the same time of voice signals. This is done by voice element by voice element as usual. Meanwhile, two voice element based backward optical flow is applied to the time series of moving pictures of mouth movement. Then the result from the voice element based forward optical flow is corrected by using the two element based backward optical flow results. Through this voice element based optical flows, Dynamic Programming: DP matching based recognition is performed. Because extracted voice elements have missing portion of elements. Furthermore, recognition needs some insertions of voice elements. DP matching allows insertion and also recognition without some

missing elements. This is referred to moving picture based recognitions, hereafter.

After all, the recognized results from moving picture based and voice signal based methods are compare and check a consistency between both results, then final recognition results are reduced.

B. Optical Flow

Optical flow is defined as object movement representations in vector form in the visual representations. From moving pictures, videos of digital images, optical flow can be extracted as vectors. There are the conventional block matching method and gradient method for extraction of optical flow. Block matching method is usually referred to "Block-based methods" which are minimizing sum of squared differences or sum of absolute differences, or maximizing normalized cross-correlation while the gradient method is used to be referred to "Differential methods" which are based on partial derivatives of the image signal and/or the sought flow field and higher-order partial derivatives. Other than these, there are "Phase correlation methods" which can get inversion of normalized cross power spectrum between two adjacent images and "Discrete optimization methods" of which the search space is quantized, and then image matching is addressed through label assignment at every pixel.

C. Input Data for Dynamic Programming: DP Matching

Figure 2 shows an example of one cut of the moving picture of mouth movements. Time series of images are acquired. Voice element can be extracted from the time series of images. From the piece of the time series of images, four feature points, top, bottom, right end, and left end are extracted as input data for DP matching.

Fig. 2. Example of a piece of moving picture of time series of images of mouth movements

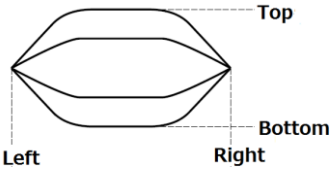

Fig. 3. Feature points as input data for DP matching

D. Fundamentals of Dynamic Programming: DP Matching

Similarity $D(A,B)$ between coded edge $[A]$ and $[B]$ is defined as follows,

$$D(A,B)=\min_{c[k]}\frac{\sum_{k=0}^{K-1}w[k]d[k]}{\sum_{k=0}^{K-1}w[k]} \quad (1)$$

where

$$A=\{a[i](i=0,\ldots,I-1)\}$$
$$B=\{b[j](j=0,1,\ldots,J-1)\} \quad (2)$$

and d[k] denote distance, as well as $w[k]$ denotes weighting coefficient,

$$w[k]=i_k-i_{k-1}+j_k-j_{k-1} \quad (3)$$

when the coded edge (K denotes the total number of edges) is represented as shown in Figure.4.

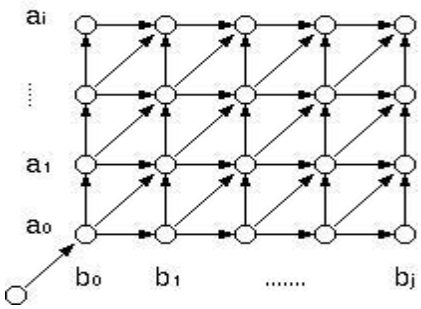

Fig. 4. Coded edge information

Subset summation of $s[c[m]]$ of numerator of equation (1) is expressed with equation (4) when k=m,

$$\begin{aligned}s[c[m]]=s[i_m,j_m]&=\min_{c[m]}\sum_{k=0}^{m}w[k]d[k]\\&=\min(\min_{c[m-1]}\sum_{k=0}^{m-1}w[k]d[k]+w[m]d[m])\\&=\min(s[c[m-1]]+w[m]d[m])\end{aligned} \quad (4)$$

If it is assumed that s is increased, then $s[c[m]]$ at k=m-1 is represented with equation (5),

$$s(i_m,j_n)=\min\begin{cases}s(i_m,j_{n-1})+d(i_m,j_n)\\s(i_{m-1},j_{n-1})+2d(i_m,j_n)\\s(i_{m-1},j_n)+d(i_m,j_n)\end{cases} \quad (5)$$

Because a, b positions are at the one of (i_{m-1},j_m), (i_{m-1},j_{m-1}), (i_m,j_{m-1}) in Fig.1. Thus total summation of s and D can be calculated if the summation of $s[c[m]]$ is reached at (i_{K-1},j_{K-1}).

Even if some of the coded edges are missing, similarity between two coded edges can be calculated results in edge image matching between the query image and the current image.

E. Detailes of Dynamic Programming: DP Matching

Initial condition is assumed to be $(x_0, x_0^{(l)})$, then $g_1^{(l)}(1,1)$ is minimum distance for $x_1 - x_1^{(l)}$ where x_i is input pattern data of voice elements while $x_i^{(l)}$ is reference voice elements. Then suffix of the input pattern data is incremented as follows,

$$g_j^{(l)}(i',i) = Min \begin{cases} g_{j-1}^{(l)}(i'-1,i) + w_j d_j(x_{i'}, x_i^{(l)}) \\ g_{j-1}^{(l)}(i'-1,i-1) + w_j d_j(x_{i'}, x_i^{(l)}) \\ g_{j-1}^{(l)}(i',i-1) + w_j d_j(x_{i'}, x_i^{(l)}) \end{cases}$$

There are three possible solutions which minimize the distance between input pattern data and the reference pattern data.

Meanwhile, $(x_0, x_0^{(l)})$ is defined as inner product (dot product) of the

$$(x, x^{(l)}) = ||x|| \cdot ||x^{(l)}|| \cos\theta$$

Then distance between two x_i and $x_i^{(l)}$ are as follows,

$$d_s = \cos\theta = \frac{(x, x^{(l)})}{||x|| \cdot ||x^{(l)}||}$$

Where $-1 \le \cos\theta \le 1$

To find the minimum distance, if the d_s is minimum when the $l=l_0$, then the input pattern data is classified to l_0. If a distortion is considered for the input pattern data due to some reasons, then d_s is no longer can be calculated with $(x_0, x_0^{(l)})$. The reason for that is some of the voice elements will be missing, or some of voice element inserted accidentally as shown in Figure 5. Therefore, distorted input pattern data (Modified pattern) has to be represented as follows,

$$x = (x_1, x_2, L, x_{i'}, L, x_{I'})^T$$

Reference patter in Figure 5 is defined as reference patter for voice elements. In this case, the following function which represents the relation between d_s and $(x_0, x_0^{(l)})$.

$$F = c(1)c(2)Lc(k)Lc(K) \qquad (k = 1,2,L,K)$$

where

$$c(k) = (i'(k), i(k))$$

This is the k-th relation between $(x_0, x_0^{(l)})$

Then the distance is rewrite with the following equation,

$$d_s^{(l)}(x) = Min\left\{\frac{\sum_{j=1}^{J} w_j d_j(x_{i'}, x_i^{(l)})}{\sum_{j=1}^{J} w_j}\right\}$$
$$(i' = 1,2,\cdots,I')(i = 1,2,\cdots,I)$$

Where wj denotes k-th weighting coefficient which allows adjustment, or normalization of the distance d_s from -1 to 1. Figure 6 shows an enlarged portion of Figure 5. Weighting coefficients can be determined as shown in Figure 6.

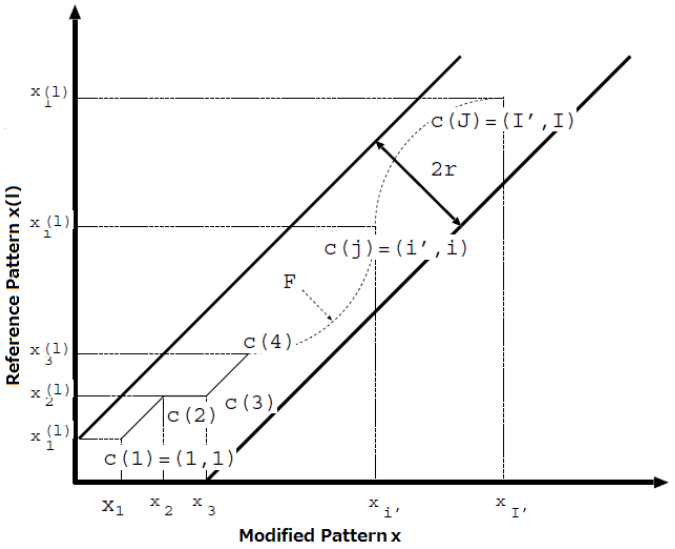

Fig. 5. Relation between reference pattern and input pattern data (Modified Pattern)

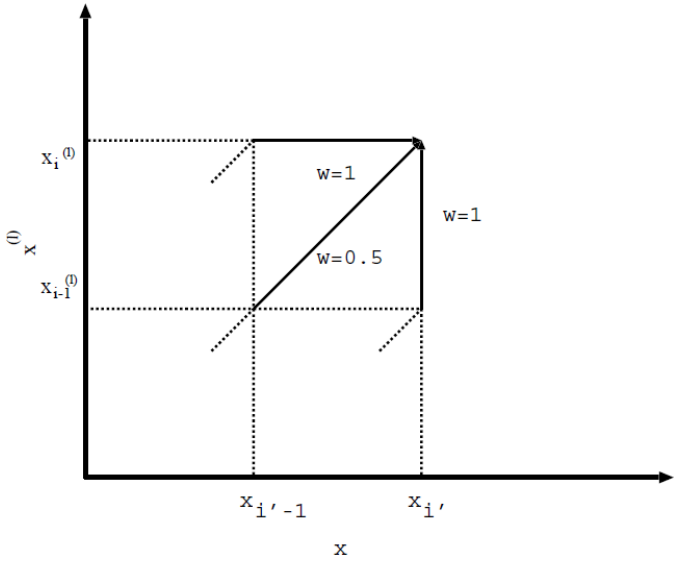

Fig. 6. Enlarged portion of Figure 5

There are some conditions for the distance definition,

Start and end of input pattern and reference pattern are corresponded,

The voice element orders have to be same for both input and reference patterns,

The corresponding reference pattern exists near by the input pattern.

Then, as shown in Figure 6, F is calculated as follows,

$$F = \begin{cases} x_{i'-1} - x_i^{(l)} \text{ and } x_{i'} - x_i^{(l)} \\ x_{i'-1} - x_{i-1}^{(l)} \text{ and } x_{i'} - x_i^{(l)} \\ x_{i'} - x_{i-1}^{(l)} \text{ and } x_{i'} - x_i^{(l)} \end{cases} \quad w_j = \begin{cases} -1 \\ 0.5 \\ 1 \end{cases}$$

Then the distance between input pattern data of voice element and the reference voice element pattern is represented as follows,

$$d^{(l)}(x) = \frac{1}{I + I'} Min \left\{ \sum_{j=1}^{J} w_j d_j(x_{i'}, x_i^{(l)}) \right\}$$

Where I and I' denotes the number of reference voice element patterns, respectively. Thus input voice element pattern is classified to the reference pattern, namely, if the d_s is minimum when the $l=l_0$, then the input pattern data is classified to l_0.

F. Voice Elements

In this paper, Japanese language recognition is focused. Japanese, in particular, the following 40 voice sounds are concerned.

Vowel: "a, i, u, e, o"

Consonant + Vowel: "ka, ki, ku, ke, ko, sa, si, su, se, so, ta, ti, tu, te, to, na, ni, nu, ne, no, ha, hi, hu, he, ho, ma, mi, mu, me, mo, ya, yu, yo,ra, ri, ru, re, ro, wa, and nn"

Voice element is defined as vowel and consonant, separately. Therefore, "a, i, u, e, o, k, s, t, n, h, m, y, r, w" are major concern. In this paper, voice recognition for these 14 vowels and consonants are concentrated.

III. EXPERIMENTS

First, the reference patters of the aforementioned 40 voice sounds are prepared with four different speakers. Sounds and moving pictures are prepared as the reference patterns.

For the optical flow based voice recognition, moving vectors of the aforementioned four features, top, bottom, left end, and right ends of mouth which are extracted from the moving pictures are used. Features are represented as the symbol. One small example of a portion of the time series of symbolized voice elements are shown in Figure 7.

In accordance with the distance, the first (L1), the second (L2), and the third (L3) candidates are determined. From the calculated distance, likelihood, or probability is also calculated for each candidate. The probability is calculated by voice element by voice element and also is evaluated for both vowels and consonants. The proposed method is based on forward and backward optical flow as explained in the second section. The probability evaluations have been done for the proposed method and compared to forward optical flow based method as well as the conventional voice recognition method.

Probability or likelihood is corresponding to the percent correct classification: PCC. If the PCC is evaluated with the first candidate only, then PCC for the conventional voice recognition method is not so good, below 43% for vowels and 14.3% for consonant + vowel while that for the proposed method with forward optical flow is 71.4% for vowel and 57.1% for consonant + vowel. Therefore, it is found that PCC is improved remarkably by taking moving picture analysis with the forward optical flow into consideration by the factor of approximately 30%.

$$\begin{matrix} 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 4 & -4 & 4 & 3 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \\ \#a & & & \\ 4 & -2 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \\ \#ra & & & \end{matrix}$$

Fig. 7. Symbolized voice elements for "a", and "ra"

TABLE I. PROBABILITY EVALUATION FOR THE FIRST TO THIRD CANDIDATES FOR THE PROPOSED AND THE METHOD WITH FORWARD OPTICAL FLOW ONLY AS WELL AS THE CONVENTIONAL VOICE RECOGNITION METHOD METHOD

		L1	L2	L3
Conventional	Vowel	42.9	71.4	71.4
	Consonant + Vowel	14.3	42.9	42.9
Forward optical flow	Vowel	71.4	85.7	100
	Consonant + Vowel	57.1	57.1	100
Forward and Backward	Vowel	90.3	95.5	100
	Consonant + Vowel	77.4	82.7	100

Furthermore, the proposed method with backward optical flow for confirmation and correction of recognized results which are obtained from the proposed method with forward optical flow only is superior to the proposed method with forward optical flow only. This implies that PCC is improved remarkably by taking confirmation and correction of recognized results which are obtained from the proposed method with forward optical flow only into account by the factor of about 20%.

PCC of vowel is always better than that of consonant + vowel, obviously. In particular for the conventional voice recognition method, there is around 30# of difference between vowel PCC and PCC of vowel + consonant.

If PCC is evaluated with the first to the third candidates, both of the proposed method with forward optical flow only

and that with forward and backward optical flow shows 100% of PCC. This implies that the effect of considering not only voice signals but also moving pictures on PCC of voice recognition is significant

As the results, it is found that the voice recognition performance can be improved by adding moving picture analysis to the voice signal analysis. This is same thing for human to human conversations. By looking at the speakers mouth movement, voice recognition is helped and reconfirmed recognized results at the same time.

ACKNOWLEDGMENT

The author would like to thank Mr. Shinji Matsuda for his effort to conduct the experiments.

REFERENCES

[1] Hongbing Hu, Stephen A. Zahorian, (2010) "Dimensionality Reduction Methods for HMM Phonetic Recognition," ICASSP 2010, Dallas, TX, http://bingweb.binghamton.edu/~hhu1/paper/Hu2010Dimensionality.pdf .(accessed on September 14 2012)

[2] Huston SJ, Krapp HG (2008). Kurtz, Rafael. ed. "Visuomotor Transformation in the Fly Gaze Stabilization System". *PLoS Biology* 6 (7): e173. doi:10.1371/journal.pbio.0060173. PMC 2475543. PMID 18651791. http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0060173.(accessed on September 14 2012)

[3] Andrew Burton and John Radford (1978). *Thinking in Perspective: Critical Essays in the Study of Thought Processes*. Routledge. ISBN 0-416-85840-6. http://books.google.com/?id=CSgOAAAAQAAJ&pg=PA77&dq=%22optical+flow%22+%22optic+flow%22+date:0-1985.(accessed on September 14 2012)

[4] David H. Warren and Edward R. Strelow (1985). *Electronic Spatial Sensing for the Blind: Contributions from Perception*. Springer. ISBN 90-247-2689-1. http://books.google.com/?id=-I_Hazgqx8QC&pg=PA414&dq=%22optical+flow%22+%22optic+flow%22+date:0-1985.(accessed on September 14 2012)

[5] S. S. Beauchemin , J. L. Barron (1995). *The computation of optical flow*. ACM New York, USA http://portal.acm.org/ft_gateway.cfm?id=212141&type=pdf&coll=GUIDE&dl=GUIDE&CFID=72158298&CFTOKEN=85078203.(accessed on September 14 2012)

[6] David J. Fleet and Yair Weiss (2006). "Optical Flow Estimation". In Paragios et al.. *Handbook of Mathematical Models in Computer Vision*. Springer. ISBN 0-387-26371-3. http://www.cs.toronto.edu/~fleet/research/Papers/flowChapter05.pdf.(accessed on September 14 2012)

[7] John L. Barron, David J. Fleet, and Steven Beauchemin (1994). "Performance of optical flow techniques". *International Journal of Computer Vision* (Springer). http://www.cs.toronto.edu/~fleet/research/Papers/ijcv-94.pdf.(accessed on September 14 2012)

[8] B. Glocker, N. Komodakis, G. Tziritas, N. Navab & N. Paragios (2008). *Dense Image Registration through MRFs and Efficient Linear Programming*. Medical Image Analysis Journal. http://vision.mas.ecp.fr/pub/mian08.pdf.(accessed on September 14 2012)

[9] Christopher M. Brown (1987). *Advances in Computer Vision*. Lawrence Erlbaum Associates. ISBN 0-89859-648-3. http://books.google.com/?id=c97huisjZYYC&pg=PA133&dq=%22optic+flow%22++%22optical+flow%22. (accessed on September 14 2012)

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science, and Technology of the University of Tokyo from 1974 to 1978 also was with National Space Development Agency of Japan (current JAXA) from 1979 to 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He was appointed professor at Department of Information Science, Saga University in 1990. He was appointed councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was also appointed councilor of Saga University from 2002 and 2003 followed by an executive councilor of the Remote Sensing Society of Japan for 2003 to 2005. He is an adjunct professor of University of Arizona, USA since 1998. He also was appointed vice chairman of the Commission “A” of ICSU/COSPAR in 2008. He wrote 30 books and published 332 journal papers