Acoustic Head Gesture Recognition and Its Applications

Akira Sasou\textsuperscript{1}, Yasuharu Hashimoto\textsuperscript{1}, Katsuhiko Sakaue\textsuperscript{1}

\textsuperscript{1} Advanced Industrial Science and Technology, AIST
{a-sasou,hashimoto.y,k.sakaue}@aist.go.jp

Abstract
When humans communicate with each other, they use not only speech but also several gestures such as facial expression, gaze, head movements, hand movements, and body posture. In this paper, we propose a novel method for recognizing head gestures that accompany speech. The proposed method tracks head movements that accompany speech by localizing the mouth position with a microphone array system. The proposed system is based only on acoustic information and never utilizes visual information. We also propose a recognition method for the mouth-position trajectory, in which Higher-Order Local Cross Correlation is applied to the trajectory. The recognition accuracy of the proposed method was on average 90.25\% for nineteen kinds of head gesture recognition tasks conducted in an open test manner, which outperformed the Hidden Markov Model-based method.

Index Terms: head gesture recognition, higher-order local cross correlation, microphone array

1. Introduction
When humans communicate with each other, they use not only speech but also several gestures such as facial expression, gaze, head movements, hand movements, and body posture. In the previous studies [1,2], a video camera was adopted for capturing head movements, in which the head pose is calculated from feature points such as the eye corners, pupils, nostrils, and so on. On the other hand, we have applied a microphone array to capture the head movements accompanying speech. The microphone array, which was originally developed for the purpose of achieving noise robust speech recognition [3], needs to localize the position of the user’s voice and the arrival directions of surrounding noises to enhance the user’s voice by beam forming. Because the localized position of the user’s voice almost corresponds with that of the mouth, the tracking of the head movements accompanying speech can be achieved by means of the microphone array. In our previous work [4], we developed a smart chair on which a microphone array is mounted so that they are mutually orthogonal. Each circuit board in a size of W130 × D10 × H5 mm has four silicon microphones soldered every 3 cm linearly. Although this microphone array has the ability to localize the user’s utterance position in three dimensions, in this paper, we utilize the X-Y coordinates only. The sampling frequency is set to 11.025 kHz. In order to distinguish the user’s voice and surrounding noises, we define the User Utterance Area (UUA), which has a volume of 20 × 20 × 20 cm\(^3\) in front of the microphone array. If a sound is localized in the UUA, the system accepts the sound as the user’s utterance. When a sound occurs outside of the UUA, the system rejects the sound as a noise. Therefore, the microphone array can easily distinguish the user’s voice from other’s voices and/or noises without training procedures such as speaker identification. We have adopted the MUltiple SIgnal Classification (MUSIC) method [5] for the User Utterance Localization (UUL).

2. Microphone Array
Figure 1 shows the three-axis microphone array we have developed, which consists of three circuit boards placed such that they are mutually orthogonal. Each circuit board in a size of W130 × D10 × H5 mm has four silicon microphones soldered every 3 cm linearly. Although this microphone array has the ability to localize the user’s utterance position in three dimensions, in this paper, we utilize the X-Y coordinates only. The sampling frequency is set to 11.025 kHz. In order to distinguish the user’s voice and surrounding noises, we define the User Utterance Area (UUA), which has a volume of 20 × 20 × 20 cm\(^3\) in front of the microphone array. If a sound is localized in the UUA, the system accepts the sound as the user’s utterance. When a sound occurs outside of the UUA, the system rejects the sound as a noise. Therefore, the microphone array can easily distinguish the user’s voice from other’s voices and/or noises without training procedures such as speaker identification. We have adopted the MUltiple SIgnal Classification (MUSIC) method [5] for the User Utterance Localization (UUL).

3. Recognition of Head Gestures
In order to achieve the recognition of a head gesture, we propose a simple and effective method for extracting a feature from the trajectory of the X-Y coordinates of the continuous user utterance positions localized by the microphone array. Hereinafter, we call this “utterance trajectory.” The feature is obtained by calculating the HLCC from the utterance trajectory.

3.1. Feature Extraction Based on HLCC
In the following, let \(x(t)\) and \(y(t)\), \(t=0,\ldots,T-1\), denote the utterance trajectory. First, the bias elements of the utterance trajectory are eliminated according to the following equations:
Second, the time length of the utterance trajectory is normalized to $T_n$:

$$x'(t) = x(t) - \bar{x}, \bar{x} = \sum_{m=0}^{\tau-1} x(t)/\tau$$

$$y'(t) = y(t) - \bar{y}, \bar{y} = \sum_{m=0}^{\nu-1} y(t)/\nu$$

Second, the time length of the utterance trajectory is normalized to $T_n$:

$$x'(t) = \{x'[t_1+1] - x'[t_N]\} \Delta t + x'[t_N]$$

$$y'(t) = \{y'[t_1+1] - y'[t_N]\} \Delta t + y'[t_N]$$

$$t_n = T / T_n, t_n = [t_1, \Delta t, t_2 - t_1]$$

The HLCC is then calculated according to:

$$f(m) = \sum_{i=0}^{K+1} \sum_{p=1}^{T_n} \{x'[t_i-1] \}^{p+1} \times \{x'[t_i - K]\}^{p_1}$$

where the matrix $m$ represents the local pattern, given by:

$$m = \begin{bmatrix} p_k & p_{k-1} & \cdots & p_0 \ q_k & q_{k-1} & \cdots & q_0 \end{bmatrix}$$

In the above equations, $(K+1)$ represents the length of the local pattern, and the order of the local pattern is defined as $\sum_{m=0}^{\nu=0} (p_i + q_i) - 1$. The HLCC feature vector is then constructed from the HLCCs calculated by using all the local patterns as:

$$f = [f(m_1), f(m_2), \cdots, f(m_D)]$$

where $D$ is the number of local patterns.

![Figure 2: Example of local patterns.](image)

Figure 2 shows the local patterns of length 3, and the orders in range from 0 to 2, where 49 patterns exist. Figure 3 shows an example of the utterance trajectories obtained by the microphone array, where the user moves his head such that the trajectory of the mouth position draws each numeral from 0 to 2. The HLCC feature vectors represented in Figure 4 were obtained by applying the 49 local patterns in Figure 2 to the modified utterance trajectories that are obtained by Eqn. (9), so that the length is normalized to $T_n = 20$.

![Figure 3: Example of utterance trajectories.](image)

3.2. Subspace Method-Based Recognition

The training procedure consists of the following steps. First, we collect a number of utterance trajectories of each head gesture. Let $f_{g,i}$ represent the HLCC feature vector calculated from the $i$th utterance trajectory of the $g$th head gesture, and let $N_g$ represent the number of utterance trajectories of the $g$th head gesture. Second, the principal component vectors are obtained by the principal component analysis to construct the subspace of each head gesture as follows. The auto-correlation matrix is given by:

$$A_g = \sum_{i=0}^{N_g} f_{g,i} f_{g,i}^T$$

Let $\lambda_{g,k}$ and $u_{g,k}$, $k = 1, \cdots, D$, represent the eigenvalues and the corresponding eigenvectors of the auto-correlation matrix respectively, where the eigenvalues are sorted in descending order. By means of the cumulative proportion given by:

$$\eta_{g,k} = \sum_{l=0}^{K=k} A_{g,l}$$

the dimension of each subspace is determined according to:

$$K_g = \min\{K | \eta_{g,K} \geq Q\}$$

With the orthonormal bases $U_g = [u_{g,1}, \cdots, u_{g,K_g}]$ of the subspace, the projection matrix is given by:

$$P_g = U_g U_g^T$$

In the recognition procedure, we first evaluate the HLCC feature vector $f$ from the utterance trajectory. The squared-norm of the projection on each subspace is then evaluated as:

$$I_g = [P_g f]^2 = [U_g^T f]^2$$

As the recognition result, we adopt the $\hat{g}$th head gesture that maximizes Eqn. (17) as:

$$\hat{g} = \arg \max_g I_g$$

4. Experimental Results

4.1. Accuracy of Head Gesture Recognition

In order to evaluate the recognition accuracy of the HLCC feature-based method, we collected nineteen kinds of head gestures from five subjects. The subjects made a fricative sound while moving their heads. The head gestures collected in this experiment are as follows. Each subject moved his head in a similar way as drawing the numerals 0 to 9, as shown in Figure 3. In addition, the head gestures made by moving the head directly in eight different directions and the head gestures.

![Figure 4: Example of HLCC feature vectors.](image)
made by randomly moving the head were collected. The subjects made the same gestures five times.

The HLCC feature vectors were obtained by using the local patterns of the lengths and the maximum orders ranging from 2 to 11 and from 1 to 3, respectively. The maximum order $N$ means that the local patterns include those of the orders in a range from $0^{th}$ to $N^{th}$. The $T_n$ in Eqn. (9) was set to 20, and the $Q$ in Eqn. (15) was set to 0.9999. The recognitions were conducted in an open test manner, that is, the training procedure was conducted based on the data of four subjects, and based on this the remaining subject was tested. Figure 7 shows the average recognition accuracy across the five subjects. In the case where the maximum order was 2 and the length was 8, the best score of 90.25% was obtained.

We also evaluated the recognition accuracy of the HMM-based method, which was conducted in the same manner as the previous experiments. Instead of the HLCC feature vector, the time series of the feature vectors was constructed from the vectors consisting of $x'(t)$ and $y'(t)$ in Eqn. (8) as the elements. The network topology of the HMMs is the left-to-right model, and the number of states and the number of mixtures were set in ranges from 10 to 30 and from 1 to 4, respectively. Figure 8 shows the results. The best score of 89.48% was obtained by using the HMMs having twenty-six states and one mixture. Even though the proposed method is very simple in comparison with the HMM-based method, its best score is comparable or even slightly better than the HMM-based method.

5. Applications

5.1. Powered wheelchair control

Various voice-driven wheelchairs have been developed to try to provide disabled people the mobility they desire [6-9]. However, problems with conventional voice-driven wheelchairs still exist. For instance, such wheelchairs use a headset microphone that can record the user’s command voice at a higher Signal-to-Noise Ratio (SNR) even with the presence of surrounding noise. However, disabled people need to put on the headset microphone each time they use the voice-driven wheelchair. When the headset microphone moves away from the user’s mouth, disabled people must also adjust the position of the headset microphone themselves. These actions are not always easy for disabled people, so we do not think headset microphones are practical. In addition, it is not easy for some disabled people to utter voice commands clearly. Such inarticulate commands cause inaccurate speech recognition.

In this application, we have applied the proposed method to an interface for controlling a powered wheelchair. The proposed interface does not require disabled people to wear any microphones or utter recognizable voice commands. Figure 7 presents the wheelchair we have developed. This microphone array system consists of two circuit boards. Each circuit board has four silicon microphones soldered every 3cm linearly. The circuit boards are placed along the diagonals of square black sponges. Considering the surrounding noise, we would like to put microphones as close to the user’s mouth as possible. However, such microphones would be dangerous for some disabled people, for instance, those having cerebral palsy with involuntary movements. The microphone should therefore be placed far enough from the user’s mouth that it does not touch the user’s head. Because the black sponges containing the microphone array circuit boards are placed on the edges of the arm rests, the user’s head never touches the microphone array system even when there are involuntary movements.

The proposed interface requires only two capabilities: the ability to move the head and the ability to utter an arbitrary sound. Some researchers developed gaze or head gesture interfaces for powered wheelchair control based on visual
5.2. Home electronic device control

A remote control or switch is widely used as an input device for operating home appliances. However, these devices are hand-controlled, so handicapped people who can hardly move their hands are unable to operate home appliances. Alternative input interfaces have been developed to avoid these difficulties. One such interface, for instance, uses camera-based recognition of gestures. Another interface uses the recognition of voice commands. The interface that uses camera-based recognition of gestures recognizes the movements of the hand and the face and executes the corresponding operations. For handicapped people who can hardly move their hands, it is popular to use movements of the face. However, when face movements are used for operations, users cannot move their heads for any other purpose, such as looking to the side. Another problem is that cameras and computers that can reliably detect face movements are comparatively expensive. An input interface using the recognition of voice commands recognizes voice commands uttered by users into a microphone and executes the corresponding operations. Users have only to speak the voice command corresponding to the operation to be executed. However, users need to be able to utter clear voice commands so that the speech recognition system can recognize them correctly. Handicapped people who can hardly speak are unable to use this interface.

Such examples suggest that traditional interfaces haven’t supported handicapped people who can hardly speak or move their hands. We think such people must have an interface that they can use. Therefore, we propose a novel interface that controls home appliances with the acoustic-based recognition of head gestures accompanying speech. Figure 8 illustrates how the proposed interface is used to control home appliances. The user can control home appliances through a GUI operated by gestures. Here, the user controls the TV, lights, and the audio system. Unlike in the interface using camera-based recognition of gestures, in the proposed interface using acoustic-based recognition of gestures, the user can freely move his/her head as long as he/she does not speak at the same time. This interface does not recognize users’ voices, therefore users are not required to utter clear voice commands. A subject who can hardly move his hands evaluated the proposed interface. The subject tried seven gestures ten times.

The average recognition rate of these gestures was 85.7%. The results confirmed the feasibility of the home appliance control interface using acoustic-based recognition of head gestures accompanying speech.

6. Conclusions

In order to achieve the recognition of head gestures accompanying speech, we have proposed a novel method that adopts a microphone array for localization of the utterance position, and that recognizes the utterance trajectory by means of an HLCC-based feature. From the experimental results, we have confirmed the feasibility of the proposed method.

7. Acknowledgements

This work was supported by KAKENHI 20700471, funded by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of the Japanese Government.

8. References