

Acoustic-Based Recognition of Head Gestures Accompanying Speech

Akira Sasou¹, Yasuharu Hashimoto¹, Katsuhiko Sakaue¹

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

Abstract

Head movements are linked not only to symbolic gestures, such as head-nodding to represent “yes” or head-shaking to represent “no,” but also to the production of suprasegmental features of speech, such as stress, prominence, and other aspects of prosody. Recent studies have shown that head movements play a more direct role in the perception of speech. 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 an 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. These gestures can help in the understanding of the spoken text. For instance, it is well known that the intelligibility of degraded auditory speech is improved when listeners view a speaker’s lip movements [1]. Another example is head movement, which are linked not only to symbolic gestures, such as head-nodding to represent “yes” or head-shaking to represent “no,” but also to the production of suprasegmental features of speech, such as stress, prominence, and other aspects of prosody [2]. Stress on a word is often accompanied by a nod of the head. A rising voice at the end of a phrase can be emphasized with a rise of the head, possibly combined with rising eyebrows [3,9]. It has been shown that the presence of visible head movements improve the intelligibility of Japanese sentences in a speech-in-noise task [4].

In the previous studies [3,5], 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 [6], 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 [7], we developed a smart chair on which a microphone array is mounted so that

the speech recognition system can understand what the user is referring to from the head orientation, which is estimated from the mouth position at the beginning of the utterance. Although the system can utilize the static orientation of the user’s head, it was unable to recognize a trajectory of the continuous head orientations that frame a head gesture.

In this paper, we propose a novel method for recognizing the head gestures accompanying speech. This proposed method, which localizes the mouth position with a microphone array, is based only on acoustic information and never utilizes visual information. We also propose a recognition method of head movement trajectory, in which Higher-Order Local Cross Correlation (HLCC) is applied to the time series of the localized mouth positions. The proposed system has the advantage of constructing a system that integrates head gesture recognition and speech recognition without using a visual device, and instead by simply extending the noise robust speech recognition system.

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 cm³ 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.

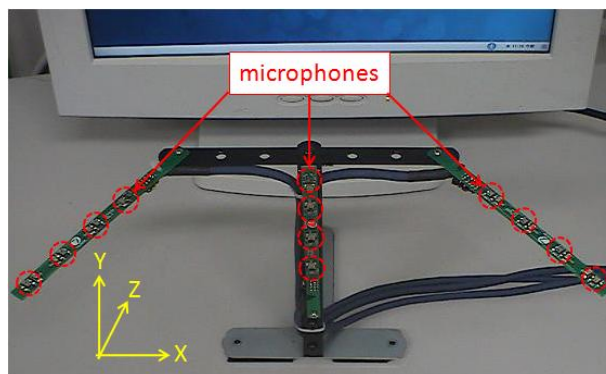


Figure 1: Three-axis microphone array.

2.1. User Utterance Localization

We have adopted the Multiple Signal Classification (MUSIC) method [8] for the User Utterance Localization (UUL). We assume that a sound occurring in the UUA is received as a spherical wave by the microphone array. The steering vectors are defined as follows:

$$\mathbf{P}_q = [Px_q, Py_q, Pz_q]^T, q = 1, \dots, 12$$

$$R_q = |\mathbf{P}_q - \mathbf{P}_0| = \sqrt{(Px_q - Px_0)^2 + (Py_q - Py_0)^2 + (Pz_q - Pz_0)^2}$$

$$\tau_q = R_q / v, g_q = g(\omega, R_q) \quad (1)$$

$$\mathbf{a}(\omega, \mathbf{P}_0) = [g_1 e^{-j\omega\tau_1}, \dots, g_{12} e^{-j\omega\tau_{12}}]^T$$

where \mathbf{P}_0 represents the position of the sound source in the UUA, $\mathbf{P}_1 \dots \mathbf{P}_{12}$ represent the positions of the microphones,

R_q is the distance between the q th microphone and the sound source, and v is the sound velocity.

The spatial correlation matrix is defined as:

$$\mathbf{R}(\omega) = (1/N) \sum_{n=1}^N \mathbf{y}(\omega, n) \mathbf{y}^H(\omega, n) \quad (2)$$

where $\mathbf{y}(\omega, n) = [Y_1(\omega, n), \dots, Y_{12}(\omega, n)]^T$, and $Y_q(\omega, n)$ represents the FFT of the n th frame received by the q th microphone. The eigenvalue decomposition of $\mathbf{R}(\omega)$ is given by:

$$\mathbf{R}(\omega) = \mathbf{E}(\omega) \mathbf{\Lambda}(\omega) \mathbf{E}^{-1}(\omega) \quad (3)$$

where $\mathbf{E}(\omega)$ denotes the eigenvector matrix, which consists of the eigenvectors of $\mathbf{R}(\omega)$ as $\mathbf{E}(\omega) = [\mathbf{e}_1(\omega), \dots, \mathbf{e}_{12}(\omega)]$, and $\mathbf{\Lambda}(\omega)$ denotes the eigenvalue matrix, which is defined as:

$$\mathbf{\Lambda}(\omega) = \text{diag}(\lambda_1(\omega), \dots, \lambda_{12}(\omega)) \quad (4)$$

$$\lambda_1(\omega) \geq \dots \geq \lambda_{12}(\omega)$$

The number of sound sources is estimated from the eigenvalues as follows. First, we evaluate the threshold value, which is defined as:

$$T_{egn}(\omega) = \lambda_1^{C_{egn}}(\omega) \times \lambda_{12}^{(1-C_{egn})}(\omega), 0 < C_{egn} < 1 \quad (5)$$

where C_{egn} is a constant that is adjusted experimentally. The number of sound sources $N_{snd}(\omega)$ is then estimated as the number of eigenvalues larger than the threshold value:

$$\lambda_1(\omega), \dots, \lambda_{N_{snd}}(\omega) \geq T_{egn}(\omega) \quad (6)$$

The eigenvectors, which correspond to these eigenvalues, form the basis of the signal subspace $\mathbf{E}_s(\omega) = [\mathbf{e}_1(\omega), \dots, \mathbf{e}_{N_{snd}}(\omega)]$.

The remaining eigenvectors $\mathbf{E}_n(\omega) = [\mathbf{e}_{N_{snd}+1}(\omega), \dots, \mathbf{e}_{12}(\omega)]$ are the basis of the noise subspace. User utterances are detected according to the following method. First, we search for the position \mathbf{P}_0 that absolutely maximizes the following value in the UUA:

$$Q(\mathbf{P}) = 1 / \sum_{\omega} |\mathbf{a}^H(\omega, \mathbf{P}) \mathbf{E}_n(\omega)|^2, \mathbf{P}_0 = \arg \max_{\mathbf{P} \in \text{UUA}} Q(\mathbf{P}) \quad (7)$$

If the absolute maximum value $Q(\mathbf{P}_0)$ exceeds the threshold value T_{usr} , we judge that the user made a sound. In the actual implementation, the user utterance position is evaluated on a grid of $10 \times 10 \times 10$ in the UUA. Thus, the utterance position is localized with an accuracy of 2 cm along each axis.

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 $y(t)$, $t=0, \dots, T-1$, denote the utterance trajectory. First, the bias elements of the utterance trajectory are eliminated according to the following equations:

$$x'(t) = x(t) - \bar{x}, \bar{x} = \sum_{t=0}^{T-1} x(t) / T \quad (8)$$

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

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

$$x''(\tilde{t}) = \{x'(t_b + 1) - x'(t_b)\} \times \Delta t + x'(t_b) \quad (9)$$

$$y''(\tilde{t}) = \{y'(t_b + 1) - y'(t_b)\} \times \Delta t + y'(t_b)$$

$$t_a = T \cdot \tilde{t} / T_n, t_b = \lfloor t_a \rfloor, \Delta t = t_a - t_b$$

The HLCC is then calculated according to:

$$f(\mathbf{m}) = \sum_{\tilde{t}=0}^{T_n-1} \{x''(\tilde{t})\}^{p_0} \{x''(\tilde{t}-1)\}^{p_1} \dots \{x''(\tilde{t}-K)\}^{p_K} \times \{y''(\tilde{t})\}^{q_0} \{y''(\tilde{t}-1)\}^{q_1} \dots \{y''(\tilde{t}-K)\}^{q_K} \quad (10)$$

where the matrix \mathbf{m} represents the local pattern, given by:

$$\mathbf{m} = \begin{bmatrix} p_K & p_{K-1} & \dots & p_0 \\ q_K & q_{K-1} & \dots & q_0 \end{bmatrix} \quad (11)$$

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

$$\mathbf{f} = [f(\mathbf{m}_1) \ f(\mathbf{m}_2) \ \dots \ f(\mathbf{m}_D)]^T \quad (12)$$

where D is the number of local patterns.

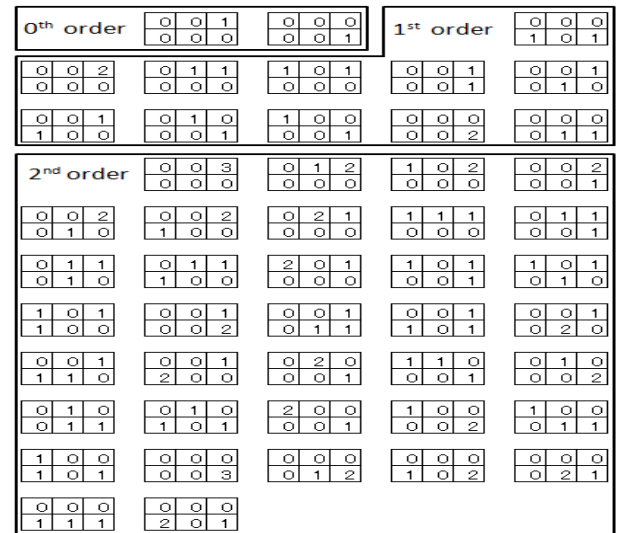


Figure 2: Example of local patterns.

Figure 2 shows the local patterns of length 3, and the orders in a range from 0^{th} to 2^{nd} , 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 $T_n = 20$.

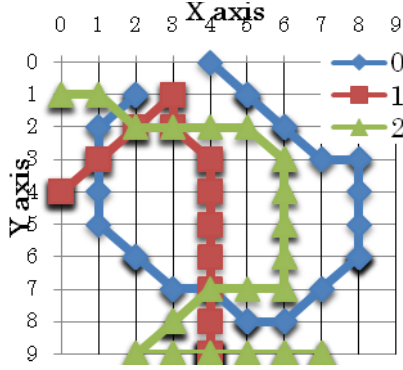


Figure 3: Example of utterance trajectories.

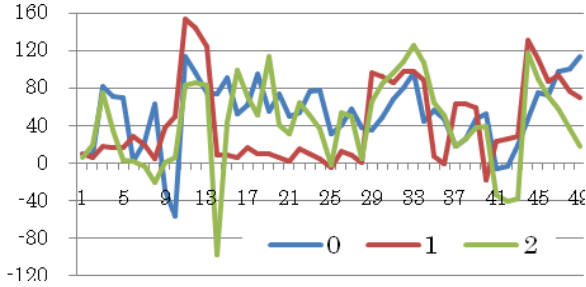


Figure 4: Example of HLCC feature vectors.

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 $\mathbf{f}_{g,i}$ 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:

$$\mathbf{A}_g = \sum_{i=1}^{N_g} \mathbf{f}_{g,i} \cdot \mathbf{f}_{g,i}^T / N_g \quad (13)$$

Let $\lambda_{g,k}$ and $\mathbf{u}_{g,k}$, $k=1, \dots, D$, 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:

$$\eta_{g,K} = \sum_{k=1}^K \lambda_{g,k} / \sum_{k=1}^D \lambda_{g,k} \quad (14),$$

the dimension of the each subspace is determined according to:

$$K_g = \min\{K | \eta_{g,K} \geq Q\} \quad (15)$$

With the orthonormal bases $\mathbf{U}_g = [\mathbf{u}_{g,1} \dots \mathbf{u}_{g,K_g}]$ of the subspace, the projection matrix is given by:

$$\mathbf{P}_g = \mathbf{U}_g \cdot \mathbf{U}_g^T \quad (16)$$

In the recognition procedure, we first evaluate the HLCC feature vector \mathbf{f} from the utterance trajectory. The squared-norm of the projection to each subspace is then evaluated as;

$$l_g = \|\mathbf{P}_g \cdot \mathbf{f}\|^2 = \|\mathbf{U}_g^T \cdot \mathbf{f}\|^2 \quad (17)$$

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

$$\hat{g} = \arg \max_g l_g \quad (18)$$

4. Experimental Results

In the following, we first present the evaluation results of the noise robustness of the UUL by means of the microphone array, and then present the recognition accuracy of head gestures based on the HLCC features, which is compared with that of the Hidden Markov Model (HMM)-based recognition method.

4.1. Noise Robustness of UUL by Microphone Array

To evaluate the noise robustness of the UUL, we recorded clean utterances of three subjects and the surrounding noises separately in a soundproof chamber. The noises were recorded by placing the loudspeaker facing the microphone array in each of four different places: in the front, the right, the back and the left, from which voices of a female and a male reading were emitted. The clean user utterances and noises were added by a computer at different SNRs ranging from 20dB to -10dB. The purpose of this experiment is to determine how much the UUL is affected by interference from the noises. In this way, we evaluated how much the positions localized from the noise-corrupted utterances differ from those of the clean utterances.

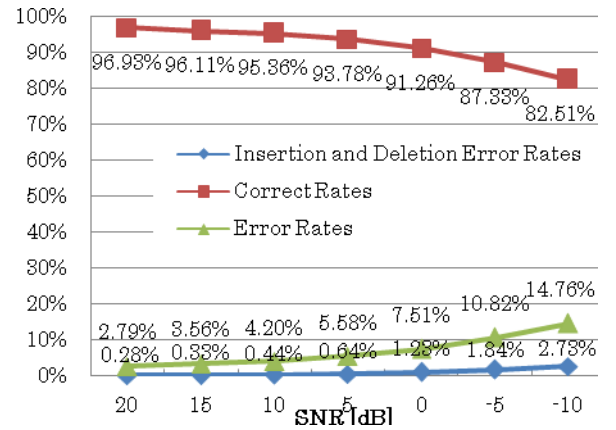


Figure 5: Accuracy of UUL in a noisy environment.

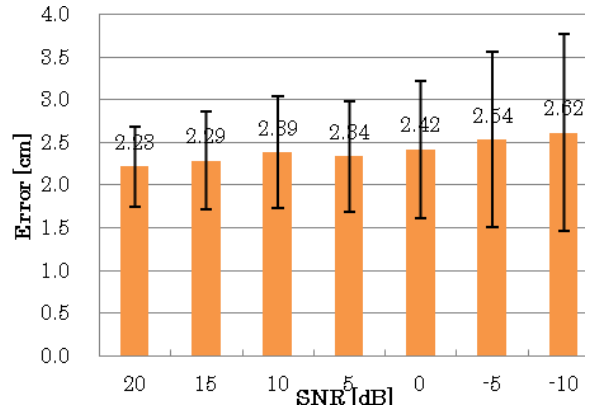


Figure 6: Error distance of UUL in a noisy environment.

Figure 5 shows the evaluated accuracy of the UUL in a noisy environment. The results are the average of all the subjects and the noises. The correct rate represents the rate at which the utterance position localized from the noise-corrupted utterance is completely the same as that of the clean utterance. The error rate represents the rate at which the localized utterance position differs from that of the clean utterance. Insertion errors occur when a user utterance is

mistakenly extracted. Deletion errors occur when the user utterance cannot be detected from the noise-corrupted utterance. Figure 6 shows the error distributions of the utterance positions localized from the clean and noise-corrupted utterances, in which the error bars represent the standard deviations. The average error distance of all the SNRs is less than approximately 2.5 cm, which nearly corresponds with the size of one grid. Therefore, although the error rates increase slightly as the SNR decreases, the actual error distance is almost negligible.

4.2. 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 moves 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 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. 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. We are now planning to develop a system that integrates head gesture recognition and speech recognition.

6. Acknowledgements

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

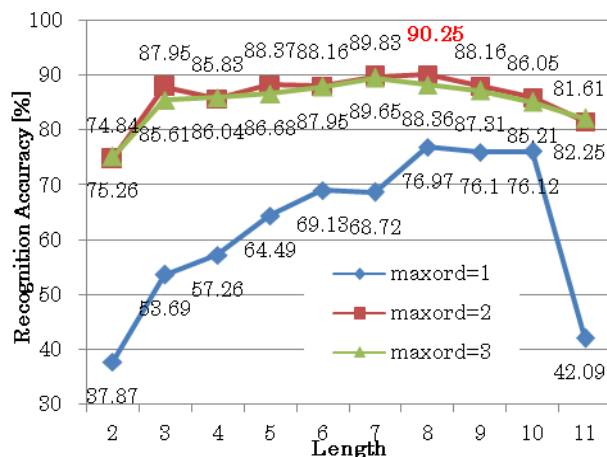


Figure 7: Recognition accuracies of the proposed method.

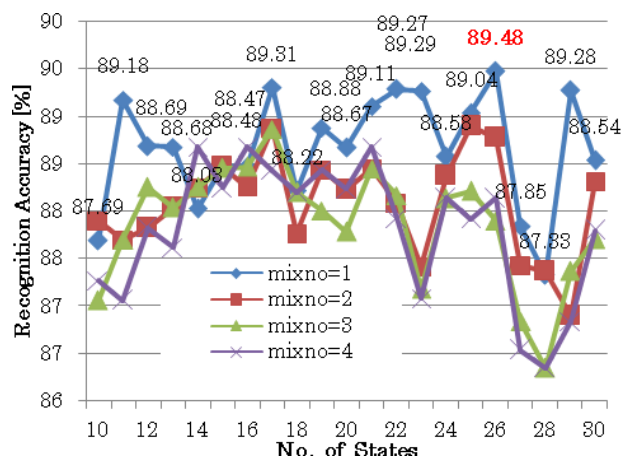


Figure 8: Recognition accuracies of the HMM-based method.

7. References

- [1] W.H.Sumby, I.Pollack, "Visual Contribution to Speech Intelligibility in Noise," J.Acoust. Soc. Am. Vol.26, No.2, pp.212-215, 1954.
- [2] U.Hardar, T.J.Steiner, E.C.Grant, F.C.Rose, "Head movement correlates of juncture and stress at sentence level," Language and Speech, Vol.26, No.2, pp.117-129, 1983.
- [3] H.P.Graf, E.Cosatto, V.Strom, F.J.Huang, "Visual Prosody: Facial Movements Accompanying Speech," Proc. of IEEE Intl. Conf. on Automatic Face and Gesture Recognition, FGR'02, 2002.
- [4] K.G.Munhall, Jeffery A.Jones, Daniel E.Callan, Takaaki Kuratate, Eric Vatikiotis-Bateson, "Visual Prosody and Speech Intelligibility," Psychological Science, Vol.15, No.2, pp.133-137, 2004.
- [5] J.W.Davis, S.Vaks, "A perceptual user interface for recognizing head gesture acknowledgements," Proc. of ACM Intl. Workshop on Perceptive user interface, pp.1-7, 2001.
- [6] A.Sasou, H.Kojima, "Noise robust speech recognition applied to voice-driven wheelchair," EURASIP Journal on Advances in Signal Processing, vol.2009, article ID 512314, 2009.
- [7] A.Sasou, "Head orientation estimation integrated speech recognition for the smart chair," Proc. of intl. symposium on Universal Communication, pp.482-489, 2008.
- [8] R.O.Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," IEEE Trans. Antennas Propag., Vol.AP-34, No.3, pp.276-280, 1986.
- [9] C.T.Ishi, C.R.liu, H.Ishiguro,N.Hagita, "Head Motion during Dialog Speech and Nod Timing Control in Humanoid Robots," Proc. of ACM/IEEE intl. conf. on Human-robot interaction, pp. 293-300, 2010