Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Phoneme Classification for Speech Synthesiser using Differential EMG Signals between Muscles

Nan BU, Toshio TSUJI, Jun ARITA, and Makoto OHGA

Abstract— **This paper proposes the use of differential electromyography (EMG) signals between muscles for phoneme classification, with which a Japanese speech synthesiser system can be constructed using fewer electrodes. In distinction from traditional methods using differential EMG signals between bipolar electrodes on the same muscle, an EMG signal is derived as differential between monopolar signals on two different muscles in the proposed method. Then, frequency-based feature patterns are extracted with filter banks, and classification of phonemes is realized by using a probabilistic neural network, which combines feature reduction and pattern classification processes in a single network structure. Experimental results show that the proposed method can achieve considerably high classification performance with fewer electrodes.**

I. Introduction

An artificial respirator is often required to assist breathing of patients with cervical spine injury, muscular dystrophy, *etc*. For these patients, speech rehabilitation is an important problem after tracheostomy, because communication is one of the critical issues related to their medical care and social interactions. To deal with this problem, a tracheostomy speaking valve was developed to redirect the air through the vocal cord. A variety of speaking valves have been proposed in the literature, and some commercial products are available [1].

Since speaking valves are inexpensive and easy for use, they have been accepted as a highly successful option for respirator-dependent patients. However, there are still some problems accompanied with the use of speaking valves, e.g., secretion accumulation, decrease of respiratory function after wearing for a few hours. Also, misuse of the speaking valve may endanger patients' lives. Although some methods, such as esophageal speech, can be used instead, the sound is rough and hoarse, and these methods are difficult to learn [2].

On the other hand, several speech synthesisers have been developed based on lip and jaw motions or electromyography (EMG) patterns of mimetic and cervical muscles [2]- [4]. In particular, Fukuda et al. proposed an EMG-based Japanese speech synthesiser system, where six Japanese phonemes (five vowels, i.e. /*a*/, /*i*/, /*u*/, /*e*/, /*o*/, and one nasal /*n*/) are classified from EMG patterns using a probabilistic neural network (NN), and then words are recognized from the series of phonemes using algorithms of the hidden Markov model (HMM) [4]. Due to the probabilistic NN and HMM algorithm, this system provides high performance of phoneme classification and word recognition, and is robust against issues such as differences among individuals and variation in temporal characteristics. However, in this system, differential EMG signals of bipolar electrodes are used, so that the number of electrodes is quite large.

To overcome this problem, the present paper proposes an alternative method for EMG pattern acquisition, which is based on differential EMG signals between muscles. In the field of EMG pattern classification, Ohga et al. first proposed this EMG acquisition method, and succeeded in classifying four or six motions of forearm for EMG-based prosthetic control using only two electrodes [5]. Distinct from traditional bipolar recording, electrodes are attached on different muscles, one electrode for each muscle, and then, differential signals between every two electrodes can be derived as input channels for classification. With this method, the number of electrodes can be reduced. Also, this paper develops a corresponding phoneme classification scheme for EMG patterns extracted. To acquire sufficient feature characteristics from the reduced EMG sources, frequency information of each channel is extracted using filter banks. It is apparent that when increasing the frequency resolution, dimensionality of the feature space would grow accordingly. The proposed speech synthesiser system incorporates a novel probabilistic NN, a reduced-dimensional log-linearized Gaussian mixture network (RD-LLGMN) [6], for classification of high-dimensional EMG patterns. With RD-LLGMN, it is expected that discriminative information can be extracted from frequency-based EMG patterns, and efficient classification of phonemes is possible.

This paper is organized as follows: In Section II, a brief review of the Japanese speech synthesiser system [4] is described. Then Section III explains details of the proposed EMG pattern acquisition method, and algorithm of the corresponding phoneme classification scheme. In Section IV, performance of the proposed method is verified with experimental results of healthy subjects and a patient with cervical spine injury. Finally, Section V concludes this paper.

II. EMG-based Speech Synthesiser System [4]

Structure of the Japanese speech synthesiser system [4] is shown in Fig. 1. This system can be divided into four parts: (1) EMG signal acquisition and feature extraction, (2) phoneme classification using an NN, (3) word recognition based on HMM, and (4) voice generation.

L pairs of Ag/AgCl electrodes are attached to the mimetic and cervical muscles, and differential EMG signals of bipolar electrodes are measured, as shown in Fig. 2. Then, the

Nan BU, Toshio TSUJI, and Jun ARITA are with the Department of the Artificial Complex Systems Engineering, Hiroshima University, Higashi-Hiroshima, 739-8527 JAPAN (e-mail: bu@ieee.org).

Makoto OHGA is with the Eastern Hiroshima Prefecture Industrial Research Institute, Fukuyama, 721-0974 JAPAN.

0-7803-8740-6/05/$20.00 ©2005 IEEE.

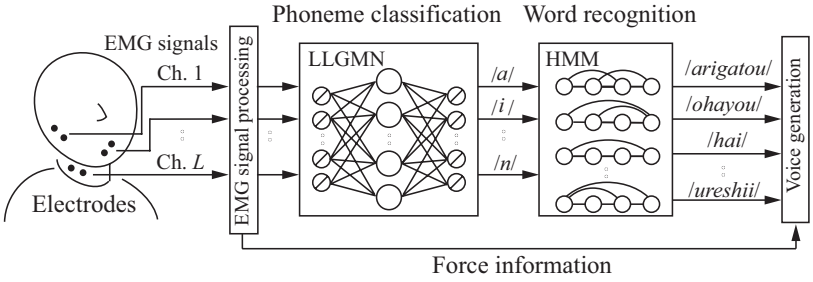


Fig. 1. Schematic view of the Japanese speech synthesiser system [4].

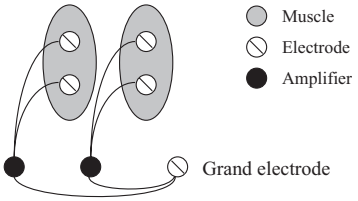


Fig. 2. Differetial EMG signals measured from the same muscles.

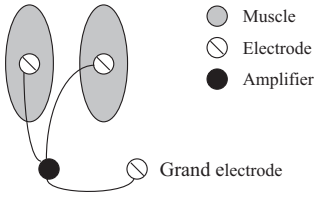


Fig. 3. Differetial EMG signal measured from the two muscles.

differential EMG signals are amplified and filtered out with a low-pass filter (cut-off frequency: 200 Hz), and digitized by an A/D converter (sampling frequency: 1 kHz).

The L channels of EMG signals are rectified and filtered out by a second-order Butterworth filter (cutoff frequency: 1 Hz). The filtered EMG signals are defined as $BEMG_l(t)$ $(l = 1, \cdots, L)$, and normalized to make the sum of L channels equal 1:

$$x'_l(t) = \frac{BEMG_l(t) - \overline{BEMG}_l^{st}}{\sum_{l'=1}^{L} BEMG_{l'}(t) - \overline{BEMG}_{l'}^{st}} \ (l = 1, \cdots, L), \tag{1}$$

where $\overline{BEMG}_l^{st}$ is the mean value of $BEMG_l(t)$ measured while relaxing the muscles. The feature vector $\mathbf{x}'(t) = [x'_1(t), x'_2(t), \ldots, x'_L(t)]^{\mathrm{T}}$ is used as input of a neural classifier. In this system, to recognize the beginning and ending of utterance, force information $F_{\mathrm{BEMG}}(t)$ is calculated as

$$F_{\mathrm{BEMG}}(t) = \frac{1}{L} \sum_{l=1}^{L} \frac{BEMG_l(t) - \overline{BEMG}_l^{st}}{BEMG_l^{max} - \overline{BEMG}_l^{st}}, \tag{2}$$

where $BEMG_l^{max}$ is the mean value of $BEMG_l(t)$ measured under the maximum voluntary contraction (MVC).

A probabilistic neural network, called a log-linearized Gaussian mixture network (LLGMN) [7], is employed for phoneme classification. This network estimates the posterior probability distribution of input features based on a Gaussian mixture model (GMM) [8] and a log-linear model. Since LLGMN integrates statistical models into the network architecture as prior knowledge, LLGMN can learn non-linear mapping between EMG patterns and Japanese phonemes using samples labeled with the corresponding phonemes. According to previous studies [2], [4], it is very difficult to recognize consonant only from mimetic and cervical EMG signals. In this system, five vowels, i.e. */a/*, */i/*, */u/*, */e/*, */o/*, and one nasal */n/*, are classified, and all the consonants are classified as corresponding vowels, for example, */ka/*, */sa/*, and */ta/* are classified as the vowel */a/*.

Due to the fact that only six phonemes can be used, HMM [9] is applied for Japanese word recognition, which has been successfully developed especially in the field of speech recognition. For recognition, one HMM is prepared for each word, for instance, */oaou/* for */ohayou/*, and */aeu/* for */taberu/*. When users utter */ohayou/*, model of */oaou/*, which consists of the sequence of vowels belonging to the word is recognized. Also, remarkable variance of utterance length exists. Since HMMs approximate the probabilistic characteristics of time series through learning, robust recognition can be achieved for words with varying temporal characteristics.

Finally, a voice synthesizing software (IBM Corp. Protalker 97) is adopted for sound generation. With this software, Japanese language can be naturally generated. Also, utterance speed, pitch frequency, and sound volume can be set freely. In this system, the sound volume is controlled corresponding to force information calculated from EMG signals.

III. Phoneme Classification using Differential EMG Signals between Muscles

In this paper, differential EMG signals between muscles are used to reduce the number of electrodes. Also, a bank of filters is used to extract frequency information, and then RD-LLGMN is utilized for phoneme classification.

A. EMG Signal Acquisition

With monopolar configuration, EMG signals are measured as shown in Fig. 3. Differential between electrodes is obtained, with which it is considered that characteristics of both muscles under the electrodes are represented. In the proposed method, S Ag/AgCl electrodes are attached to the mimetic and cervical muscles, where each muscle is attached with one electrode. EMG signals are recorded

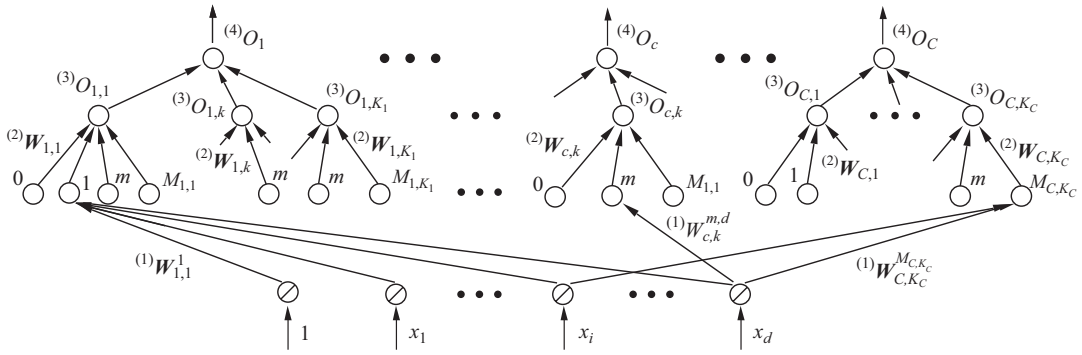


Fig. 4. Structure of RD-LLGMN.

with sampling frequency of 1 kHz, the difference between potentials of every two electrodes is computed, so that there are $S(S-1)/2$ channels of EMG signals available. L channels of EMG signals, $(L \leq S(S-1)/2)$, are then fed into the feature extraction process.

B. Feature Extraction

Since differential is derived from electrodes on different muscles, spatial information may be partially lost. A bank of Z band-pass filters, $(\mathrm{BPF}_i, i = 0, \cdots, Z-1)$, is applied to L channels to extract frequency information for compensation. The bandwidth of the ith filter is set as follows:

$$\mathrm{BPF}_i : 20 + \sigma i \text{ [Hz]} \sim 20 + \sigma(i+1) \text{ [Hz]}, \tag{3}$$

where $\sigma = U/Z$, and U is the frequency range under consideration. After the filter-bank stage, the number of input channels, denoted as d, becomes $L \times Z$, and the raw EMG signals of each channel are rectified and filtered by a low pass filter (cut-off frequency: 1 Hz). The filtered EMG signals are defined as $EMG_i(t)$ $(i = 1, \cdots, d)$, and normalized to make the sum of d channels equal 1.

$$x_i(t) = \frac{EMG_i(t) - \overline{EMG}_i^{st}}{\sum_{i=1}^{d} EMG_i(t) - \overline{EMG}_i^{st}} \quad (i = 1, \cdots, d), \tag{4}$$

where $\overline{EMG}_i^{st}$ is the mean value of $EMG_i(t)$, which is measured while relaxing the muscles. Then the normalized patterns, $\mathbf{x}(t) = [x_1(t), x_2(t), \cdots, x_d(t)]^{\mathrm{T}}$, are used as input to RD-LLGMN.

C. Phoneme Classification

RD-LLGMN [6] is used for phoneme classification, which provides a novel approach of feature reduction for finding discriminant features of a reduced size and calculates posterior probabilities for classification. There are two basic ideas in RD-LLGMN: 1) orthogonal transformation, which projects the original input space into a lower-dimensional space, and 2) Gaussian mixture models, which estimate probability distribution of patterns in the projected lower-dimensional space. This network combines the feature reduction process with the classification part, and is trained in a fashion of minimum classification error (MCE) learning [10], which enables the classification part to achieve lower error probability.

RD-LLGMN is a four-layer NN, the structure of which is shown in Fig. 4. Given an input vector $\mathbf{x} \in \Re^d$, the first layer consists of $d+1$ units, where one unit has a bias input of 1, and an identity function is used for activation of each unit. Let ${}^{(1)}O_i$ $(i = 0, \cdots, d)$ denote the output of the ith unit in the first layer, we have

$${}^{(1)}O_i = \begin{cases} 1, & i = 0 \\ x_i, & i = 1, 2, \cdots, d \end{cases} \tag{5}$$

where x_i $(i = 1, 2, \cdots, d)$ is the element of $\mathbf{x}$.

In the second layer, the unit $\{c, k, 0\}$, $(c = 1, \cdots, C;$ $k = 1, \ldots, K_c)$, is a bias unit, and its output ${}^{(2)}O_{c,k}^0 = 1$; the unit $\{c, k, m\}$ $(m = 1, \cdots, M_{c,k})$ receives the output of the first layer weighted by ${}^{(1)}W_{c,k}^m$, where C is the number of classes under consideration; K_c the number of components of the Gaussian mixture distribution in class c; and $M_{c,k}$ the number of dimension of component k in class c. The input ${}^{(2)}I_{c,k}^m$ and the output ${}^{(2)}O_{c,k}^m$, for $(m \neq 0)$, are defined as follows:

$${}^{(2)}I_{c,k}^m = \sum_{i=1}^{d} {}^{(1)}O_i \, {}^{(1)}W_{c,k}^{m,i}, \tag{6}$$

$${}^{(2)}O_{c,k}^m = ({}^{(2)}I_{c,k}^m)^2. \tag{7}$$

Through this layer, vector $\mathbf{x} \in \Re^d$ is projected into $M_{c,k}$-dimension spaces, $M_{c,k} < d$.

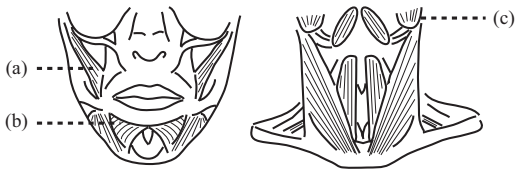
The unit $\{c, k\}$ in the third layer sums up outputs of the second layer weighted by coefficients ${}^{(2)}W_{c,k}^m$. The relationships between the input of unit $\{c, k\}$ in the third layer $({}^{(3)}I_{c,k})$ and the output $({}^{(3)}O_{c,k})$ are defined as

$${}^{(3)}I_{c,k} = \sum_{m=0}^{M_{c,k}} {}^{(2)}O_{c,k}^m \, {}^{(2)}W_{c,k}^m, \tag{8}$$

$${}^{(3)}O_{c,k} = \frac{\exp[{}^{(3)}I_{c,k}]}{\sum_{c'=1}^{C} \sum_{k'=1}^{K_{c'}} \exp[{}^{(3)}I_{c',k'}]}. \tag{9}$$

In this layer, RD-LLGMN calculates posterior probability of each Gaussian component using reduced-dimensional features.

The fourth layer consists of C units corresponding to the number of classes. Unit c sums up outputs of K_c components $\{c, k\}$ in the third layer. The function between the input and



(a) M. depressor labii inferioris (c) M. masseter
(b) M. zygomaticus major

Fig. 5. Mimetic and cervical muscles used in the experiments.

the output is described as

$$^{(4)}O_c = {}^{(4)}I_c = \sum_{k=1}^{K_c} {}^{(3)}O_{c,k}. \tag{10}$$

After only optimizing the weight coefficients with an MCE-based training algorithm, the output of RD-LLGMN, $^{(4)}O_c$, can estimate the posterior probability of class c.

In the proposed phoneme classification method, we assumed that the amplitude level of EMG signals changes in proportion to muscle force. A power level is defined as

$$F_{\mathrm{MEMG}}(t) = \frac{1}{S}\sum_{s=1}^{S} \frac{MEMG_s(t) - \overline{MEMG}_s^{st}}{MEMG_s^{max} - \overline{MEMG}_s^{st}}, \tag{11}$$

where $MEMG_s(t)$ indicates the filtered signal (cut-off frequency: 1 Hz) of rectified raw EMG directly measured from the electrode s ($s = 1, \cdots, S$), $\overline{MEMG}_s^{st}$ is the mean value of $MEMG_s(t)$, which is measured while relaxing the muscles, and $MEMG_s^{max}$ is the mean value of $MEMG_s(t)$ measured under MVC. $F_{\mathrm{MEMG}}(t)$ indicates the force information, and is used to recognize whether the motion has really happened or not, by comparing $F_{\mathrm{MEMG}}(t)$ with a predefined threshold M_d.

Entropy of RD-LLGMN's output is also calculated to prevent risk of misclassification. The entropy is defined as

$$H(t) = -\sum_{c=1}^{C} {}^{(4)}O_c(t) \log {}^{(4)}O_c(t). \tag{12}$$

If the entropy $H(t)$ is less than a threshold H_d, the specific motion with the largest probability is determined according to the Bayes' decision rule. If not, the determination is suspended.

IV. Experiments

Japanese phoneme classification experiments were conducted to examine performance of the proposed method. Four subjects (A, B, C: healthy, D: a patient with cervical spine injury) participated in these experiments.

A. Experimental Condition

Three Ag/AgCl electrodes (SEB120, GE marquette Corp.) were attached to subject's mimetic and cervical muscles (M. Depressor Labii Inferioris (DLI), M. Zygomaticus Major (ZM), and M. Masseer (MA); see Fig. 5). EMG signals measured from three muscles were recorded using monopolar electrodes (sampling frequency: 1 kHz). Differential between DLI and ZM was used as the input channel one, differential between DLI and MA as the channel two, and differential between ZM and MA as the channel three. The frequency range U was set at 250 Hz, and the number of band-pass filters Z was 6, so that the dimension of input features for RD-LLGMN d was 18. Parameters of GMM in RD-LLGMN were set as: $C = 6$, $K_c = 1$ ($c = 1, \cdots, 6$). Dimensions of the reduced subspaces $M_{c,k}$, ($c = 1, \cdots, C$; $k = 1$), were set as $M = 9$. In the training phase, 50 EMG patterns were extracted from EMG signals of each phoneme, and teacher signals consisted of $C \times 50$ patterns. The determination thresholds were set as $M_d = 0.08$, and $H_d = 0.5$.

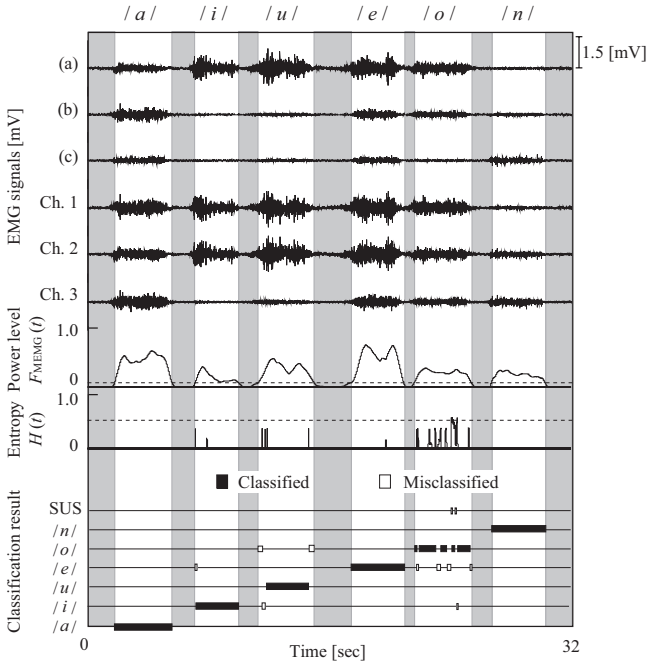


Fig. 6. Examples of the classification results (subject D) using the proposed method. (a: DLI, b: ZM, c: MA)

B. Phoneme Classification Results

An example of the classification results (subject D) is shown in Fig. 6. In this figure, three channels of raw monopolar EMG signals, three channels of differential EMG signals, the force information $F_{\mathrm{MEMG}}(t)$, the entropy $H(t)$, and the classification results are plotted. The gray areas indicate that no utterance occurred because the force information F_{MEMG} was less than M_d.

Misclassification is found in the utterance of /o/, and the beginning and the ending of utterance /u/. The classification rate is 90.6% in this experiment. Compared to healthy subjects, amplitude of EMG signals measured from muscles is lower. Since this patient eats soft meals everyday, it is considered that the muscles around the jawbone such as the muscle of masseter degrade. Also, for each misclassified utterance, the entropy is high. Misclassification could be reduced using an appropriately modulated threshold H_d.

C. Comparison Experiments

Accuracy of the classification results for four subjects was investigated as well. To verify the proposed method,

comparison experiments were conducted with four phoneme classification methods based on bipolar EMG recording:

- **BPNN**: Feature vector $\mathbf{x}' \in \Re^3$, defined as (1) was used as input signals, and a back propagation neural network (BPNN) [8] was applied for classification.
- **LLGMN**: Feature vector $\mathbf{x}' \in \Re^3$ was classified with the probabilistic NN, LLGMN [7]. This is the method used in the previously proposed speech synthesiser system [4].
- **BPNN with PCA**: The bank of six filters was applied to three channels of differential EMG signals in the same way as described in III-B, then a feature extraction process, principle component analysis (PCA) [8], was used to reduce the dimensionality to nine. After the PCA process, BPNN was applied.
- **LLGMN with PCA**: LLGMN with the PCA was used for classification.

LLGMN [7] is a three-layer feedforward probabilistic NN based on GMM. Number of units in the input layer of LLGMN was equal to the dimension of input vector. Units in the hidden layer correspond to the Gaussian components in GMM, the number of which was set as one. The output layer had C units, and each unit output posterior probability for the input pattern. On the other hand, BPNN had four layers, and units of the input layer and the hidden layers were set as 3, 10, and 10 in the BPNN method, and 9, 30, and 30 in the BPNN with PCA method, respectively. The number of units in the output layer was six. Each output of BPNN corresponds to one phoneme, and it was normalized to make the sum of all outputs equal 1.0, so that it can be regarded as the posterior probability of each phoneme. The same determination thresholds, M_d and H_d, were used for classification based on BPNN and LLGMN. LLGMN is trained with a maximum likelihood learning [7]. The training of BPNN continued until the mean squared error became less than 0.1, where the learning rate was 0.01. However, if the mean squared error after 50,000 iterations was still higher than 0.5, the learning procedure was stopped.

For all methods, five sets of randomly chosen initial weights were used to train each sample data. EMG signals measured for about 30 seconds (six phonemes) were tested. Table I shows mean values and standard deviations of the classification rates using five methods. It can be seen that RD-LLGMN using the proposed EMG pattern acquisition method outperformed all other methods, with electrodes of half amount. Comparing the classification results of BPNN and BPNN with PCA, frequency components extracted may provide important information for phoneme classification. However, similar difference cannot be found in the comparison between LLGMN and LLGMN with PCA.

TABLE I

COMPARISON OF CLASSIFICATION RATES BETWEEN THE PROPOSED METHOD AND OTHER METHODS.

Type of the methods		BPNN	LLGMN	BPNN with PCA	LLGMN with PCA	RD-LLGMN
The number of electrodes		6	6	6	6	3
Subject A	*C.R.*	43.7	40.3	68.9	94.8	88.4
	S.D.	6.7	0.1	0.9	0.2	8.5
Subject B	*C.R.*	39.8	79.1	65.0	69.5	69.6
	S.D.	13.4	0.3	1.8	0.9	6.2
Subject C	*C.R.*	33.1	77.2	52.2	93.5	92.2
	S.D.	12.0	0.1	1.8	0.1	3.4
Subject D	*C.R.*	59.9	89.1	53.6	74.4	90.3
	S.D.	1.8	0.1	7.1	0.3	5.4

C.R. : Classification rate [%] *S.D.* : Standard deviation [%]

V. CONCLUSION

This paper proposes a novel phoneme classification method for Japanese speech synthesiser system. This method uses differential EMG signals between muscles, and classification can be achieved based on fewer electrodes. To acquire sufficient feature characteristics from the reduced EMG sources, filter banks are applied to extract frequency information. With the probabilistic NN, RD-LLGMN, discriminative information is extracted from frequency-based EMG patterns with large dimensions, and efficient classification of phonemes is possible.

To examine the discrimination accuracy of the proposed method, phoneme classification experiments and comparison experiments have been carried out with four subjects. In the experiments, relatively high classification rates of the proposed method using less number of electrodes were confirmed.

In the future research, we would like to improve the preprocessing method of EMG signals, such as modulation of the parameters of filter banks and the low-pass filtering. Also, locations of electrodes and selection of monopolar channels should be investigated.

REFERENCES

[1] A.H. Shikani, J. French, and A.A. Siebens, "New unidirectional airflow ball tracheostomy speaking valve," *Otolaryngology-Head and Neck Surgery*, Vol. 123, No. 1, pp. 103-107, 2000.

[2] R.L. Goode, "Artificial laryngeal devices in post-laryngectomy rehabilitation," *The Laryngoscope*, Vol. 85, pp. 677-689, 1975.

[3] E.M. Mullar, "Strain gauge transduction of lips and jaw motion in the midsagittal plane: Refinement of a prototype sysytem," *Journal of the Acoustical Society of America*, Vol. 65, No. 2, pp. 481-486, 1979.

[4] O. Fukuda, S. Fujita, and T. Tsuji, "Substitute vocalization system based on EMG signals," *The IEICE Trans. on Information and Systems*, Vol. J86-D-II, No. 7, pp. 1-7, 2003. (in Japanese)

[5] M. Ohga, M. Takeda, A. Matsuba, A. Koike, and T. Tsuji, "Development of a five-finger prosthetic hand using ultrasonic motors controlled by two EMG signals," *Journal of Robotics and Mechatronics*, Vol. 14, No. 6, pp. 565-571, 2002

[6] N. Bu, and T. Tsuji, "Multivariate pattern classification based on local discriminant component analysis," *Proc. of the 2004 IEEE Int. Conf. on Robotics and Biomimetics*, Paper-ID 290, 2004.

[7] T. Tsuji, O. Fukuda, H. Ichinobe, and M. Kaneko, "A log-linearized Gassian mixture network and its application to EEG pattern classification," *IEEE Trans. on Systems, Man, and Cybernetics-Part C: Application and Reviews*, Vol. 29, No. 1, pp. 60-72, 1999.

[8] C. Bishop, "Neural Networks for Pattern Recognition," New York: Oxford University Press, 1995.

[9] L.R. Rabiner, "A tutorial on hidden Markov model and selected application in speech recognition," *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.

[10] B-H. Juang, and S. Katagiri: "Discriminative Learning for Minimum Error Classification," *IEEE Trans. on Signal Processing*, Vol. 40, No. 12, pp. 3043-3054, 1992.