International Journal of Computational Intelligence and Applications
Vol. 7, No. 1 (2008) 1–15

Imperial College Press
www.icpress.co.uk

A SPEECH SYNTHESIZER USING FACIAL EMG SIGNALS

TOSHIO TSUJI
Department of Artificial Complex Systems Engineering
Hiroshima University, Higashi-Hiroshima 739-8527, Japan
tsuji@bsys.hiroshima-u.ac.jp

NAN BU
National Institute of Advanced Industrial Science and Technology
Tosu 841-0052, Japan

JUN ARITA
Department of Artificial Complex Systems Engineering
Hiroshima University, Higashi-Hiroshima 739-8527, Japan

MAKOTO OHGA
Eastern Hiroshima Prefecture Industrial Research Institute
Fukuyama 721-0974, Japan

Received 28 November 2006
Revised 27 April 2007
Accepted 4 May 2007

This paper proposes a novel phoneme classification method using facial electromyography (EMG) signals. This method makes use of differential EMG signals between muscles for phoneme classification, which enables a speech synthesizer to be constructed using fewer electrodes. The EMG signal is derived as a differential between monopolar electrodes attached to two different muscles, unlike conventional methods in which the EMG signal is derived as a differential between bipolar electrodes attached to the same muscle. Frequency-based feature patterns are then extracted using a filter bank, and the phonemes are classified using a probabilistic neural network, called a reduced-dimensional log-linearized Gaussian mixture network (RD-LLGMN). Since RD-LLGMN merges feature extraction and pattern classification processes into a single network structure, a lower-dimensional feature set that is consistent with classification purposes can be extracted; consequently, classification performance can be improved. Experimental results indicate that the proposed method with a fewer number of electrodes can achieve a considerably high classification accuracy.

Keywords: Facial EMG signals; speech recognition; probabilistic neural networks.

1. Introduction

As a side effect of laryngectomy or tracheostomy, patients sometimes suffer from a loss of phonation function. The rehabilitation of speech and voice is an important challenge to these patients since communication is a critical issue related to their medical care and social interactions. A number of voice rehabilitation methods have

been investigated over the years, and some artificial larynges and speaking valves have become commercially available.[1,2] However, these methods still have some disadvantages such as poor sound quality, the need of frequent maintenance, and inconvenience of use (especially in daily life).[2–4]

Many researchers have reported that the electromyography (EMG) signals from the body's facial and cervical muscles can be used for speech recognition.[5–16] To our knowledge, Sugie and Tsunoda[5] proposed the first EMG-based speech recognition system, in which three channels of EMG signals from the muscles around the mouth were used to discriminate five Japanese vowels with an automaton. Further, the prototype of a real-time speech synthesizer was constructed as a speech prosthesis for patients who have lost their phonation capabilities. Although the average classification rate achieved was 64%, it was proved that the EMG signals obtained from the facial muscles contain information useful for speech recognition.

Recently, significant improvements have been achieved in the EMG-based speech recognition. Chan *et al.* reported an improved classification performance for a 10-word vocabulary using five channels of EMG signals measured with 10 electrodes.[6,7] This method was further enhanced by combining recognition results based on acoustic and EMG signals with a multiexpert system.[8] In Ref. 9, five Japanese vowels were recognized using three EMG channels (six electrodes), and the classification accuracy exceeded 90% for all the subjects. Kumar *et al.*[10] employed a back-propagation neural network (NN) to classify five English vowels using three EMG channels (six electrodes), an average classification accuracy of 88% was reported. Further, Maier-Hein *et al.*[11] confirmed an average classification accuracy of 97.3% for ten English digits using seven EMG channels (14 electrodes). Based on the method of Maier-Hein *et al.*, investigations involving large vocabularies were conducted.[12,13]

In particular, Fukuda *et al.* proposed an EMG-based speech synthesizer system in which six Japanese phonemes (five vowels, namely, /*a*/, /*i*/, /*u*/, /*e*/, /*o*/, and one nasal /*n*/) are classified from five EMG channels (10 electrodes) using a probabilistic NN; then, words are recognized from the series of phonemes using algorithms of the hidden Markov model (HMM).[14] Due to the probabilistic NN and HMM algorithm, this system provides high-accuracy phoneme classification and word recognition, and it is robust against issues such as the differences among individuals and variations in temporal characteristics.

However, these studies used the differential EMG signals obtained from bipolar electrodes; consequently, the number of electrodes was fairly large. This is undesirable in practical applications and from the viewpoint of being less noticeable. Jorgensen *et al.*[15] developed a system using two pairs of electrodes, and the authors indicated that as few as one electrode pair located diagonally between the cleft of the chin and the larynx would suffice for recognizing a small vocabulary.[16] However, since the cervical muscles around the larynx are used, these methods may not be applicable for patients after laryngectomy or tracheostomy.

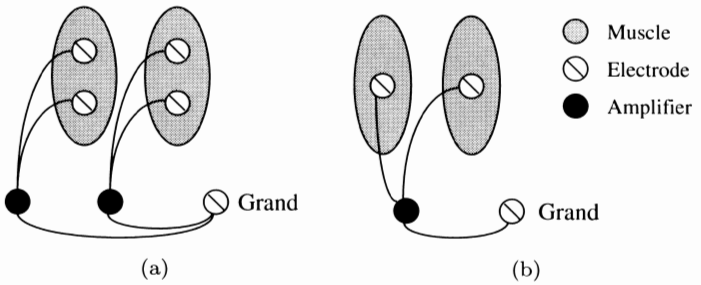

Fig. 1. Electrode configurations: (a) differential EMG signals measured from the same muscles (bipolar), (b) differential EMG signal measured from two muscles (monopolar).

In order to achieve acceptable classification accuracy by using fewer electrodes, one is required to conform to the following criteria:

(1) sufficient feature characteristics should be extracted for speech recognition while reducing the number of electrodes, and
(2) an effective pattern classification tool needs to be provided.

To tackle these problems, the present paper proposes a novel method for EMG-based speech recognition. This method consists of an EMG acquisition method based on differential EMG signals between muscles and a classifier using a probabilistic NN. Unlike conventional bipolar recording configurations, the electrodes are attached to different muscles, one electrode for each muscle. Subsequently, differential signals between every two electrodes can be derived as input channels for classification (see Fig. 1).[17] This method can reduce the number of electrodes.[18] In order to acquire sufficient feature characteristics from the reduced EMG sources, the frequency content of each channel is extracted using a filter bank.

The dimensionality of the feature space should grow with an increasing frequency resolution. The proposed method incorporates a novel probabilistic NN, called a reduced-dimensional log-linearized Gaussian mixture network (RD-LLGMN),[19] for the classification of high-dimensional EMG patterns. Two basic concepts of this probabilistic NN are (1) an orthogonal transformation that projects the original input space into a lower-dimensional space and (2) the Gaussian mixture model (GMM) that estimates the probability distribution of patterns in the projected lower-dimensional space. This network combines the feature extraction process with the classification part, and is trained in the manner of minimum classification error (MCE) learning,[20] which enables the classification part to realize a low error probability. The proposed EMG-based speech recognition method is expected to extract discriminative information from frequency-based EMG patterns and enable an efficient classification of phonemes using fewer electrodes.

This paper is organized as follows. Section 2 explains the details of the proposed method. In Sec. 3, the performance of the proposed method is evaluated

with experimental results of healthy subjects and a laryngectomized patient. Comparison experiments are then presented in Sec. 4. Finally, Sec. 5 concludes this paper.

2. Methods

2.1. *EMG signal acquisition and feature extraction*

By using a monopolar configuration, the EMG signals are measured as shown in Fig. 1(b). The differential between the 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 facial muscles; one electrode is attached to one muscle. The EMG signals are recorded at a sampling frequency of 1 kHz. The difference between the potentials of every two electrodes is computed; as a result, 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.

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

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

where $\sigma = U/Z$. U indicates the frequency range under consideration, and is set as 250 Hz in this study. 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 $\mathrm{EMG}_i(t)$ ($i = 1, \ldots, d$), and normalized to make the sum of d channels equal to 1.

$$x_i(t) = \frac{\mathrm{EMG}_i(t) - \overline{\mathrm{EMG}}_i^{st}}{\sum_{i=1}^{d} \mathrm{EMG}_i(t) - \overline{\mathrm{EMG}}_i^{st}} \quad (i = 1, \ldots, d), \tag{2}$$

where $\overline{\mathrm{EMG}}_i^{st}$ is the mean value of $\mathrm{EMG}_i(t)$, which is measured when the muscles are relaxing. Then, the normalized patterns, $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_d(t)]^{\mathrm{T}}$, are used as the input features for phoneme classification.

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

$$F_{\mathrm{EMG}}(t) = \frac{1}{S} \sum_{s=1}^{S} \frac{\mathrm{MEMG}_s(t) - \overline{\mathrm{MEMG}}_s^{st}}{\mathrm{MEMG}_s^{\max} - \overline{\mathrm{MEMG}}_s^{st}}, \tag{3}$$

where $\mathrm{MEMG}_s(t)$ indicates the filtered signal (cut-off frequency: 1 Hz) of rectified raw EMG directly measured from the electrode s ($s = 1, \ldots, S$), $\overline{\mathrm{MEMG}}_s^{st}$ is the mean value of $\mathrm{MEMG}_s(t)$, which is measured when the muscles are relaxing, and

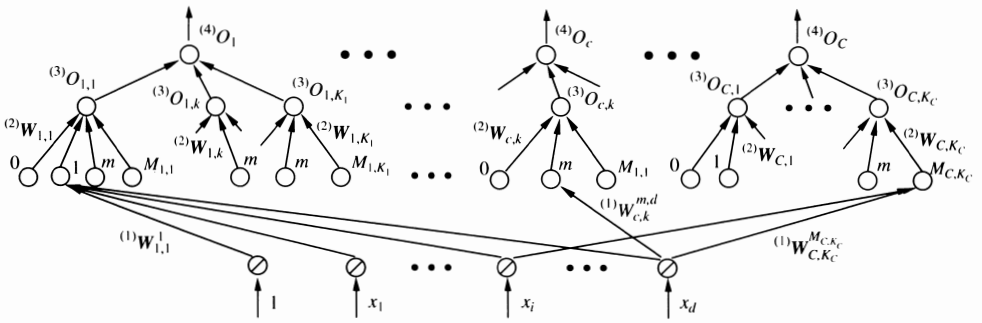

Fig. 2. Structure of RD-LLGMN.

$\text{MEMG}_s^{\max}$ is the mean value of $\text{MEMG}_s(t)$ measured under the maximum voluntary contraction (MVC). $F_{\text{EMG}}(t)$ indicates the force information, and it is used to recognize whether the motion has really occurred or not by comparing $F_{\text{EMG}}(t)$ with a predefined threshold M_d.

2.2. *Phoneme classification*

RD-LLGMN[19] is used for phoneme classification. It provides a novel approach for feature reduction by finding the discriminant features of a reduced size and calculates the posterior probabilities for classification.

RD-LLGMN is a four-layer NN; its structure is shown in Fig. 2. 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 the activation of each unit. Let $^{(1)}O_i$ $(i = 0, \ldots, d)$ denote the output of the ith unit in the first layer, we have

$$^{(1)}O_i = \begin{cases} 1, & i = 0, \\ x_i, & i = 1, 2, \ldots, d, \end{cases} \tag{4}$$

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

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

$$^{(2)}I^m_{c,k} = \sum_{i=1}^{d} {}^{(1)}O_i \, {}^{(1)}W^{m,i}_{c,k}, \tag{5}$$

$$^{(2)}O^m_{c,k} = (^{(2)}I^m_{c,k})^2. \tag{6}$$

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

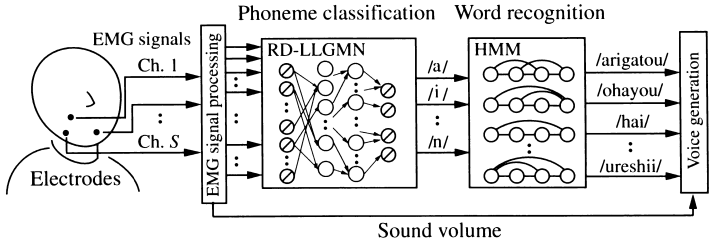

Fig. 3. Schematic view of the proposed speech synthesizer system.

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

$$ {}^{(3)}I_{c,k} = \sum_{m=0}^{M_{c,k}} {}^{(2)}O^m_{c,k}\, {}^{(2)}W^m_{c,k}, \tag{7} $$

$$ {}^{(3)}O_{c,k} = \frac{\exp[{}^{(3)}I_{c,k}]}{\sum_{c'=1}^{C}\sum_{k'=1}^{K_{c'}} \exp[{}^{(3)}I_{c',k'}]}. \tag{8} $$

In this layer, RD-LLGMN calculates the 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 the K_c components $\{c, k\}$ in the third layer. The function between the input and the output is described as

$$ {}^{(4)}O_c = {}^{(4)}I_c = \sum_{k=1}^{K_c} {}^{(3)}O_{c,k}. \tag{9} $$

Only after the weight coefficients are optimized using an MCE-based training algorithm, the output of RD-LLGMN, ${}^{(4)}O_c$, can estimate the posterior probability of class c.

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

$$ H(t) = -\sum_{c=1}^{C} {}^{(4)}O_c(t) \log {}^{(4)}O_c(t). \tag{10} $$

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; otherwise, the determination of the motion is suspended.

2.3. *Speech synthesizer*

Based on the proposed method, a speech synthesizer system is constructed, as shown in Fig. 3. According to the results of phoneme classification, words are recognized

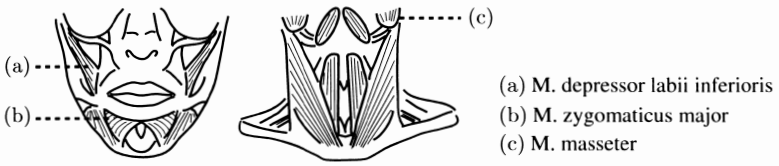

Fig. 4. Location of the target muscles.

using the HMM-based method proposed by Fukuda *et al.*[14] Finally, voice generation is achieved using synthesizer software. In this system, the sound volume is controlled according to the force information calculated from the EMG signals.

According to the previous studies, it is very difficult to recognize consonant only from EMG signals.[14] In the proposed system, five vowels, namely, /*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[21] is applied for Japanese word recognition, which has been successfully developed especially in the field of speech recognition. For word recognition, one HMM is prepared for each word, for instance, /*oaou*/ for /*ohayou*/ and /*aeu*/ for /*taberu*/. When users utter /*ohayou*/, the corresponding model /*oaou*/, which consists of the sequence of vowels belonging to the word, is recognized. Further, the utterance lengths vary remarkably. Since HMMs approximate the probabilistic characteristics of time series through learning, stable recognition can be achieved for words with varying temporal characteristics.

3. Evaluation

Japanese phoneme classification experiments were conducted to examine performance of the proposed method. Five subjects (A–D: healthy, E: a laryngectomized patient) participated in these experiments. The subjects were asked to utter six phonemes ($C = 6$) for approximately 30 s in the order of /*a*/, /*i*/, /*u*/, /*e*/, /*o*/, and /*n*/. From the 11 trials conducted, the training trial was randomly selected, and the other 10 trials were used for testing purposes.

Only three Ag/AgCl electrodes (SEB120, GE Marquette Corp.) are attached to the subject's facial muscles (M. Depressor Labii Inferioris (PLI), M. Zygomaticus Major (ZM), and M. Masster (MA); see Fig. 4). The differential between DLI and ZM was used as input channel one, differential between DLI and MA as channel two, and differential between ZM and MA as channel three. The number of band-pass filters Z was six, therefore the dimension of the input features for RD-LLGMN d was 18. The parameters of the GMM in RD-LLGMN were set as: $C = 6$, $K_c = 1$ $(c = 1, \ldots, 6)$. The dimensions of the reduced subspaces $M_{c,k}$ $(c = 1, \ldots, C; k = 1)$ were set as $M = 9$. In the training phase, 50 EMG patterns were extracted from

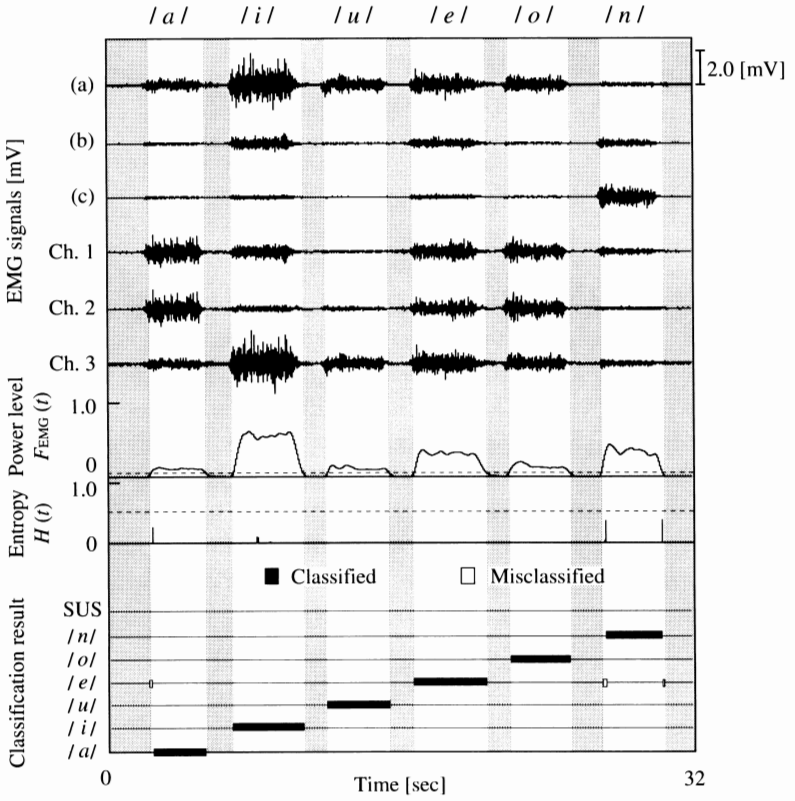

Fig. 5. Examples of the classification results (subject A) [(a): DLI, (b): ZM, (c): MA].

the 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 $H_d = 0.5$.

Figure 5 depicts an example of subject A's classification results of the *best* trial, which provides the best classification rates among all the test trials. In this figure, three channels of the raw monopolar EMG signals, three channels of the differential EMG signals, the force information $F_{\mathrm{EMG}}(t)$, the entropy $H(t)$, and the classification results have been plotted. The gray areas indicate that no utterance occurred because the force information F_{EMG} was less than M_d. Although misclassification can be observed in the beginning of the utterance of /a/, and the beginning and ending of the utterance of /n/, the classification result of RD-LLGMN is relatively stable, and a high classification rate of 98.8% was realized in this experiment.

Further, the phoneme classification experiments were conducted with a laryngectomized patient (subject E). Figure 6 shows an example of the classification results of the *best* test trial. Misclassification is found in the utterance of /o/ and the beginning and ending of the utterance of /u/. In this experiment, the classification

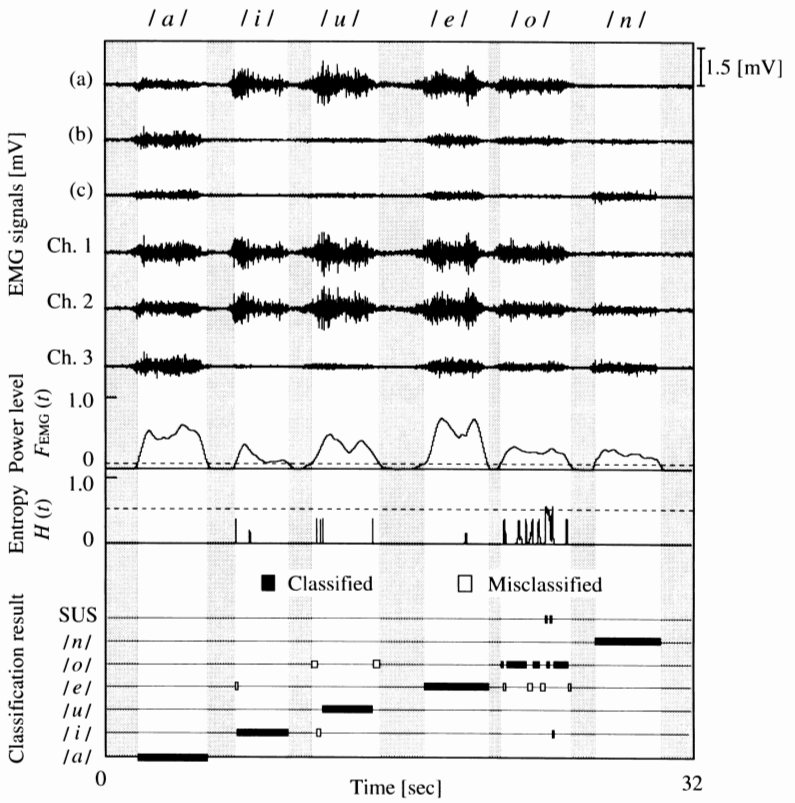

Fig. 6. Examples of the classification results (subject E) [(a): DLI, (b): ZM, (c): MA].

rate was 90.6%. As compared to the healthy subjects, the amplitude of the EMG signals measured from the muscles is low. Since the patient eats soft meals everyday, it is considered that the muscles around the jawbone, such as the muscle of masseter, are degraded. Moreover, for each misclassified utterance, the entropy is high. Misclassification could be reduced using an appropriately modulated threshold H_d.

4. Comparison Experiments

To verify the proposed method, comparison experiments were conducted. A neural classifier, log-linearized Gaussian mixture network (LLGMN),[22] was applied for phoneme classification. LLGMN is a probabilistic NN, which estimates the posterior probability distribution of the input features based on a GMM and a log-linear model. LLGMN has been used in the previous research.[14] For the details of LLGMN, please refer to the literature. In addition, a method based on feature extraction with LLGMN was utilized. A feature extraction process, principle

component analysis (PCA),[23] was used to reduce the dimensionality of the input features. After the PCA process, LLGMN was applied for phoneme classification. For simplicity, two methods used in the comparison experiments are referred to as LLGMN and LLGMN with PCA hereafter.

4.1. *Variation in classification performance with various conditions*

First, the phoneme classification results obtained using RD-LLGMN and LLGMN with PCA under various conditions are presented here. $\mathbf{x} \in \Re^d$ defined in Eq. (2) was used as the input signal. In the PCA part, the original features are projected into a lower-dimensional space on directions that correspond to the M highest eigenvalues of the covariance matrix.[23] Feature vectors extracted with these M directions are then fed into LLGMN.

LLGMN[22] is a three-layer feedforward probabilistic NN based on GMM. The number of units in the input layer of LLGMN was set equal to M. The units in the hidden layer correspond to the Gaussian components in GMM, the number of which was set as one. The output layer had six units, and each unit outputs the corresponding posterior probability for the input pattern. The same determination thresholds, M_d and H_d, were used for the classification based on LLGMN with PCA. LLGMN was trained with a maximum likelihood learning.[22]

In the comparison experiments, the classification rates of two methods are evaluated by varying the dimensionality of input EMG features d and an extraction rate (denoted as β), which is the ratio of M to d. The dimensionality of the input EMG features d is changed by changing the number of filters Z from one to six. Five sets of randomly selected initial weights were used to train each classifier. Figures 7 and 8 show the mean values and standard deviations of the classification results for the *best* test trial (subject A) for different parameter combinations, namely,

$$d \times \beta : \begin{cases} \beta \in \left[\frac{1}{3}, \frac{2}{3}, 1\right], & (d \in [3, 9, 15]), \\ \beta \in \left[\frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, 1\right], & (d \in [6, 12, 18]). \end{cases} \tag{11}$$

It should be noted that the directions of the axes of d and β are reversed in the figures showing standard deviations for improving their clarity. From these figures, it can be observed that RD-LLGMN achieved higher classification rates than LLGMN with PCA. Since PCA and LLGMN are separately optimized based on different training criteria, the extracted features may not always be consistent with the purpose of classification, and their classification accuracy was poorer than that of RD-LLGMN. Further, we can observe that when β increases, the classification rates of the two methods increase slightly. This is due to the fact that more information is used for pattern classification. However, computational complexity and time used

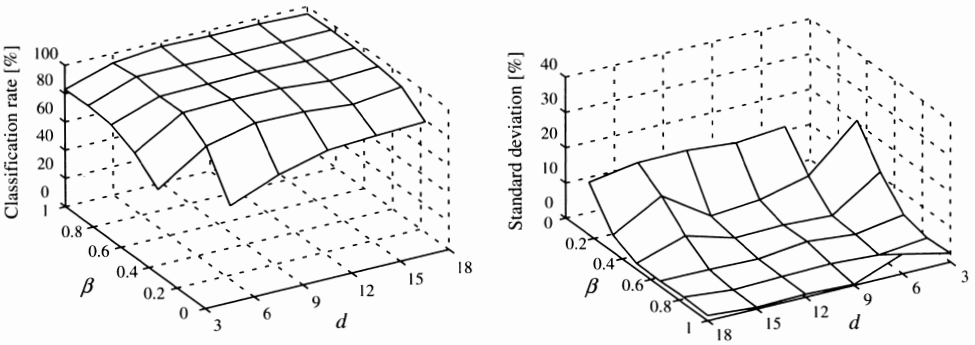

Fig. 7. Classification results using RD-LLGMN (subject A).

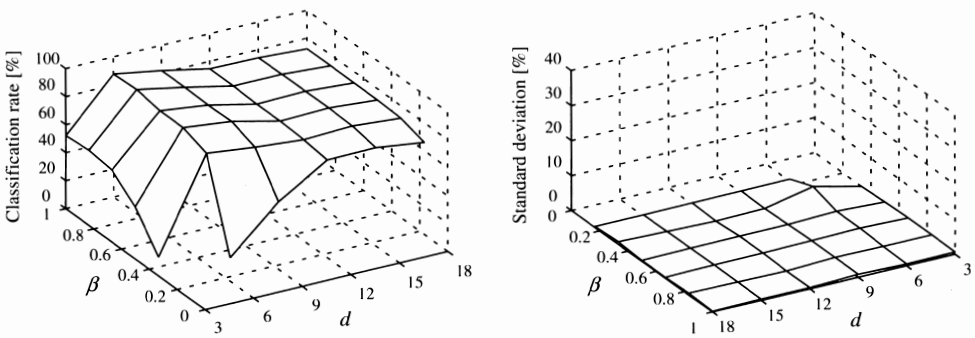

Fig. 8. Classification results using LLGMN with PCA (subject A).

for training are significantly increased. On the other hand, when we increase d, a similar trend can be observed for RD-LLGMN. In contrast, classification rates of LLGMN with PCA method decrease for approximately 15%. When increasing d, the entropy of LLGMN's output increased. This implies that the classification became more ambiguous; therefore, the classification results were suspended, resulting in a decrease in the classification rates of LLGMN with PCA.

4.2. *Comparison between methods using differential EMG signals between muscles*

Comparison experiments were then conducted with LLGMN, LLGMN with PCA, and RD-LLGMN using differential EMG signals between muscles. For LLGMN, three channels of differential EMG signals between the muscles were rectified and filtered by a second-order Butterworth filter (cutoff frequency: 1 Hz). Note that the number of electrodes used for EMG acquisition is three. The filtered EMG signals

Table 1. Results of comparison between three methods using differential EMG signals between the muscles. Only three electrodes were used for EMG acquisition.

	Methods		
	LLGMN	LLGMN with PCA	RD-LLGMN
Subject A	68.6 ± 8.0	65.6 ± 6.5	$\mathbf{84.0 \pm 8.9}$
Subject B	32.6 ± 8.5	37.2 ± 4.4	$\mathbf{72.3 \pm 7.4}$
Subject C	80.4 ± 9.3	72.0 ± 6.9	$\mathbf{83.7 \pm 7.9}$
Subject D	30.3 ± 3.8	15.7 ± 2.9	$\mathbf{60.6 \pm 6.1}$
Subject E	29.5 ± 8.1	60.7 ± 7.7	$\mathbf{80.4 \pm 8.9}$

Mean $\pm$ S.D. [%].

are defined as $\mathrm{CEMG}_l(t)$ $(l = 1, 2, 3)$ and are normalized to make the sum of the three channels equal to 1.0:

$$x'_l(t) = \frac{\mathrm{CEMG}_l(t) - \overline{\mathrm{CEMG}}_l^{st}}{\sum_{l'=1}^{3} \mathrm{CEMG}_{l'}(t) - \overline{\mathrm{CEMG}}_{l'}^{st}} \quad (l = 1, 2, 3), \tag{12}$$

where $\overline{\mathrm{CEMG}}_l^{st}$ is the mean value of $\mathrm{CEMG}_l(t)$ measured while the muscles are relaxing. LLGMN used the feature vector $\mathbf{x}' \in \Re^3$ as the input. In case of LLGMN with PCA, the bank of six filters was applied to the three channels of differential EMG signals in the same way as described in Sec. 2.1. Then, PCA was used to reduce the dimensionality to nine. The number of units in the input layer of LLGMN was equal to the dimension of the input vector.

For all the methods, five sets of randomly selected initial weights were used for training. Ten test trials were conducted; in each test trial, the EMG signals were measured for approximately 30 s (six phonemes). Table 1 shows the mean values and standard deviations of the classification rates of the 10 test trials using the three methods. The ranges of the classification rates (%) with the proposed method are [65.9, 98.8] (subject A), [55.9, 86.1] (subject B), [59.8, 96.5] (subject C), [51.3, 76.0] (subject D), and [60.6, 90.6] (subject E). The examples shown in Figs. 5 and 6 are the best classification results of subjects A and E. It is evident that the proposed method outperformed the other methods.

5. Conclusion

This paper proposes a novel phoneme classification method for speech synthesizer using facial EMG signals. This method uses differential EMG signals between muscles, and classification can be achieved based on fewer electrodes. In order to acquire sufficient feature characteristics from the reduced EMG sources, a filter bank is used to extract the frequency information. Employing 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 using three electrodes for

EMG measurements have been carried out with five subjects. In the experiments, relatively high classification rates of the proposed method using a small number of electrodes were confirmed. Furthermore, with the same number of electrodes, the proposed method outperforms the other methods.

In this paper, RD-LLGMN is used for phoneme classification. For EMG-based speech recognition, many researchers have applied NNs, such as multi-layer perceptron (MLP) and radial-basis function (RBF) networks, for classification purposes.[9,10,15,16] However, with these traditional neural classifiers, training process becomes complicated when dealing with large-dimensional data. A large training dataset is always required to estimate the parameters in NNs. Usually, feature extraction is conducted prior to a classification process in order to find a compact feature set to avoid exhaustive computation. Unfortunately, the classification schemes based on a feature extractor with a classifier suffer from some intrinsic limitations: the feature extractor and the classifier are separately optimized; moreover, they usually have different training criteria.[19] In contrast, RD-LLGMN merges the feature extraction and pattern classification processes into a single network structure, and the parameters are modulated with a criterion to minimize the error probability. It is expected that RD-LLGMN would yield better classification performance. In this paper, comparison results between the proposed method and the LLGMN-based methods have been provided. A further investigation to compare MLP- and RBF-based classification methods with the proposed method is envisaged.

In the future research, we would like to improve the pre-processing method of EMG signals, such as the modulation of the parameters of the filter banks and low-pass filtering. Further, the locations of electrodes and the selection of monopolar channels should be investigated. From Fig. 7, it can be observed that the standard deviations of the proposed method are larger than those of LLGMN with PCA when β is small. A detailed investigation is required to evaluate the stability of the classification results of the proposed method. Further work would be needed to compare the proposed method with other previously proposed EMG-based speech recognition methods.

Acknowledgements

The authors would like to acknowledge Mr. Hiromi Koseki and Dr. Osamu Fukuda for their kind and encouraging support. Also, this work is partially supported by the 21st Century COE Program of JSPS (Japan Society for the Promotion of Science) on *Hyper Human Technology Toward the 21st Century Industrial Revolution.*

References

1. A. H. Shikani, J. French and A. A. Siebens, New unidirectional airflow ball tracheostomy speaking valve, *Otolaryngol. Head Neck Surg.* **123** (2000) 103–107.

2. P. Jassar, R. J. England and N. D. Stafford, Restoration of voice after laryngectomy, *J. R. Soc. Med.* **92** (1999) 299–302.
3. T. Most, Y. Tobin and R. C. Mimran, Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production, *J. Commun. Disord.* **33** (2000) 165–181.
4. D. H. Brown, F. J. M. Hilgers, J. C. Irish and A. J. M. Balm, Postlaryngectomy voice rehabilitation: State of the art at the millennium, *World J. Surg.* **27** (2003) 824–831.
5. N. Sugie and K. Tsunoda, A speech prosthesis employing a speech synthesizer — Vowel discrimination from perioral muscle activities and vowel production, *IEEE Trans. Biomed. Eng.* **32** (1985) 485–490.
6. A. D. C. Chan, K. Englehart, B. Hudgins and D. F. Lovely, Myo-electric signals to augment speech recognition, *Med. Biol. Eng. Comput.* **39** (2001) 500–504.
7. A. D. C. Chan, K. Englehart, B. Hudgins and D. F. Lovely, Hidden Markov model classification of myoelectric signals in speech, *IEEE Eng. Med. Biol. Mag.* **21**(5) (2002) 143–146.
8. A. D. C. Chan, K. B. Englehart, B. Hudgins and D. F. Lovely, Multiexpert automatic speech recognition using acoustic and myoelectric signals, *IEEE Trans. Biomed. Eng.* **53**(4)(2006) 676–685.
9. H. Manabe, A. Hiraiwa and T. Sugimura, Unvoiced speech recognition using EMG — Mime speech recognition, *Proc. ACM Conf. Human Factors Comput. Syst.* (2003), pp. 794–795.
10. S. Kumar, D. K. Kumar, M. Alemu and M. Burry, EMG based voice recognition, *Proc. 2004 Int. Conf. Intell. Sensors, Sensor Networks Inform. Process* (2004), pp. 593–597.
11. L. Maier-Hein, F. Metze, T. Schultz and A. Waibel, Session independent non-audible speech recognition using surface electromyography, in *Proc. 9th IEEE Workshop Automat. Speech Recog. Understand.* (2005), pp. 331–336.
12. M. Walliczek, F. Kraft, S.-C. Jou, T. Schultz and A. Waibel, Sub-word unit based non-audible speech recognition using surface electromyography, *Proc. 9th Int. Conf. Spoken Lang. Process.* (2006), pp. 1487–1490.
13. S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft and A. Waibel, Towards continuous speech recognition using surface electromyography, *Proc. 9th Int. Conf. Spoken Lang. Process.* (2006), pp. 573–576.
14. O. Fukuda, S. Fujita and T. Tsuji, Substitute vocalization system based on EMG signals, *IEICE Trans. Inform. Syst.* **J86-D-II** (2003) 1–7 (in Japanese).
15. C. Jorgensen, D. D. Lee and S. Agabon, Sub-auditory speech recognition based on EMG signals, *Proc. 2003 Int. Joint Conf. Neural Networks* (2003), pp. 3128–3133.
16. B. J. Betts, K. Binsted and C. Jorgensen, Small-vocabulary speech recognition using surface electromyography, *Interact. Comput.* **18** (2006) 1242–1259.
17. 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, *J. Robot. Mechatron.* **14** (2002) 565–571.
18. N. Bu, T. Tsuji, J. Arita and M. Ohga, Phoneme classification for speech synthesiser using differential EMG signals between muscles, *Proc. 27th Int. Conf. IEEE Eng. Med. Biol. Soc.* (2005), pp. 5962–5966.
19. N. Bu and T. Tsuji, Multivariate pattern classification based on local discriminant component analisys, *Proc. IEEE Int. Conf. Robot. Biomimet.* (2004), pp. 924–929.
20. B.-H. Juang and S. Katagiri, Discriminative learning for minimum error classification, *IEEE Trans. Signal Process.* **40** (1992) 3043–3054.
21. L. R. Rabiner, A tutorial on hidden Markov model and selected application in speech recognition, *Proc. IEEE* **77** (1989) 257–286.

22. T. Tsuji, O. Fukuda, H. Ichinobe and M. Kaneko, A log-linearized Gaussian mixture network and its application to EEG pattern classification, *IEEE Trans. Syst. Man Cybern. Appl. Rev., Part C* **29** (1999) 60–72.
23. C. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, New York, 1995).