J. Biomed. Eng. Res : 101-109, 2006

# HMM-Based Automatic Speech Recognition using EMG Signal

Ki-Seung Lee

*Department of Electronic Engineering, Konkuk University*
*(Received April 10, 2006. Accepted May 22, 2006)*

## Abstract

It has been known that there is strong relationship between human voices and the movements of the articulatory facial muscles. In this paper, we utilize this knowledge to implement an automatic speech recognition scheme which uses solely surface electromyogram (EMG) signals. The EMG signals were acquired from three articulatory facial muscles. Preliminary, 10 Korean digits were used as recognition variables. The various feature parameters including filter bank outputs, linear predictive coefficients and cepstrum coefficients were evaluated to find the appropriate parameters for EMG-based speech recognition. The sequence of the EMG signals for each word is modelled by a hidden Markov model (HMM) framework. A continuous word recognition approach was investigated in this work. Hence, the model for each word is obtained by concatenating the subword models and the embedded re-estimation techniques were employed in the training stage. The findings indicate that such a system may have a capacity to recognize speech signals with an accuracy of up to 90%, in case when mel-filter bank output was used as the feature parameters for recognition.

Key words : surface EMG signals, automatic speech recognition, hidden markov model

## I. INTRODUCTION

Automatic speech recognition (ASR) is a technique that translates the incoming speech signals into their contextual information automatically. The existing ASR systems mainly depend on acoustic signal patterns which are less advantageous in the presence of ambient noise. The electromyogram (EMG) signals from the articulatory facial muscles can be considered as a secondary source of speech information [1] and can be used to design a new type of ASR system. Besides its noisy-robustness of EMG-based ASR system, it is useful for the persons who underwent laryngectomy caused by laryngeal cancer or accidents [1, 2]. Underlying principle is that different phonemes are produced by different vocal articulations, and hence, we could classify phonemes and ultimately words by using EMG signals. This kind of ASR system has been discussed in recent literatures, where EMG signals were solely used in recognition procedure [2-5] and where EMG signals were used as an auxiliary signal in addition to speech signals [6-8] .

The problems of EMG-based ASR system can be formulated as how to detect input signals (EMG signals) and how to build a relationship between contextual information and corresponding EMG signals. Determining adequate locations of the facial muscles from the standpoint of maximizing the recognition accuracy of the EMG-based ASR system is not trivial, although the role of each facial muscle is well defined. In the previous works, the locations of EMG electrodes were heuristically determined [2-4, 6-8]. The facial muscles employed in these works include mentalis, depressor anguli oris, masseter [4, 6], digastricus, zygomaticus major, orbiculas oris [3]. In [11], EMG signals were collected from the areas close to the larynx and throat. Sensing method for collecting EMG signals are also important issue. Since the usage of the invasive electrode is less comfortable for users, surface EMG signals were commonly used [3, 4, 6, 8].

The second problem associated with a EMG-based ASR system can be thought as building a mapping rule that maps a given sequence of EMG signals into a sequence of context words (or phonemes). Since it is not easy to model raw EMG signal itself, it is necessary to convert the raw EMG signals into the feature parameters and to define the models for the parameters prior to building a mapping rule. Several parameters have been employed in EMG-base ASR systems, including mel frequency cepstral coefficient (MFCC) [3], root mean square (RMS) [4], auto-regressive (AR) coefficient [6],

Corresponding Author : Ki-Seung Lee
Deptartment of Electronic Engineering, Konkuk University,
1 Hwayang-dong, Gwangjin-gu, Seoul, 143-701, Korea
TEL : 02-450-3489 / FAX : 02-3437-5235
E-mail : kseung@konkuk.ac.kr

Coiflet wavelet transformation (CWT) coefficient [8] and discrete wavelet transformation (DWT) coefficient [11]. It was reported that wavelet transformation yielded superior results to the others, but differences in recognition performance were not remarkable [6]. In the presented works, we will find the appropriate feature parameters that yield the highest recognition accuracy.

The model for EMG parameters is based on their statistical characteristics. It is well known that a sequence of speech signals can be modelled by quasi-stationary random process [12]. An well-known statistical tool for modelling such a kind of time-varying random process is hidden Markov model (HMM) [9]. Since both EMG signals and the corresponding speech signals are inherently originated from the same context, a sequence of EMG signals has been also modelled by HMM [3, 6, 8]. In the HMM-based ASR systems, recognition is performed by finding the template having the maximum likelihood to the observation sequence which corresponds to incoming EMG signals.

In case where the task focuses on the isolated word recognition problem, the system can be more easily implemented by using a simple model. An example of this approach was proposed by Kumar [4], where artificial neural network (ANN) was used to classify a given sequence of EMG signals into 5 five English vowels. In [7], linear discriminant analysis (LDA) accompanied by principle component analysis (PCA) was carried on the time-normalized AR coefficients to recognize an isolated digit number. However, these approaches have their inherent limitations associated with isolated word recognition, that is, only one word can be recognized at a time.

Although HMM can be suitable for describing inter-phoneme or inter-word dependencies [9], the previous EMG-based ASR systems involved with HMM mostly dealt with the isolated word recognition tasks. To date, a connected word recognition task which recognizes more than two words simultaneously, has not been discussed in the area of EMG-based ASR, in spite of its important role in real-word situation. The works proposed herein focus on the connected word recognition task, where each template word is modelled by a sequence of subwords which typically corresponds to a phoneme sequence. The proposed EMG-based ASR method was applied to connected digits recognition task. The validity of the proposed method is evaluated in the context of correct word recognition rate. The various experiments were also performed to find the optimal configuration of underlying HMM in the sense of maximizing the recognition accuracy. Several issues related with implementing the EMG-based ASR systems for connected word recognition were also discussed, including the estimation method of the connected subwords' models.

## II. METHODS

A block diagram of the proposed EMG-based ASR system is shown in Fig. 1. There are two stages; In the training stage, the EMG signals are first obtained, an analysis is performed on these EMG samples to derive the feature parameters

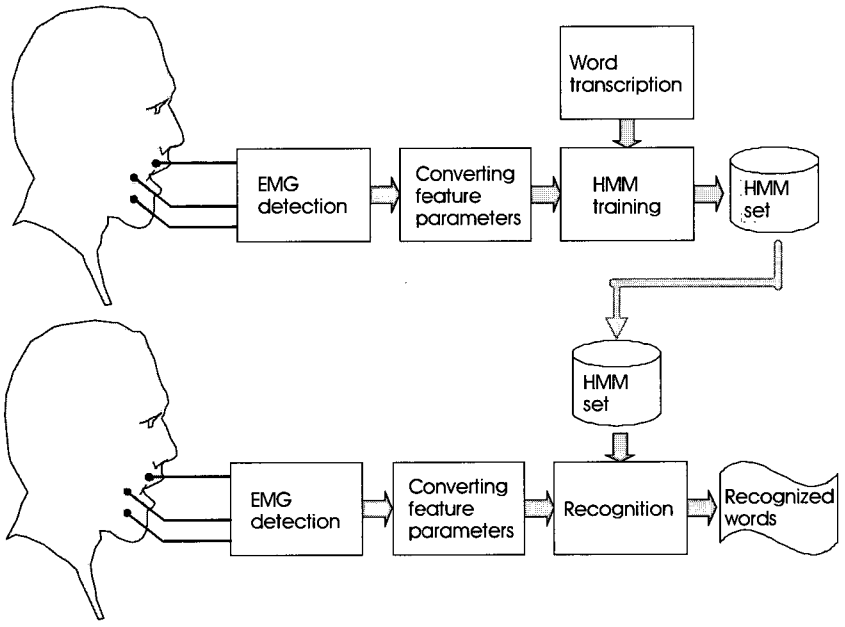


Fig. 1. Block digram of the proposed EMG-based ASR system. Top: Offline (training) procedure, Bottom: Online (recognition) procedure.

to be used as the input signals for ASR system. In this work, various feature parameters, which have been widely employed in speech-based ASR systems are involved in determining the appropriate feature parameters for EMG-based ASR system. The parameters describing HMM are estimated by using the features from the EMG signals and their orthographic word transcription.

In the online stage, the feature parameters which are identical to those in the training stage were derived from the incoming EMG signals. The feature parameters were then fed into the recognition procedure, which is based on a maximum likelihood (ML) estimation. Each part of the proposed system is described in more detail in the following sections.

### A. EMG Detection

Since there is no explicit relationship between the functions of each facial muscle and recognition performance, it is very difficult to find the optimal locations of the facial muscles in the sense of maximizing recognition accuracy in an analytical way. Hence, the locations of EMG electrodes were determined heuristically, based on a trial-and error approach. In this work, surface EMG signals were obtained from three articulatory muscles of the face: the levator anguli oris, the zygomaticus major, and the depressor anguli oris. The levator anguli oris originates from canine fossa immediately below the infraorbital of the maxilla (bone) and has a role in raising the skin tissue upwards from the corners of the mouth. The zygomaticus major originates from the zygomatic bone, and has a role in drawing the angle of the mouth upward and outward. The depressor anguli oris originates from the mandible and inserts skin at an angle of mouth and pulls corner of mouth downward. A detailed description of EMG detection will be appeared III.

### B. Feature Parameters

The features are extracted from the windowed EMG signals. A 100 msec length Hanning window was used to compute and extract the feature parameters at 20 msec intervals. The following features were taken into consideration for the candidate observation vectors for the subsequent HMM; log mel-filter bank channel outputs (FBANK), linear mel-frequency spectrum (MELSPEC), linear predictive coefficients (LPC), linear predictive reflection coefficients (LPREFC), mel-frequency cepstral coefficients (MFCC), LPC cepstral coefficients (LPCEPSTRA). FBANK and MELSPEC are given by output of mel scale filter bank. Mel scale is given by

$$Mel(f) = 2595\log_{10}(1+\frac{f}{700}) \tag{1}$$

where $f$ is frequency in Hz. Mel-scale filter bank analysis reflects the human auditory system [12]. Although EMG signals are not perceived by human ear, the usefulness of mel-scale filter bank analysis was confirmed in an EMG-based ASR system [3].

LPC is used for defining AR model of a given sequence. The AR model is given by

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + e(n) \tag{2}$$

where $a_i$ is the $i$-th LPC and e(n) is uncorrelated white noise. LPC $a_i$ is obtained by minimizing the mean square error of e(n). LPC is widely used in most speech application including speech coding, speech recognition, and speech enhancement [10, 12]. LPC was also employed in the EMG-based ASR systems [6, 11].

MFCC is given by the discrete cosine transform (DCT) coefficients of the log filter bank amplitude.

$$c_i = \sqrt{\frac{2}{N}}\sum_{j=1}^{N} m_j \cos(\frac{\pi i}{N}(j-0.5)) \tag{3}$$

where $c_i$ is the $i$-th MFCC and $N$ is the number of filter banks. Note that $m_j$ is the output of the $j$-th filter bank amplitude, obtained by the filter bank (1).

The LPCESTRA is defined as the natural logarithm of the LPC all-pole filter function. Let us consider a power series expansion for the logarithm function of the LPC all-pole filter.

$$\ln[1/(1+\sum_{i=1}^{p} a_i z^{-i})] \cong \sum_{n=1}^{N_C} c_n z^{-n} \tag{4}$$

where $c_n$ is the n-th LPC cepstral coefficient, $N_C$ is the order of LPC cepstral coefficients and z is a complex number. Note that $a_i$ is the $i$-th LPC coefficient described in (2). Hence, the LPC cepstral coefficients can be obtained from the LPC coefficients. A more detailed description for each parameter including LPREFC would be found in [12].

For most ASR systems, other useful features include signal power in dB, which replaces the 0-th order LPC cepstral coefficient, and the first and second temporal derivatives of the feature vector called the delta and delta-delta coefficients, respectively. These derivatives help give an estimate of the temporal variations in the signal. These features can be applied to the underlying EMG-based ASR system. However, the experimental results showed that the performance improvements

were not as high as in the case of the speech-based ASR systems. Accordingly, in our approach, delta and delta-delta coefficients were not employed in HMM estimation and recognition.

### C. Building HMM

HMMs represent a stochastic process that takes time series data as input, and outputs the probability that the data is generated by the model. HMMs have been successfully employed in most speech recognition tasks [9, 15]. An HMM is comprised of $N$-states, with each state $j$ containing an output probability distribution $b_j(\mathbf{o}_t)$, which determines the likelihood of generating observation $\mathbf{o}_t$ in state $j$ at time $t$. The probability of an observation moving from state $i$ to state $j$ is defined by the transition probability $a_{ij}$. The output distribution function of a state is represented as a multivariate Gaussian distribution or discrete distribution. The former is referred to as continuous density HMM and the latter is referred to as discrete density HMM. A typical HMM is shown in Fig. 2. The states are shown as circles, and the non-zero transition probabilities as arrows. Zero probability transitions are not shown.

Given T independent observations $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T\}$, a state sequence $\mathbf{q} = \{q_1, q_2, ..., q_T\}$, and the model $\lambda$ describing HMM, the probability that $\mathbf{O}$ and $\mathbf{q}$ occur simultaneously is given by

$$p(\mathbf{O}, \mathbf{q} \mid \lambda) = b_{q_1}(\mathbf{o}_1) a_{q_1 q_2} b_{q_2}(\mathbf{o}_2) ... a_{q_{T-1} q_T} b_{q_T}(\mathbf{o}_1) = \pi_{q_1} \prod_{t=1}^{T-1} a_{q_t q_{t+1}} b_{q_{t+1}}(\mathbf{o}_{t+1}) \tag{5}$$

where is $\pi_{q_1} = b_{q_1}(\mathbf{o}_1)$. If the number of states is $N$, there are $N^T$ such state sequences. The probability of $\mathbf{O}$ (given the model) is obtained by summing the joint probability (5) over all possible state sequences $\mathbf{q}$, giving

$$p(\mathbf{O} \mid \lambda) = \sum_{\text{all } \mathbf{q}} p(\mathbf{O}, \mathbf{q} \mid \lambda) \tag{6}$$

Now, the problem of HMM parameter estimation is then formulated as follows:

$$\lambda^* = \arg\max_{\lambda} p(\mathbf{O} \mid \lambda) \tag{7}$$

i.e., finding the HMM parameters in which the likelihood of the underlying model can be maximized, given the training data. This maximization problem can be solved by the Baum-Welch re-estimation formulas, based on the expectation -maximization (EM) algorithm [13].

In isolated word recognition, the HMM parameters are independently estimated for each word by using sufficient observation sequences for the words to be recognized. Whereas any word is represented by a series of subwords in connected word recognition. (A typical example of subwords is phoneme). Hence, building an individual word model can be accomplished by concatenating the models for each subwords. Accordingly, all the words to be recognized should have their orthographic phonetic transcriptions, which can be described in a pronunciation dictionary. The pronunciation dictionary employed in this work is shown in Table 1, which is composed of 10 Korean isolated digits. Note that phonemic symbols in the right column of the table followed the TIMIT or NIST format [17]. According to the table, it can be understood that the required subwords for the underlying EMG-based ASR system include {p, ah, L, oh, s, g, uh, iy, L, ch, yu, K, M, ng, sp, sil}, where "sp" (=short pause) is employed to separate two neighboring words, and "sil" (=silence) is employed to separate two neighboring sentences.

In connected word recognition, since all words are composed of the subwords' models, the HMMs are obtained by the subword-by-subword basis schemes. A straight forward way to achieve this is that subwords' HMMs are obtained from the labelled EMG stream, where a sequence of the EMG signals is segmented by its phonetic transcription, under assumption that each subword corresponds to each phoneme. In this case, HMM parameter estimation for a specific subword is performed on the observations from the EMG signals labelled by the same subword. However, hundreds of hours are required to label a large corpus for HMM training by hand [14]. Another drawback associated with hand labelling is that the results lack consistency because of the subjective decisions involved in the process. Thus, the hand labelling of a large amount of the EMG signals would be undesirable.

In this work, we trained individual subword model by using the embedded re-estimation technique [9]. An iterative algorithm was developed for designing a set of the subwords' HMMs having the maximum likelihood to the training EMG data. This iterative algorithm consists of segmenting the training EMG data using an a prior set of subwords' HMMs, and updating the set of subwords' HMMs using the labelled EMG sequence. This process is repeated until some convergence threshold is reached. Before iteration, initial subword boundaries should be provided. In this work, initial subword boundaries are obtained by uniform segmentation of the training data, in which each segment duration is given by entire duration of the training data divided by the total number of the subwords. A simple example of the embedded re-estimation procedure is shown in Fig. 3. This example illustrates how HMMs for subwords "ch", "iy" and "L" are

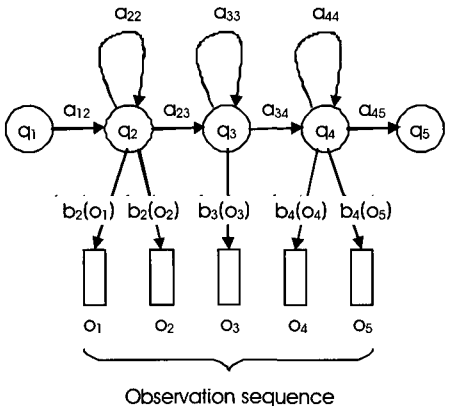


Fig. 2. A typical HMM.

estimated by using EMG sequence "ch+iy+L+iy+L+iy+ch+iy+L" (7-1-2-7-2 in Arabic numerals). In the segmentation stage, a forced alignment technique [15] was applied, that finds the subword boundaries yielding the maximum likelihood of the training sequence with respect to the corresponding phonetic transcription. The embedded re-estimation technique doesn't require hand-labelled EMG sequence. Hence, any hand-intervention is not necessary in constructing a set of subwords' HMMs.

Table 1. Pronunciation dictionary for 10 Korean digits.

| Word(in English) | Pronunciation(in Korean) |
|---|---|
| one | iy+L |
| two | iy |
| three | s+ah+M |
| four | s+오 |
| five | oh |
| six | yu+K |
| seven | ch+iy+L |
| eight | p+ah+L |
| nine | g+uh |
| zero | g+oh+ng |
| Special word | Description |
| sil | silence |
| sp | short pause |

### D. Recognition

In isolated word recognition, recognition is performed by finding the template having the maximum likelihood to the incoming observation sequence. In connected word recognition, a template is represented by a series of the subwords' models. Hence, recognition is performed by finding a word template that yields the maximum likelihood of the incoming observation sequence.

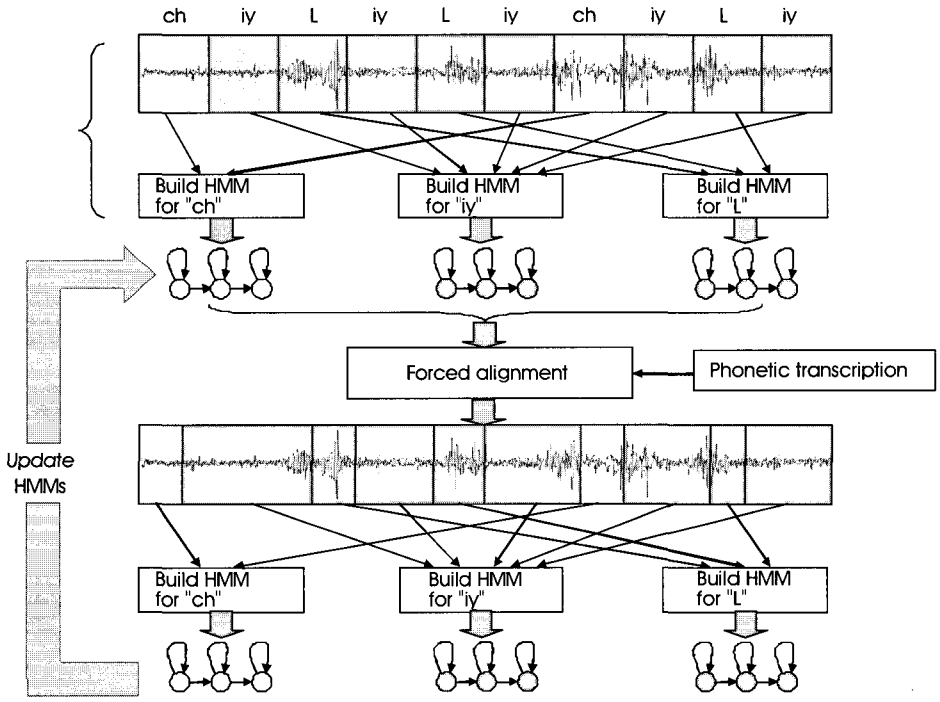


Fig. 3. An example of embedded re-estimation procedure for an EMG sequence "ch+iy+L+iy+L+iy+ch+iy+L+iy" (7-1-2-7-2). Trained HMMs for this example iuclude "ch", "iy" and "L".

$$\Lambda^* = \arg\max_{\Lambda} p(\mathbf{O} \mid \Lambda) \tag{8}$$

where $\Lambda$ denotes the word's HMM that is given by concatenating a number of subwords' HMMs, i.e.

$$\Lambda = \lambda_{sw_1} \oplus \lambda_{sw_2} \oplus \ldots \oplus \lambda_{sw_K} \tag{9}$$

where $\lambda_{sw_k}$ denotes the HMM for the $k$-th subword and $\oplus$ denotes concatenation. $K$ is the total number of the subwords in the word. Hence, a connected word recognition problem can be thought as finding a subword sequence, in which the overall likelihood of the underlying subword sequence is maximized. In this maximization problem, the number of possible subword sequences is limited by the number of the words to be recognized.

Since the likelihood $p(\mathbf{O} \mid \Lambda)$ is evaluated for all possible combinations of the $K$-subwords, maximization of Eq. (8) requires huge computations. To reduce the number of computations, Viterbi decoding [9] is often employed to solve the problem (8).

### E. Experimental Validation

Two Korean male subjects participated in this study. A ten-word vocabulary consisting of the words "zero [g+oh+ng]" to "nine [g+uh]" was used. For each subject, 25 series of 10 words, 4 series of 7 words and 2 series of 8 words were constructed. Hence, the total number of words employed in the recognition experiments is 588. The order of the words in each series was randomly permuted. Subjects were asked to speak each word in a consistent manner, minimizing the variation in volume and speaking style. The data set was split into the training corpus and the test corpus. Each corpus has 46 series and 16 series, respectively.

The location of each electrode for each muscle is shown in Fig. 4. Each EMG signal was collected using pairs of Ag-AgCl button electrodes (3M Red Dot, USA). Electrodes were 4.4 cm in diameter. The reference electrodes was located at the back of the neck. Before acquiring EMG signals, EMG target locations were cleaned with alcohol wet swabs.

A pre-amplifier with a Gain of 1000 was placed for each EMG channels, which was implemented by a high-precision instrumentation amplifier, AD620 (Analog Devices, USA). To minimize motion artifacts and aliasing, a band-pass filter with low corner (-3dB) 50Hz and with high corner (-3dB) frequency of 500Hz was implemented. To increase the robustness against power line noise, a notch filter with notch frequency of 60Hz shouldalso be involved. However, since it has been known that dominant energy of EMG lies in the 50-150 Hz range [16], a coarsely designed analog notch filter would affect the performance of the underlying ASR system. Hence, the notch filter was implemented in digital domain, with an advantage of high cut-off characteristics. The EMG signal was sampled at 1000Hz with 12 bits precision.

The number of mel-scale filter channels for the parameters FBANK, MELSPEC and MFCC was determined under the constraint that the highest center frequency of the filter bank is less than the Nyquist frequency (=500Hz). As for LP-based parameters, the order was set to 5. The order of the cepstral parameters including LPCEPSTRA and MFCC was also set to 5. For both cases, the order was heuristically determined.

In training HMM parameters, a typical left-to-right model was used. The distribution of acoustic features were modelled using mixtures of diagonal Gaussians. HTK [18] was used to build all phonemes' HMMs and to perform forced alignment.

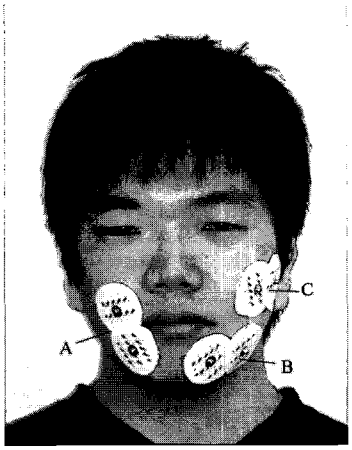


Fig. 4. The locations of electrodes for detecting the EMG signals of the articulatory facial muscles. A: levator anguli oris, B: depressor anguli oris, C: zygomaticus major.

## III. RESULTS

Fig. 5 is a plot the recognition rate as a function of the number of states for the various feature parameters. Recognition accuracy of MELSPEC is always higher than other cases, regardless of the number of states. The maximum recognition rate of MELSPEC is 90.43% in case when the number of states is 5. This result is somewhat similar to [6], where discrete wavelet transform (DWT) coefficients were employed. This can be interpreted by the fact that both FBANK and DTW coefficients are given by the outputs from the sub-band analysis filters.

Although it can be seen that the maximum accuracy was achieved when the number of states is 7 in some cases, no explicit consistency between recognition accuracy and the

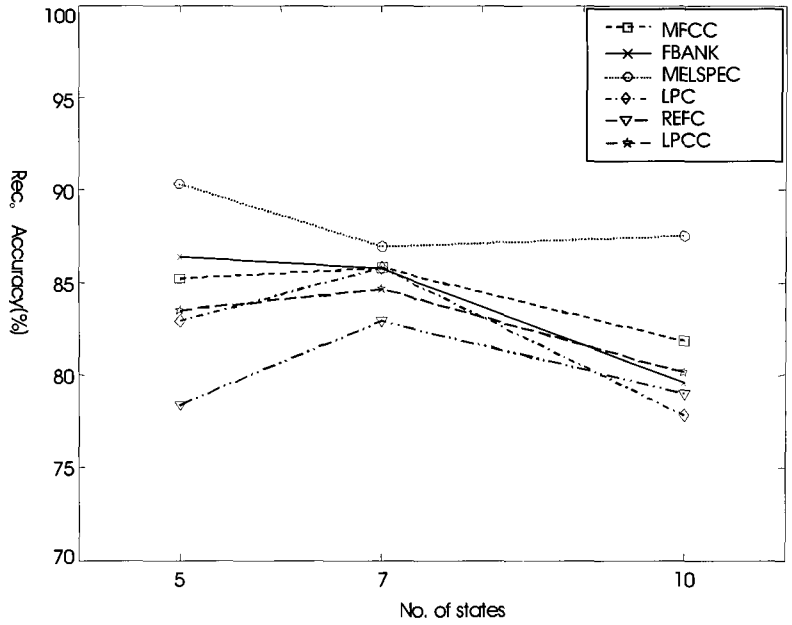


Fig. 5. Recognition performance for the various feature parameters, according to the number of states.

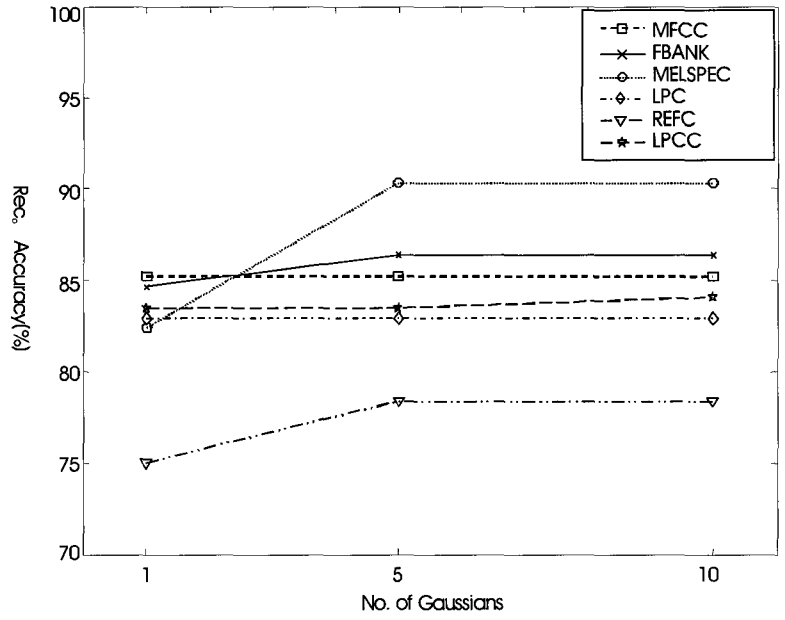


Fig. 6. Recognition performance for the various feature parameters, according to the number of Gaussians.

explicit consistency between recognition accuracy and the number of states was observed in Fig. 5. This is due to that the same number of states is assigned to every subwords, regardless of the characteristics of the underlying subword. It can be said that the number of states is closely related with the dynamic characteristics of the future parameters. Accordingly, it can be inferred that performance improvements can be achieved by assigning the different number of states to the different phoneme.

Fig. 6 is a plot the recognition rate as a function of the number of Gaussians employed to model the output probability of each state for the various feature parameters. The results shown in Fig. 6 were obtained when the number of states was 5. The results are similar to those of Fig. 5. MELSPEC yields the maximum recognition rate over the other parameters, except when the number of Gaussian is 1,

## IV. DISCUSSION

It is noteworthy that the parameters based an LP-analysis (including LPC, LPCEPSTRA and REFC) revealed the inferior performance comparing to the other parameters. A possible explanation for this bad performance of the LPC-based parameters is that AR model is not a good choice for EMG-production model.

It was known that the recognition accuracy of 95-93% was generally obtained, in case of speech-based connected digits recognition tasks. The recognition accuracy obtained herein is very close to that of the most speech-based recognition system. Although the EMG-based ASR system is comparable in terms of the global recognition performance, local recognition performances, which is measured by the individual recognition rate according to the word, are somewhat different each other. Accordingly, we employed the confusion matrix, which contains information about actual and predicted classification done by the recognizer to inspect the recognition performance for each word.

The confusion matrix from the proposed ASR system is shown in Table 2. The word "5 [oh]" is sometimes misclassified as "6 [y+uh+K]" . This is mainly due to the fact that the two words "5 [oh]" and "6 [y+uh+K]" are mainly distinguished by tongue position, not by lip shape. Unfortunately, major EMG signals associated with tongue movements were not involved in this work. Whereas differences in acoustic domain between these two words are not as much as in EMG-domain. Hence, it can be said that speech-based ASR has the superior performance to EMG-based ASR in this case.

It can be inferred that misclassification between "1 [iy+L]" and "2 [iy]" is also caused by tongue position. Another reason for this misclassification can be explained by the fact that the final consonant "L" has normally very short duration. Hence, very short duration of the phoneme "L" causes phoneme deletion error that leads to misclassification between "1 [iy+L]" and "2 [iy]". The words "3 [s+ah+M]" and "4 [s+ah]" are also distinguished by short duration of the final consonant "M". This may cause misclassification between these two words. In Table 2, however, error-free results were achieved for these two words. A possible reason for this result is that lip is completely closed when we finish pronouncing the word "3 [s+ah+M]", whereas lip is opened when we finish pronouncing the word "4 [s+ah]". Thus, the lip shapes are quite different each other in this case.In other words, these two words have acoustically similar characteristics. This leads to misclassification between these two words in speech-based ASR.

Differences in prosody are not so major factors influencing recognition performance. This was confirmed by the two facts. One is that although there were remarkable differences in speaking style between the two subjects (one subject uttered slowly and softly, the other uttered rapidly and loudly) recognition performance was uniform across the subjects (89.9% vs 90.9%). The other is that relationship between the utterance loudness which is measured by average RMS of speech waveforms and recognition accuracy was shown to be very small. This was measured by the chi-square value of a statistical significance test. A low chi-square value of 2.11 was obtained from the experiments which clearly accept the hypothesis that high/low volume speech signals are independent in terms of recognition accuracy of EMG-based ASR.

Table 2. Confusion matrix for the word recognition results.

| | | Recognized word | | | | | | | | | | | Rec. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | del | rate |
| Word Spoken | 1 | 10 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 83.3 |
| | 2 | 2 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 86.7 |
| | 3 | 0 | 0 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | 4 | 0 | 0 | 0 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| | 5 | 0 | 0 | 0 | 1 | 14 | 2 | 1 | 0 | 0 | 0 | 3 | 77.8 |
| | 6 | 0 | 1 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 90.9 |
| | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 1 | 100 |
| | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 0 | 0 | 0 | 100 |
| | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 13 | 0 | 0 | 92.9 |
| | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 1 | 90.9 |
| | Ins | 0 | 11 | 0 | 3 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | |

("del" means deletion errors and "Ins" means insertion errors)

## V. CONCLUSION

In this work, an EMG-based ASR scheme is proposed and the performance of the proposed scheme was evaluated. The major contribution of this work is applying a subword-based recognition approach to the continuous word recognition problem under the condition that EMG signals are solely employed. An embedded re-estimation scheme, which has been used in many speech-based continuous word recognition tasks, was employed to estimate the HMMs. The findings here show that a subword-based approach can be a solution for the connected word recognition problem of EMG-based ASR system. This was confirmed by the experimental results, yielding up to 90% recognition accuracy where recognition was performed on the series of digit numbers. Another contribution is that the linear mel-scale filter bank outputs produced the superior performance comparing to the other feature parameters.

The resultant recognition accuracy of the proposed EMG-based ASR system is comparable with that of the speech-based ASR systems. However, the task was limited to recognizing series of digit numbers. To increase the usefulness of the EMG-based ASR system, practical aspects should be considered. For example, the number of words to be recognized should be increased. Moreover, the system can recognize more complicated sentences which are often used in real-life situation. Another weak point of the proposed scheme is that the locations of electrodes were not determined optimally. Hence, a detailed analysis of relationship between the function of each facial muscle and the phonemes spoken would be desirable. This will give a clue to determine the optimal locations for collecting the surface EMG signals. Our future studies will focus on these issues.

## REFERENCES


[1] F. Grandori, P. Pinelli, P. Ravazzani, F. Ceriani, G. Miscio, F. Pisano, R. Colombo, S. Insalaco, and G. Tognola, "Multiparametric analysis of speech production mechanisms," *IEEE EMB Magazine*, vol. 13, issue 2, pp. 203-209, 1995.

[2] E.A. Goldstein, J.T. Heaton, J.B. Kobler, G.B. Stanley, and R.E. Hiiman, "Design and implementation of a hands-free electolarynx device controlled by neck strap muscle electromyographic activity," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 325-332, 2004.

[3] H. Manabe, and Z. Zhang, "Multi-stream HMM for EMG-based speech recognition," in *Proc. 26th Annual International Conference of the IEEE EMBS,* San Francisco, CA, USA, 2004, pp. 4389-4392.

[4] S. Kumar, D.K. Kumar, M. Alemu, and M. Burry, "EMG based voice recognition," in *Proc. 2004 Intelligent Sensor, Sensor Networks and Information Processing Conference*, 2004, pp. 597-596.

[5] H.-J. Park, S.-H. Kwon, H.-C. Kim, and K.-S. Park, "Adaptive EMG-driven communication for the disability," in *Proc. 1st Joint BMES/EMBS Conference,* Atlanta, GA, USA, 1999, pp. 656.

[6] A.D.C. Chan, K. Englehart, B. Hudgins, and D.F. Lovely, "Hidden Markov Model classification of myoelectics signals in speech," *IEEE EMB Magazine*, vol. 9, pp. 143-146, 2002.

[7] A.D.C. Chan, K. Englehart, B. Hudgins, and D.F. Lovely, "Myoelectric signals to augment speech recognition," *Med. Biol. Eng. Comput.*, pp. 500-504, 2001.

[8] R.S. Kumaran, K. Narayanan, and J.N. Gowdy, "Myoelectric signals for multimodal speech recognition," in *Proc. 2005 EUROSPEECH,* Lisboa, Portugal, 2005, pp. 1189-1192.

[9] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ, US A: Prentice-Hall, 1993.

[10] G.M. White and R.B. Neely, "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-24, no. 2, pp. 183-188, 1976.

[11] C. Jorgensen and D.D. Lee, and S. Agabon, "Sub auditory speech recognition based on EMG signals," in *Proc. the InternationalJoint Conference on Neural Network*, vol. 4, 2003, pp. 3128-3133.

[12] L.R. Rabiner, and R.W. Schafer, *Digital Processing of Speech Signal*, Englewood Cliffs, NJ, USA : Prentice Hall, 1978.

[13] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society*, vol. 39, pp. 1-38, 1977.

[14] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS system," in *Proc. the Joint Meeting of ASA, EAA, and DAGA*, Berlin, Germany, March 1999.

[15] L.R. Rabiner, J.G. Wilpon and F.K. Soong, "High performance connected digit recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, issue 8, pp. 1214-1225, 1989.

[16] K. Ogino and W.M. Kozak, "Spectrum analysis of surface electromyogram," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Boston, MA, USA, 1983, pp. 1114-1117.

[17] B. Fisher, Tsylb2-1.1 Syllabification software, Available: http://www.nist.gov/speech/tools, August, 1996.

[18] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland, HTK Speech Recognition Toolkit, Available: http://htk.eng.cam.ac.uk