

# Speech Recognition Using EMG; Mime Speech Recognition

*Hiroyuki Manabe, Akira Hiraiwa, Toshiaki Sugimura*

NTT DoCoMo Multimedia laboratories  
manabe@mml.yrp.nttdocomo.co.jp

## Abstract

The cellular phone offers significant benefits but causes several social problems. One such problem is phone use in places where people should not speak, such as trains and libraries. A communication style that would not require voiced speech has the potential to solve this problem. Speech recognition based on electromyography (EMG), which we call "Mime Speech Recognition" is proposed. It not only eases communication in socially sensitive environments, but also improves speech recognition accuracy in noisy environments. In this paper, we report that EMG yields stable and accurate recognition of 5 Japanese vowels uttered statically without generating voice. Moreover, the ability of EMG to handle consonants is described, and the feasibility of basing comprehensive speech recognition systems on EMG is shown.

## 1. Introduction

With the appearance of the cellular phone, user communication style has evolved greatly. Users can communicate anytime, anywhere, and with anyone. However, this is creating social problems. These benefits lead the user into talking where no one should talk, i.e. libraries, theaters, and restaurants. Especially in Japan, the use of cellular phones in public places, such as trains and buses, is developing into a serious social problem. Though there are several reasons, a key reason is that loud voices upset the surrounding passengers. A survey conducted in 2000 by the Japanese government found that 80% or more of passengers regard the use of cellular phones and PHS units in public spaces as unpleasant.

There are some approaches to solve the problem. A new communication style that does not require voiced speech is one of the approaches. If speech without generating voice could be recognized by using a technique such as lip-reading, this problem could be reduced. We call speech recognition that makes this communication style possible "Mime Speech Recognition". Mime Speech Recognition also has the possibility of solving other weakness present in the conventional speech recognition techniques and acting as an utterance tool for a cordectomy patient.

The accuracy of speech recognition is strongly degraded in noisy environments. Since it is obvious that cellular phones will be used in various noise environments, robust speech recognition in noisy environments is strongly desired. Several approaches have been proposed to improve recognition accuracy in such environments. One approach is based on using the signal components that are robust against noise in addition to the voice signal. More robust speech recognition can be realized by unifying the information detected by Mime

Speech Recognition and that acquired by conventional speech recognition.

Mime speech Recognition can be realized in the following ways. When speaking, the vocal and articulation organs are moved, which does not depend whether with generating voice or without generating voice. This implies that speech can be recognized by observing the movements or activities of such organs. Three methods can be used to observe their movements. One method is to use image capture and processing. Another method is to use EMG. The other method is to use a magnetic field.

We examined the possibility of speech recognition from captured EMG signals towards the realization of communication via cellular phones that does not require the speech to be voiced.

## 2. Image vs. EMG vs. Magnetic Field

The use of image processing to realize speech recognition is generally called lip-reading or speech-reading. It has been pointed out that it is not only hearing-impaired people who perform lip-reading; healthy people also do it unconsciously. Unfortunately, lip-reading fails to discriminate between some phonemes. When image processing is used, only those body motions and changes that appear on the exterior of the body can be detected. It is virtually impossible to reliably detect body motions and changes within the body, such as those of the tongue, all of which play critical roles in forming utterances. Furthermore, image processing is computationally expensive and so does not suit small terminals such as cellular phones. Moreover, the camera used to capture the facial images must be fixed at some distance from the face, which suggests very poor wearability and convenience.

EMG offers several important features. EMG is able to detect the activities of the muscles involved with utterance formation. The first benefit of EMG is the ability to observe internal muscle states. The second benefit is ease of detection; all that is needed is the placement of skin electrodes close to the muscle of interest. This benefit is very important if we are to modify a cellular phone to support Mime speech Recognition. Since most users hold the cellular phone against or close to his face, it seems possible to detect EMG by building electrodes into the cellular phone.

The other method is based on magnetic fields. This method detects the form and motion of the articulation organ or vocal organ. Actual observations are performed using fMRI and a magnetic sensor. Although its spatial resolution is high, current measurement devices are too large for inclusion in a cellular phone. Furthermore, there is also the problem that markers must be stuck on the body parts suggesting extremely poor wearability.

The above discussion indicates that EMG is the best way of

enhancing cellular phones to realize Mime Speech Recognition.

### 3. Related Works

EMG represents the activities of muscles and so has been the subject of research on speech recognition. Sugie et al. tried to recognize speech using EMG as a speech prosthesis for handicapped people[1]. From the EMG measured at three places on the face, they ascribed muscle activities into binary states; they tried to recognize 5 Japanese vowels using an automaton and obtained the accuracy of 64% (average). Morse et al. tried to recognize 10 English words and achieved the accuracy of 60%[2]. The EMG observed from three places on the neck and one place on the temple was broken down into frequency components and analyzed using a neural network. Chan et al. used EMG to augment speech recognition[3]. These efforts show that EMG can be use for speech recognition.

### 4. Approach

It is, in a strict sense, impossible to perform all the operations needed to utter all phonemes while not uttering any voice at all. However, if the voiced speech is extremely quite, the EMG should be enough to permit adequate recognition accuracy.

#### 4.1. Words vs. phonemes

There are two approaches toward realizing Mime speech Recognition. One approach is to set anticipated words[2][3]. This approach is expected to yield high recognition accuracy even if the number of targeted words is low. However, this approach is strongly language dependent and the use of unusual or new words is inhibited. The other approach is based on phoneme recognition[1]. While there are many phonemes that must be recognized, the recognition process is independent of specific words and specific languages. If it is considered that development of the conventional speech recognition is based on the feature for every phoneme, also when EMG is used, it will be necessary to catch the feature for every phoneme. We decided to take the latter approach. As the first step, we measured EMG during Japanese utterances.

#### 4.2. EMG during utterance

We observed the following characteristics of utterances. Fig. 1 plots the distance between the center of the upper lip, and the center of the lower lip, observed by an optical motion capture system, and the surface EMG observed from under the jaw of a subject who started with a closed mouth, relaxed facial muscles for about 1.5sec., and uttered /kaka/ without generating voice. In Fig. 1 the lip movement corresponding to uttering first /k/ is seen from 1.5 to 2sec. and that corresponding to uttering second /k/ is at 4sec. The surface EMG measured under the jaw represents the sum of signals emitted from the muscles for opening the jaw, and the muscles for moving the tongue. Fig. 1 shows three points. One is that EMG is preceded by movement operation, and so is distinguishable. The second is that when opening the mouth and moving the tongue to utter /k/, large amplitude EMG is seen; this indicates that the muscles for opening the jaw and moving the tongue work suddenly. The third is that when the

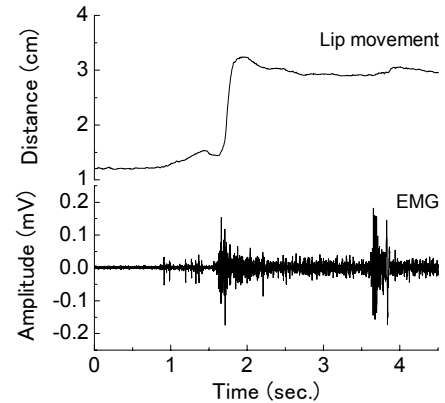


Fig. 1: The lip movement and measured EMG.

second /k/ was uttered, although there was almost no change in the distance between upper lip and lower lip, large amplitude EMG is seen which corresponds to the operation of moving the tongue. These results confirm that EMG, unlike image processing, can clearly capture the state of the tongue. This is significant, especially when attempting speech recognition.

#### 4.3. Japanese Utterances

Japanese has 5 vowels; /a/, /i/, /u/, /e/, and /o/. Consonants are not uttered continuously and are always followed by a vowel. This makes it obvious that vowels play an important role in Japanese speech recognition. As the first step we attempted vowel recognition. Uttering Japanese vowels involves three characteristic operations. They are pursing the lips (especially for /u/ and /o/), lifting the corners of the mouth (/i/), and opening the jaw (/a/, /e/, and /o/). These operations require activation of the orbicularis oris, zygomaticus major, and digastricus.

Muscles are strengthened in two cases: to hold the same condition, and to change the condition. In stream utterance, in order to change the condition, the muscles always need to change the strength levels of the muscles. In this case, the strength required to utter a specific phoneme is not fixed, it depends on the immediately prior location of the articulation organ. As is widely known, EMG represents the strength levels generated by the muscles, so in order to recognize continuous speech from EMG, it is necessary at least to recognize the immediately prior condition. Since this paper addresses the recognition of vowels uttered in isolation, the recognition of consonants and stream utterances are future goals.

### 5. EMG for isolated vowels

We captured the EMG when uttering the 5 Japanese vowels statically without generating voice, and used the EMG to recognize which vowel was uttered[4].

#### 5.1. EMG capture

We captured EMG from three areas: on the lip, on the cheek, and under the jaw. These areas overlay the three key muscles mentioned above. EMG was captured by pressing electrodes mounted on subject's fingers to his face (Fig. 2). This measurement style strongly supports integration with a wearable terminal. We measured EMG while three subjects uttered the 5 vowels for 12 seconds statically without



Fig. 2: Arrangement of electrodes.

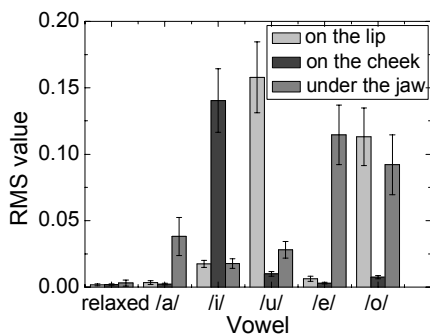


Fig. 3: RMS values at uttering vowels.

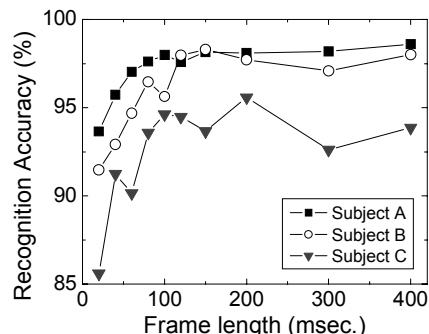


Fig. 4: Recognition accuracy of vowels.

generating voice. Each trial was repeated 10 times. The signal detected with the electrodes was amplified 1000 times. Sampling frequency was 2000Hz, A/D conversion resolution was 16bits. Measured EMG was changed into RMS (Root Mean Square) values. Fig. 3 shows the average RMS (the standard deviation is given by the error bar) for each vowel; the frame length was 100ms and the frame period was 20ms. "Relaxed" which represents silent sections in conventional speech recognition research is also shown. Fig. 3 shows that the 5 vowels and "relaxed" produce unique EMG characteristics that can be well discriminated.

## 5.2. Recognition of Vowel EMG

Since the results shown in Fig. 3 indicate clear separation, we thought that any recognition algorithm would work. We decided to use a neural network given its ease of construction and its ability to offset individual differences. We used a three-layer neural network with 3 input layer nodes, 6 hidden layer nodes, and 5 output layer nodes, each of which corresponds to one of the 5 Japanese vowels. If none of the output layer nodes fire, the recognition result is "relaxed". Ten sets of EMG were measured using 12 seconds per vowel. To train the neural network, 9 sets of EMG were used, the remaining 1 set was used as recognition data. The neural network was trained 1000 times using the back propagation algorithm. The training of the neural network and EMG recognition was repeated 10 times, once for each of the 10 measured sets. The average value of recognition was taken as the final recognition result. This recognition completely disregarded EMG time correlation and recognition was performed using the data of just each one frame of three channels. Two subjects were males, 1 was female, and all were in their twenties. Fig. 4 shows the recognition accuracy total for each vowel and "relaxed" when the frame period is fixed at 20ms while the frame length was changed from 20ms to 400ms. The average accuracy exceeded 95% for all subjects with frame length of 200ms. This result shows that we can recognize the 5 Japanese vowels by using EMG. Based on this result, we built a system that performs recognition in real time with the frame length of 400ms and frame period of 200ms. That is, even if we reduce the influence of EMG generated when the vowel uttered changes, we can still achieve high recognition accuracy.

## 6. EMG for consonants

It was mentioned above that the operation used to utter a consonant greatly depends on the prior utterance operation. To simplify this test, we decided to measure EMG when the

consonant was preceded and followed by the same vowel. The same EMG measurement conditions used for isolated vowel recognition were used. The vowel was /a/, and the consonants were /k/, /s/, /t/, /n/, /h/, /m/, /j/, and /r/ which are the basic Japanese consonants. We recorded the signals when the vowel was uttered statically for 12sec. without generating voice, and the consonant is uttered every one second.

### 6.1. EMG for consonants

An example of the measured raw EMG, captured when uttering /ama/, is shown in Fig. 5. Note that the scales in Fig. 5 have been expanded for clarity; EMG on the cheek is expanded 5 times and EMG from under the jaw is doubled. Three kinds of features corresponding to uttering /m/ are seen in the measured EMG from 0.45 to 0.7sec. The first feature is large amplitudes, for example EMG on the lips from 0.45 to 0.6sec. The second feature is depression of amplitude; for example, EMG from under the jaw from 0.45 to 0.55sec. The third feature involves the time difference of the above mentioned features caught in different channels; for example, there is about 0.1sec. between the large amplitude of EMG on the lip and that of EMG from under the jaw. These features are certainly not observed for all consonants and in all EMG channels. This is clearly seen for consonant /h/. However, one or more of these 3 features were seen in at least one or more of the channels for the other consonants.

Prior to analysis, the measured EMG was processed. In most cases, the large amplitude continues for 100ms to 200ms. We decided to extract the measured EMG with 50ms frame length and 20ms frame cycle, and from them calculate RMS values. Fig. 6 shows calculated RMS for /ama/, which corresponds to Fig. 5. In Fig. 6, the EMG of 2 channels is expanded and displayed as in Fig. 5.

### 6.2. Large amplitude

Large amplitude was seen most clearly in the observed EMG, see Fig. 6. When uttering a consonant, the speech muscles work suddenly. The large amplitude seen in observed EMG can be interpreted as a result of sudden activity of these muscles. This characteristic of large amplitude was seen for all consonants examined, except for /h/, in at least one channel.

### 6.3. Depression of amplitude

Depression of amplitude was observed frequently, especially in the EMG measured from under the jaw. In Fig. 6, the RMS value measured from under the jaw at around 0.45sec. shows this characteristic. When the level of EMG for the vowel was large, this depression was especially noticeable.

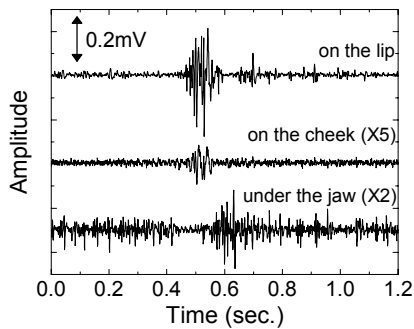


Fig. 5: Measured EMG at uttering /ama/.

#### 6.4. Time difference between the features

The first and the second feature are the features observed within one channel. This third feature is acquired by observing those features over two or more channels. That is, a difference is seen for the time when the large amplitude and depression of amplitude is observed between channels. In Fig. 6, it is clear that there is about 0.1sec. between the peaks of RMS measured on the lip and RMS measured from under the jaw.

### 7. Discussion

Here we consider the idea that the EMG features observed when a consonant is uttered depend on what kind of speech mechanism was used. Large amplitude indicates the sudden activation of muscles. In order to utter a consonant just after uttering a vowel, the speech muscles must work quickly to reset the articulation organ. All the basic Japanese consonants examined in this paper, except /h/, require modification of the tongue and closing of the jaw. These activities use muscles under the jaw, and indeed large signal amplitudes were observed in the EMG measured from under the jaw. Moreover, especially in order to utter /m/, it is necessary to close lips and large amplitude is observed in EMG measured on the lip. The second feature, the drop in amplitude, represents a temporary weakening of muscle activity. To close a jaw that is currently open in order to utter a consonant, the muscles that open the jaw need to weaken (in terms of their activities) temporarily. Next, in order to open the jaw, those muscles need to generate high levels of strength. This indicates that depression of amplitude should be observed just before a large amplitude event and naturally suggests the validity of the third feature, the timing of events. Consider, as an example, the utterance of /ama/. In order to utter /m/ after uttering /a/, it is necessary to close the jaw and lips, both of which are currently open. In order to perform this operation, the muscles that open the jaw must weaken, and the muscles for closing the lips must contract strongly. At this time, large amplitude is observed in the EMG measured on the lip, and depression of amplitude is observed in the EMG measured from under the jaw. At the next instance, in order to utter /m/, it is necessary to open the jaw and the lips. At this time, large amplitude is observed in EMG measured from under the jaw. In this sequence, there is noticeable time difference between the EMG measured from under the jaw, and the EMG measured on the lip.

In the trials made for this paper, most features were not stable during the measurements. We believe that there are two main causes. The first is the difficulty of moving the speech organs without vocalization. It is possible that the subjects were not fluent in this style of communication. This problem can be solved by training and is not considered to be serious handicap

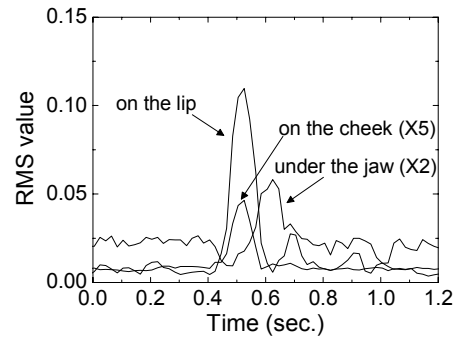


Fig. 6: Calculated RMS at uttering /ama/.

to adoption. The second cause is rather immature measurement system used. The areas measured this time were areas which is effective in recognition vowels. It differs from the activities of muscles which are required to utter vowels, and those required to utter consonants. Therefore, it is necessary to catch the activities of the muscles which become important to utter consonants. Moreover, although RMS values of EMG were used in this paper, it is necessary to confirm the optimum features, such as frame length and frame period.

While a lot more work is needed, our work has clarified the following. The three characteristics of EMG seen when uttering many consonants suggest that consonant recognition is possible. Moreover, EMG information can be used to enhancement the performance of conventional speech recognition techniques.

### 8. Conclusions

Mime speech Recognition, which can solve the social problem caused by the use of cellular phones in public spaces, was proposed; it does not require the speech to be voiced. We captured the EMG from three places on the face while uttering the 5 Japanese vowels in isolation. Trials showed that the vowels can be recognized with high accuracy when using short frame lengths. EMG was also captured when uttering consonants. The results showed that the three main characteristics of EMG, large amplitude, depression of amplitude, and event sequences, suggest the possibility that consonants can be recognized.

### 9. References

- [1] N. Sugie, K. Tsunoda, "A Speech Prosthesis Employing a Speech Synthesizer – Vowel Discrimination from Perioral Muscle Activities and Vowel Production", *IEEE Trans. Biomedical Engineering*, Vol. BME-32, No. 7, pp. 485-490, 1985.
- [2] S. Morse, Y. N. Gopalan, M. Wright, "Speech recognition using myoelectric signals with neural networks", *Proc. 13th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vol. 13, No. 4, pp.1877-1878, 1991.
- [3] A. D. C. Chan, K. Englehart, B. Hudgins, D. F. Lovely, "Myo-electric signals to augment speech recognition", *Medical & Biological Engineering & Computing 2001*, pp. 500-504, 2001.
- [4] H. Manabe, A. Hiraiwa, T. Sugimura, "Unvoiced Speech Recognition using EMG", *extended abstracts of CHI2003*, pp. 794-795, 2003