Frame-Based SEMG-to-Speech Conversion

Yuet-Ming Lam
Department of Computer Science
and Engineering,
The Chinese University of Hong Kong,
Shatin, Hong Kong.
Email: ymlam@cse.cuhk.edu.hk

Philip Heng-Wai Leong
Department of Computer Science
and Engineering,
The Chinese University of Hong Kong,
Shatin, Hong Kong.
Email: phwl@cse.cuhk.edu.hk

Man-Wai Mak
Department of Electronic
and Information Engineering,
The Hong Kong Polytechnic University,
Hung Hom, Hong Kong.
Email: enmwmak@polyu.edu.hk

Abstract— **This paper presents a methodology that uses surface electromyogram (SEMG) signals recorded from the cheek and chin to synthesize speech. A neural network is trained to map the SEMG features (short-time Fourier transform coefficients) into vector-quantized codebook indices of speech features (linear prediction coefficients, pitch, and energy). To synthesize a word, SEMG signals recorded during pronouncing a word are blocked into frames; SEMG features are then extracted from each SEMG frame and presented to the neural network to obtain a sequence of speech feature indices. The waveform of the word is then constructed by concatenating the pre-recorded speech segments corresponding to the feature indices. Experimental evaluations based on the synthesis of eight words show that on average over 70% of the words can be synthesized correctly and the neural network can classify SEMG frames into seven phonemes and silence at a rate of 77.8%. The rate can be further improved to 88.3% by assuming medium-time stationarity of the speech signals. The experimental results demonstrate the feasibility of synthesizing words based on SEMG signals only.**

I. Introduction

Speech is the most natural way of communication among humans. The speech production process involve the contraction of the lungs, the vibration of the vocal cords and the resonance of the air stream in the vocal tract. Unfortunately, there are situations in which communication through speech is impossible or inappropriate. For example, people suffering from the side effect of laryngectomy surgeries or vocal cord damage are not able to produce normal speech. Speech communication can also be affected by a number of factors; for instance, background noise can degrade the quality of the produced speech, resulting in poor intelligibility.

To address some of the limitations of speech communication, non-acoustic communication systems that use surface electromyogram [1] signals to recognize speech have been proposed. For example, Jorgensen et al. [2] used two-channel SEMG signals from the larynx to recognize six words. In Chan et al. [3], an SEMG-based recognizer for the digits '0' to '9' was proposed to complement the speech in some noisy environments. In their system, five pairs of Ag-AgCl button electrodes are placed under a pilot's oxygen mask to record the SEMG signals from the primary facial muscles. SEMG-based phoneme recognition was presented in [4], where each phoneme was regarded as an isolated word to perform recognition.

The work mentioned earlier demonstrate the feasibility of recognizing speech based on SEMG signals. However, most of the proposed approaches focused on recognizing or classifying the SEMG signals into a limited set of words. These approaches are similar to conventional isolated word recognition systems in that there must be sufficient silence intervals before and after the speech signals, i.e., the words must be segmented and isolated from each other before recognition can be taken place. These word-recognition systems have difficulties in recognizing continuous speech, and the recognition accuracy can be affected by the duration of the words.

Instead of recognizing isolated words from SEMG signals, this paper proposes to synthesize speech directly from SEMG signals. In particular, features are extracted from SEMG signals and converted into speech features in a frame-by-frame basis. The conversion is done via a multi-layer neural network trained to classify the short-time Fourier transformation parameters of SEMG signals into phonetic classes. Because the features are extracted from phonemes and conversion is done at the frame level, the proposed methodology has the potential to be applied to recognize continuous speech with unlimited vocabulary .

The organization of this paper is as follows. In Section II, the design methodology, including data acquisition, feature selection, neural network training and speech synthesis methodology, are introduced. Classification results and the synthesized speech are presented in Section III, the correlation between SEMG frame size and recognition performance of the neural network will also be addressed. Finally, conclusions are drawn in Section IV.

II. Methodology

A. Data Acquisition

Two channels of SEMG signals were collected as shown in Figure 1. The first channel was collected from the cheek about 2.5cm from the nose and another was collected from the chin. An additional electrode was attached to the forehead as a reference point. Speech was recorded using a microphone. The SEMG signal was amplified with a gain of 1000. Both the amplified SEMG signal and speech were recorded concurrently using a National Instruments Inc. PCI6024E PCI data acquisition card [5] at a sampling rate of 8000Hz.

1-4244-0173-9/06/$20.00 ©2006 IEEE.

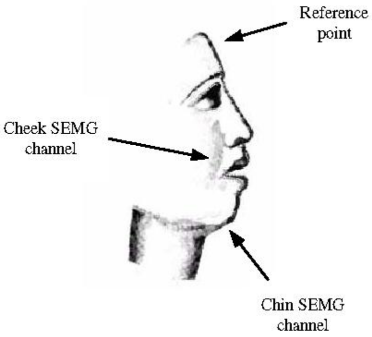

Fig. 1. Electrode placement: SEMG signals were collected from the cheek and chin, forehead was used as reference point.

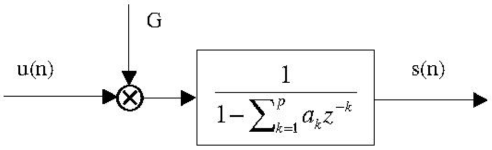

Fig. 2. Linear predictive coding model.

The training set consists of data samples of seven phonemes: *ae, iy, ao, uw, sh, f* and *s*. Each data sample was recorded in a twenty-second period, during which the speaker repeatedly pronounced a given phoneme. The training data comprises the SEMG and speech signals of four such data samples for each phoneme.

Both phonemes and words were used for testing. The words are *shaw, she, ash, shoe, see, saw, fee* and *off*, whose phonetic transcriptions are formed by concatenating the training phonemes. The testing data consists of one data sample of each phoneme and one data sample of each word. The recorded speech of the testing set was used as references for performance evaluation.

B. Speech Feature Selection

Figure 2 shows the linear predictive coding [6] model, where $u(n)$ is an excitation signal containing pitch information and G is the gain term representing the speech energy. In this work, the input speech was blocked into frames using a frame size of 22.5ms, and there was no overlapping between frames. This scheme is commonly used in speech coding [7]. For each speech frame, the LP coefficients, pitch, and energy were extracted and used as speech features.

C. SEMG Feature Selection

The short-time Fourier transform (STFT) [8] coefficients were used as SEMG features. The SEMG signals for each channel was also blocked into frames. For each frame, the frequency spectrum from 1 Hz to 450 Hz is calculated from each SEMG frame and divided into 10 equal frequency bands, each with a bandwidth of 45 Hz. The frequency components in each band were summed to give one STFT coefficient corresponding to that band. Ten STFT coefficients were thus obtained for each frame. As a result, 20 STFT coefficients were extracted from the two SEMG channels.

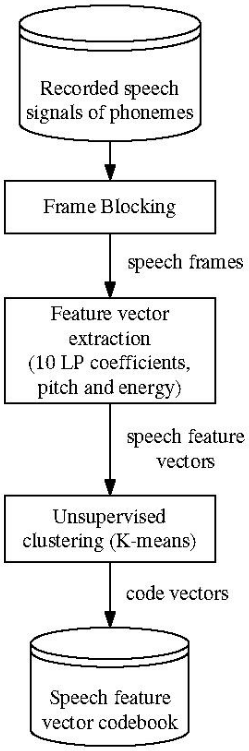
Recorded speech signals of phonemes
Frame Blocking
speech frames
Feature vector extraction (10 LP coefficients, pitch and energy)
speech feature vectors
Unsupervised clustering (K-means)
code vectors
Speech feature vector codebook

Fig. 3. Creation of speech feature vector codebook: LP coefficients, pitch, and energy were extracted from each speech frame. Unsupervised clustering was performed to extract the representative features vector for each phoneme. These features vectors form a codebook.

The SEMG frame size should be chosen carefully, because it affects the frequency resolution [9]. If a small frame size is used, better time resolution can be obtained but this results in poor frequency resolution. On the other hand, using larger frame sizes can improve the frequency resolution but with a loss in information between adjacent frames. In this paper, correlation between frame size and performance is analyzed.

D. Neural Network Training

Figures 3 and 4 illustrate the feature set construction and neural network training. As shown in Figure 3, the speech signals of training phonemes were blocked into frames and the LP coefficients, pitch, and energy were extracted and concatenated to form speech feature vectors for each frame. Unsupervised clustering, using the k-means algorithm, was performed to extract the representative feature vectors for each phoneme. These extracted feature vectors form a codebook.

After forming the speech feature vector codebook, vectors for training a neural network can be constructed as shown in Figure 4. As mentioned earlier, the SEMG signals and speech were recorded concurrently. The speech signals of training phonemes were blocked into frames, and the extracted speech

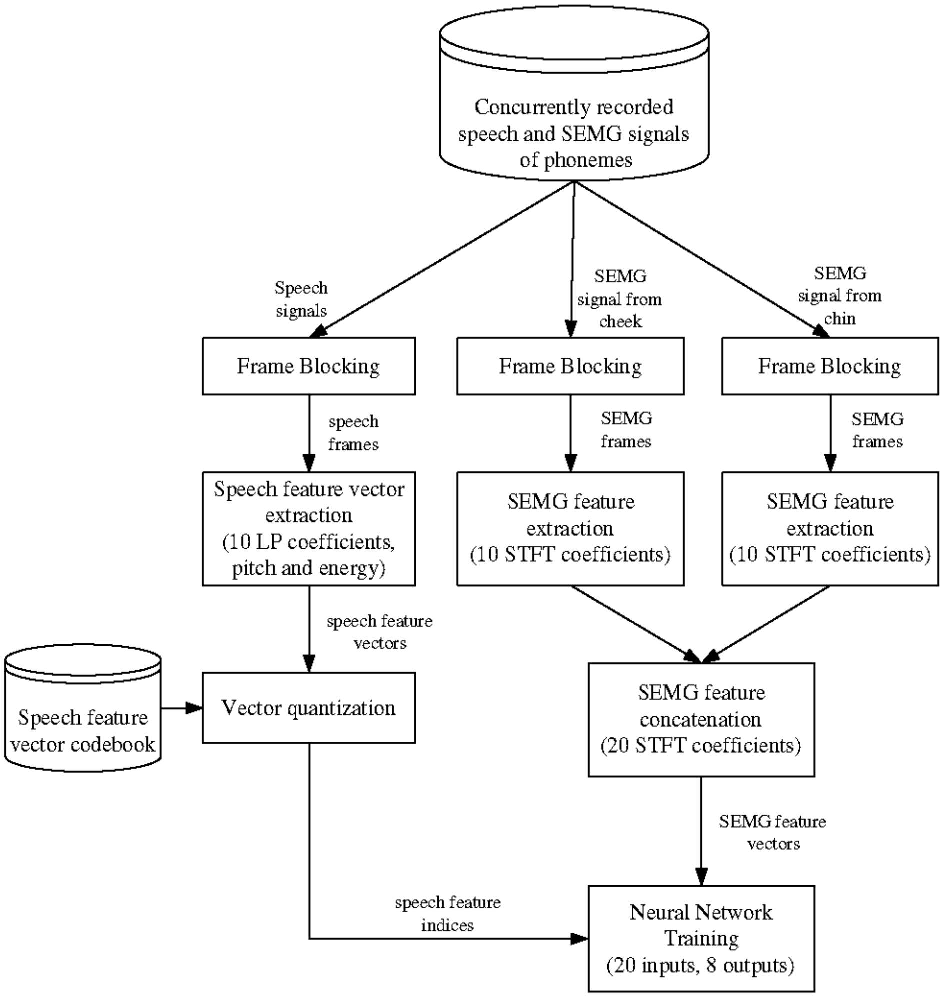
Concurrently recorded speech and SEMG signals of phonemes
Speech signals
SEMG signal from cheek
SEMG signal from chin
Frame Blocking
Frame Blocking
Frame Blocking
speech frames
SEMG frames
SEMG frames
Speech feature vector extraction (10 LP coefficients, pitch and energy)
SEMG feature extraction (10 STFT coefficients)
SEMG feature extraction (10 STFT coefficients)
speech feature vectors
Speech feature vector codebook
Vector quantization
SEMG feature concatenation (20 STFT coefficients)
SEMG feature vectors
speech feature indices
Neural Network Training (20 inputs, 8 outputs)

Fig. 4. Frame based feature set construction and neural network training: Features were extracted from each SEMG and speech frame. Speech feature vectors were vector quantized to obtain the feature indices; the neural network was trained using SEMG feature vectors as input and the corresponding speech features indices as output.

feature vectors from speech frames were vector quantized using the speech feature vector codebook. Thus, each speech frame was represented by a codebook index. As the codebook was formed by the representative feature vectors of phonemes, this codebook index indicates to which phoneme a speech frame belongs. The SEMG signals of training phonemes were also blocked into frames, and the STFT coefficients were extracted from two SEMG channels and concatenated to form an SEMG feature vector. Each of the SEMG feature vectors was paired with the corresponding speech feature index to form an input-target training pair. A two-layer feed-forward backpropagation neural network [10], which takes an SEMG feature vector as input and produced one of eight possible outputs (including silence and seven phonemes), was trained using the input-target pairs.

E. Speech Synthesis

Speech can be synthesized from an SEMG signal as shown in Figure 5. This synthesis method can be applied not only to phonemes, but also to words or even sentences. In this work, SEMG signals of phonemes or words were recorded from both channels and blocked into frames. The SEMG feature vectors were extracted. The trained neural network was then used to classify these vectors and to produce sequences of speech feature indices.

Classification performance of the neural network can be improved by using an error correction technique. The idea is based on the observation that voiced speech signals are fairly stationary over a medium period of time (on the order of 200ms or more [6]). The error correction technique scans the produced sequence of speech feature indices over a window of 10 indices, i.e. 225ms, and replace these 10 indices by the one with highest occurrence frequency within the window.

After performing error correction on the sequence of speech feature indices, a concatenation synthesis method [11] was applied to reconstruct the target speech in a frame-by-frame basis. Based on the error corrected speech feature indices, target phoneme frames were loaded from the pre-recorded set

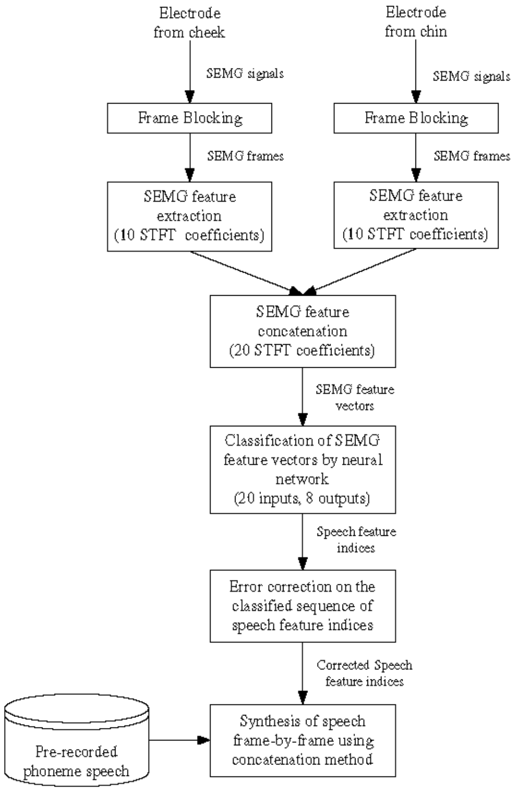

Fig. 5. Speech (phonemes, words or sentences) synthesis based on SEMG signals: SEMG feature vectors were extracted from each SEMG frame and classified into speech feature indices. Speech was synthesized frame-by-frame based on the error-corrected speech feature indices.

and concatenated to form the complete speech. The transition between phonemes was smoothed using overlap and add method.

III. EXPERIMENTS AND RESULTS

A. SEMG Frame Size

To find the SEMG frame size that balances the trade-off between the time and frequency resolution, classification performance of the neural network for SEMG frames of different sizes were analyzed. Classification performance was evaluated using SEMG feature vectors extracted from one data sample of each phoneme, and the average classification rates for SEMG frame sizes from 22.5ms to 202.5ms is shown in Figure 6. A clear trend can be seen in this figure: the classification rate is higher for larger SEMG frame sizes and becomes saturated for frame sizes larger than 112.5ms. Although a slightly higher classification rate was obtained at 202.5ms, frame size 112.5ms was chosen for further experiment as smaller frame size gives better time resolution.

B. Error correction

Table I shows the confusion matrix for phonemes' SEMG feature vector classification at a frame size of 112.5ms. The rows show the results of the neural network classification, and the columns are the feature vectors of the original speech.

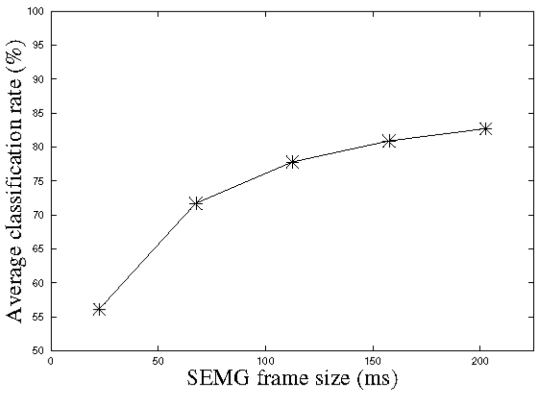

Fig. 6. Average classification rate of SEMG feature vectors for different SEMG frame sizes.

TABLE III
SYNTHESIS RESULT FOR WORDS

Words	Number of words involved	Number of words synthesized correctly
she	8	7
ash	7	6
shaw	9	9
see	8	6
saw	9	5
shoe	8	2
fee	8	4
off	7	6
Total	64	45

The average classification rate is 77.8%. The error correction process was applied to correct misclassification errors and the results after applying correction are shown in Table II. Although more voiced SEMG feature vectors are misclassified as silence, the overall classification rates for all phonemes are improved and the average classification rate is improved to 88.3%.

C. Speech synthesis

Words were synthesized using the trained neural network and the error correction technique. Input SEMG signals for words were blocked into frames. SEMG feature vectors were extracted and the trained neural network was used for classification. Error correction process was applied and words were synthesized by the concatenation method. One 20-second sample of each word was used for this experiment, Table III shows the results obtained. The percentage of words synthesized correctly, i.e., distinguishable, is 70.3%. Figure 7 shows the spectrograms of four synthesized instances. One can see that the synthesized instances and the reference speech have similar characteristics despite the words have not been involved in the training process. Although the synthesized instances and the reference speech may not perfectly aligned, for example, some silence frames before the reference word *ash* are synthesized as phoneme *ae* as shown in sub-figures (1b) and (2b) of Figure 7, the intelligibility of the synthesized

TABLE I

CONFUSION MATRIX FOR SEMG FEATURE VECTOR CLASSIFICATION AT SEMG FRAME SIZE 112.5MS

Neural Network classification	Reference Speech							
	Silence	*ae*	*iy*	*ao*	*uw*	*sh*	*f*	*s*
Silence	87.8%	3.5%	8.3%	4.9%	3.2%	14.3%	8.7%	17.9%
ae	1.4%	85.3%	0.0%	0.2%	6.1%	1.4%	5.7%	0.0%
iy	1.4%	0.0%	66.9%	0.5%	0.0%	0.0%	0.2%	2.2%
ao	1.7%	0.3%	9.1%	88.1%	0.0%	1.5%	0.2%	0.2%
uw	1.2%	6.7%	0.0%	0.0%	71.6%	7.6%	13.4%	0.0%
sh	1.1%	1.9%	0.0%	6.3%	5.3%	73.9%	2.4%	0.0%
f	1.3%	2.3%	0.0%	0.0%	13.4%	1.1%	69.2%	0.0%
s	4.1%	0.0%	15.7%	0.0%	0.4%	0.2%	0.2%	79.7%

TABLE II

CONFUSION MATRIX AFTER APPLYING ERROR CORRECTION TO THE CLASSIFIED SPEECH FEATURE INDICES

Neural Network Classification	Reference Speech							
	Silence	*ae*	*iy*	*ao*	*uw*	*sh*	*f*	*s*
Silence	91.8%	4.7%	8.3%	5.5%	8.3%	17.6%	14.8%	8.5%
ae	1.4%	95.3%	0.0%	0.0%	0.6%	0.0%	0.0%	0.0%
iy	1.2%	0.0%	80.7%	0.0%	0.0%	0.0%	0.0%	0.0%
ao	1.0%	0.0%	0.7%	92.7%	0.0%	0.0%	0.0%	0.0%
uw	1.0%	0.0%	0.0%	0.0%	90.9%	0.3%	3.8%	0.0%
sh	1.1%	0.0%	0.0%	1.9%	0.2%	82.1%	0.0%	0.0%
f	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	81.5%	0.0%
s	1.8%	0.0%	10.3%	0.0%	0.0%	0.0%	0.0%	91.5%

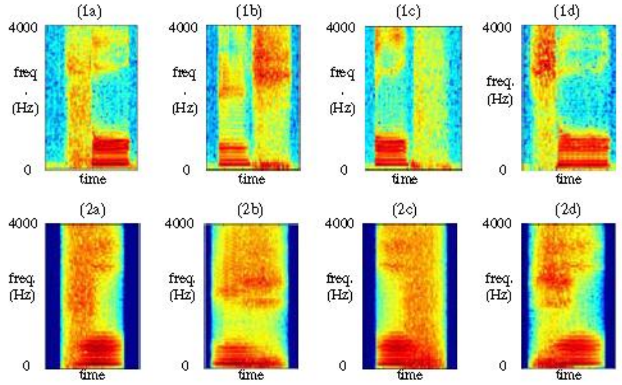

Fig. 7. Spectrograms of four synthesized instances. (1a) - (1d) are reference speech of *saw*, *ash*, *off* and *shaw* respectively, and (2a) - (2d) are the corresponding synthesized speech.

instances are not affected.

IV. CONCLUSIONS

This work has demonstrated the feasibility of synthesizing speech from SEMG data using a phoneme-based feature extraction and frame based feature conversion methodology. The effect of varying the SEMG frame size was studied, and a frame size 112.5ms was found to provide a good balance between time and frequency resolution. It was further shown that the quality of the synthesized speech can be improved by utilizing knowledge concerning the medium-term stationarity of speech. We believe the proposed methodology can be potentially scaled up to an unlimited vocabulary continuous speech synthesis system.

REFERENCES

[1] S. Kumar and A. Mital, *Electromyography in Ergonomics*. Taylor & Francis Ltd., 1996.

[2] C. Jorgensen, D.D. Lee, and S. Agabon, "Sub auditory speech recognition based on emg signals," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 4, 2003, pp. 3128–3133.

[3] A.D.C. Chan, K. Englehart, B. Hudgins, and D.F. Lovely, "Hidden markov model classification of myoelectric signals in speech," in *Proceedings of the 23rd Annual International Engineering in Medicine and Biology Society*, vol. 2, 2001, pp. 1727–1739.

[4] C. Jorgensen and K. Binsted, "Web browser control using emg based sub vocal speech recognition," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 2005, pp. 294c–294c.

[5] *6023E/6024E/6025E User Manual*, National Instruments Inc., december 2000 Edition.

[6] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall PTR, 1993.

[7] T.E. Tremain, "The government standard linear predictive coding algorithm: Lpc-10," *Speech Technology*, pp. 40–49, April 1982.

[8] L. Deng and D. O'Shaughnnessy, *Speech processing: A Dynamic and Optimization-Oriented Approach*. Marcel Dekker Inc., 2003.

[9] J.S. Karlsson, B. Gerdle, and M. Akay, "Analyzing surface myoelectric signals recorded during isokinetic contractions," *IEEE Engineering in Medicine and Biology Magazine*, no. 6, pp. 97–105, Nov-Dec 2001.

[10] M. Chester, *Neural Networks: A Tutorial*. PTR Prentice-Hall Inc., 1993.

[11] A. Breen, "Speech synthesis models: a review," *Electronics & communication engineering journal*, pp. 19–31, February 1992.