Non-acoustic Communication with Speech Smoothing

Yuet Ming Lam

Faculty of Information Technology,
Macau University of Science and Technology.
ymlam@must.edu.mo

Abstract *This paper presents a technique to synthesize speech from SEMG signals using a frame-by-frame basis. SEMG signals are firstly enframed and classified into a number of phonetic classes by a neural network, then the produced sequences of phonetic indices are translated to acoustic signals by concatenating their corresponding pre-recored speech segments. A significant advantage of the proposed synthesis based approach compared with previous recognition based approach is that, human is intelligent enough to recognition the synthesized speech although there is errors in it. Experimental evaluations based on the synthesis of eight words show that on average over 73.4% of the words can be synthesized correctly and the neural network can classify the SEMG frames of seven phonemes at a rate of 81.9%. The accuracy can be increased to 88.6% by using a glitch removal technique to smooth the produced sequence of phonetic indices. The results show that the phoneme-frame based speech synthesis technique can be applied to SEMG-based non-acoustic communication.*

Keywords *Non-acoustic communication, surface electromyogram signals, neural network.*

1. Introduction

Speech is the most natural way of self-expression and communication among humans. The speech production process involves the contraction of lungs to produce an air stream, the vibration of vocal cords to produce voiced excitation, and the resonance of the air stream in the vocal tract. However, there are situations in which communication via 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, because vocal cord vibration plays a vital role in the speech production process. A noisy environment may also degrade the quality of the produced speech and results in lower intelligibility. Moreover, using speech may be undesirable in some situations, e.g. in the military operations.

To address some of these limitations of speech communication, non-acoustic communication systems that using surface electromyogram [1] signals to recognize speech have been proposed. Some proposed systems focus on recognizing words from isolated SEMG signals [2] [3] [4] [5] [6] [7]; these systems demonstrate the feasibility of using SEMG signals to improve the performance of conventional speech recognition systems in some noisy environments [2] [8]. The improvement is mainly due to the fact that SEMG signals are immune to acoustic noise. It has also been demonstrated that SEMG-based speech recognition is applicable to computer-human

DOI : 10.5121/sipij.2012.3101

interface [9] [10] [11]. SEMG-based phoneme recognition is presented in [12] [13] [14] [15] [16] [17], but the recognition is done by regarding each phoneme as an isolated word.

The previously proposed SEMG-based speech recognition systems show the feasibility of recognizing speech based on SEMG signals. However, most of the proposed methodologies focus on classifying the SEMG signals into a limited set of words. These approaches share similarities with isolated-word speech recognition systems in that there must be sufficient silence intervals before and after the speech signals, i.e., the words must be labeled and isolated from each other. These word recognition systems have difficulties in recognizing continuous speech, and the recognition accuracy depends largely on the word duration. To overcome this limitation, Sugie et al. [18] proposed to recognize phonemes in a frame-by-frame basis. Where SEMG signals are blocked into frames and each frame is classified into one of the five vowels or silence. Using this approach, labeling and isolation are not required. However, the recognition accuracy is low because the recognition is based on the active/inactive states of the SEMG channels. Moreover, the feasibility of applying this methodology to word recognition is not addressed.

This paper proposes to synthesize speech, including phonemes, words or even sentences, directly from SEMG signals using a phoneme-frame based feature extraction and conversion approach. In particular, the training data set consist of phonemes only, words are synthesized to evaluate the performance. To synthesize a word, features are extracted from frame blocked SEMG signals and classified into a number of phonetic classes, the classification is done by a neural network which is trained using features extracted from parallel recorded SEMG and speech signals when pronounce phonemes, the produced sequence of phonetic class number are mapped to acoustic signals by concatenating corresponding pre-recorded speech. Because the features are extracted from phonemes and conversion is done at the frame level, the proposed method is potentially applicable to continuous speech synthesis. Moreover, a significant advantage of the proposed synthesis based approach compared with previous recognition based approach is that, human is intelligent enough to recognition the synthesized speech although there is errors in it. The contributions of this paper are as follows:

- A phoneme-frame based technique to synthesize speech from SEMG signals which can potentially achieve unlimited vocabulary.
- An analysis of the SEMG features and a feature reduction scheme.
- A glitch removal technique to improve the classification accuracy of phoneme frames.

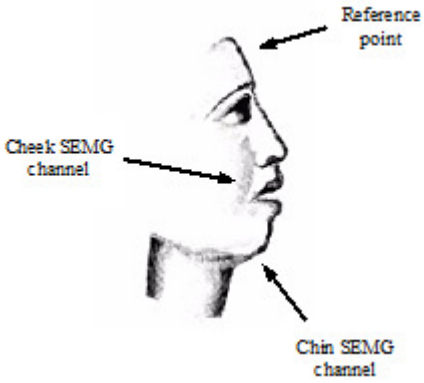

Figure 1: Electrode placement: One channel is from the cheek, the other is from the chin, an additional electrode is attached to the forehead as a reference point.

2 Methodology

2.1 Data Acquisition

Two channels of SEMG signals are collected as shown in Figure 1. The first channel is obtained from the cheek, which is about 2.5cm from the nose. The second channel is obtained from the chin. An additional electrode is attached to the forehead as a reference point. The SEMG signal is amplified with a gain of 1000. Both the amplified SEMG signal and speech are recorded concurrently using a PCI data acquisition card [19] at a sampling rate of 8000Hz.

The training data set consists of data samples of seven phonemes: *ae, iy, ao, uw, sh, f* and *s*. Each data sample is recorded in a twenty-second period, during which the speaker repeatedly pronounces one of the seven phonemes. The training data set for each phoneme is comprised of the SEMG and speech signals of four such data samples. Since SEMG signals from each channel were blocked into frames every 22.5ms, 24864 SEMG frames of each channel are involved in training.

Both phonemes and words are used for testing. The words are *shaw, she, ash, shoe, see, saw, fee* and *off*, whose phonetic transcriptions are formed by concatenating the seven phonemes. The testing data set consists of one data sample of each phoneme (6216 frames) and one data sample of each word (64 words). The phoneme and word samples are used to evaluate the performance of the neural network classification and the accuracy of speech synthesis respectively. The recorded speech is used as a reference for performance evaluation.

2.2 Speech Feature Selection

The input speech is blocked into 22.5ms frames, and there is no overlapping between frames. This scheme has been used in speech coding standard [20]. For each speech frame, ten linear predictive (LP) coefficients, pitch and root mean square value are extracted and concatenated as speech feature vector.

Table I: STFTC number and corresponding frequency region.

STFTC number	Frequency region
STFTC 1	1Hz – 45Hz
STFTC 2	46Hz – 90Hz
STFTC 3	91Hz – 135Hz
STFTC 4	136Hz – 180Hz
STFTC 5	181Hz – 225Hz
STFTC 6	226Hz – 270Hz
STFTC 7	271Hz – 315Hz
STFTC 8	316Hz – 360Hz
STFTC 9	361Hz – 405Hz
STFTC 10	406Hz – 450Hz

2.3 SEMG Feature Selection

The SEMG signals for each channel are blocked into frames, and for each SEMG frame, the short-time Fourier transform [21] coefficients (STFTCs), root mean square value (RMSV), and zero-crossing rate (ZCR) are extracted and used as SEMG features.

• STFTC: The frequency spectra from 1Hz to 450Hz are calculated from each SEMG frame and divided into ten equal frequency sections; the bandwidth of each section is 45Hz. The frequency components in each section are summed to give one STFTC corresponding to that section, which result in ten STFTCs. The STFTC numbers and their corresponding frequency regions are shown in Table I. Totally, twenty STFTCs are extracted from the two SEMG channels.

• RMSV: The root mean square value is calculated for each SEMG frame according to the following equation:

$$RMSV = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x(i)^2} \tag{1}$$

where $x(i)$ is the ith SEMG sample within the frame and N is the frame length. Two RMSVs are extracted from the two SEMG channels.

• ZCR: The zero-crossing rate is known as the number of time-domain zero-crossing within a particular duration of signal, divided by the length of that duration. Two ZCRs are extracted from the two SEMG channels.

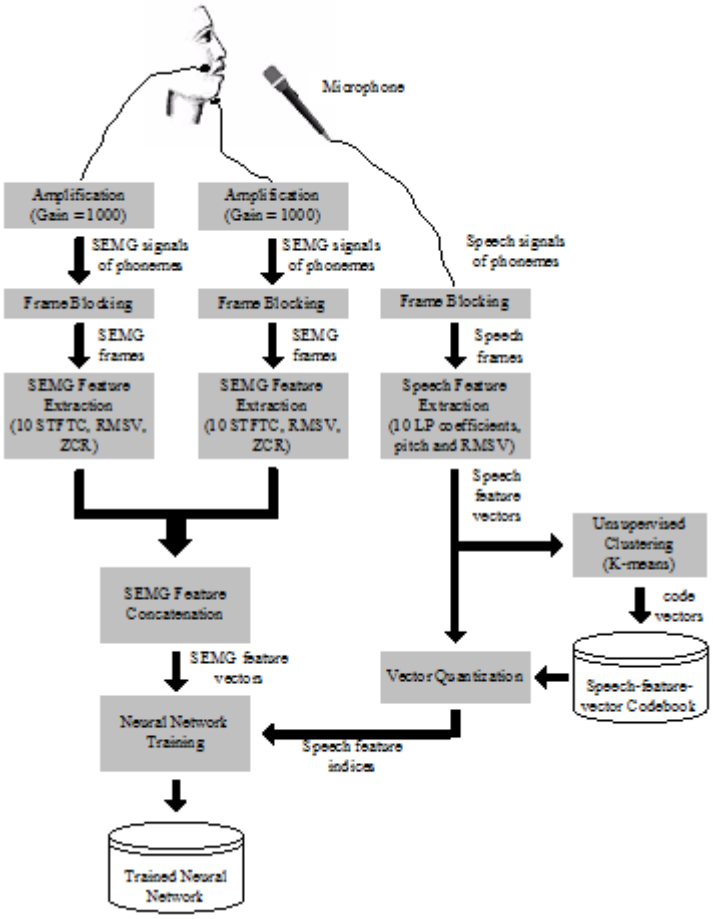

Figure 2: Phoneme-based feature extraction and neural network training: Concurrently recorded speech and SEMG signals are blocked into frames. Speech feature vectors extracted from speech frames are quantized using a speech-feature-vector codebook to give one speech-feature index for each speech frame. The SEMG feature vectors extracted from SEMG frames are paired with corresponding speech feature indices to form input-target training pairs. The neural network is trained using these input-target pairs.

This SEMG feature selection scheme leads to twenty-four features, including twenty STFTCs, two RMSVs and two ZCRs. To analysis the effect of different frequency regions on recognition performance, the symmetric divergences [22] between different STFTCs are calculated. Symmetric divergence is a separability measure of two distributions, and the divergence between two classes, ω_1 and ω_2, is calculated as follows:

$$DIV_{12} = \frac{1}{2} tr(\sum_1^{-1} \sum_2 + \sum_2^{-1} \sum_1 - 2I) + \frac{1}{2}(\mu_1 - \mu_2)'(\sum_1^{-1} + \sum_2^{-1})(\mu_1 - \mu_2) \qquad (2)$$

where μ_1 and $\sum_1$ are the mean and covariance of class ω_1, and μ_2 and $\sum_2$ are the mean and covariance of class ω_2. The average divergence is calculated as follows:

$$DIV_{avg} = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} DIV_{ij}}{\sum_{i=1}^{N-1} i} \tag{3}$$

where N is the number of classes. In this paper, N is equal to 8, which includes silence and 7 phonemes.

The SEMG frame size should be chosen carefully, because it affects the frequency resolution [23]. 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 size can improve the frequency resolution, but results in a loss of information between adjacent frames. In this paper, correlation between frame size and performance is analyzed.

2.4 Neural Network Training

As shown in Figure 2, the speech signals for the training phonemes are blocked into frames, and the LP coefficients, pitch, and root mean square value are extracted and concatenated to form speech feature vectors. As shown by the dash line in the figure, unsupervised clustering, based on the K-means algorithm, is used to extract the representative feature vectors for the phonemes and silence. The extracted feature vectors form a speech-feature-vector codebook.

After forming the speech-feature-vector codebook, the training vectors for the neural network can be constructed as shown in Figure 2. It is noted that only phonemes are involved in training. The speech signals of the training phonemes are blocked into frames and the extracted speech feature vectors are quantized using the speech-feature-vector codebook. Thus, each speech frame is represented by a codebook index. Because the codebook is formed by the representative speech feature vectors, the speech feature index indicates to which phoneme a speech frame belongs. The SEMG signals of training phonemes are also blocked into frames, and the STFTCs, RMSVs, and ZCRs are extracted from two SEMG channels and concatenated to form an SEMG feature vector. Each of the concatenated SEMG feature vectors is paired with the corresponding speech feature index to form an input-target training pair. A three-layer MLP (Multilayer Perceptrons) [24], which take an SEMG feature vector as input and produces one of eight possible speech codebook indices (silence and seven phonemes) as output, is trained using the input-target pairs.

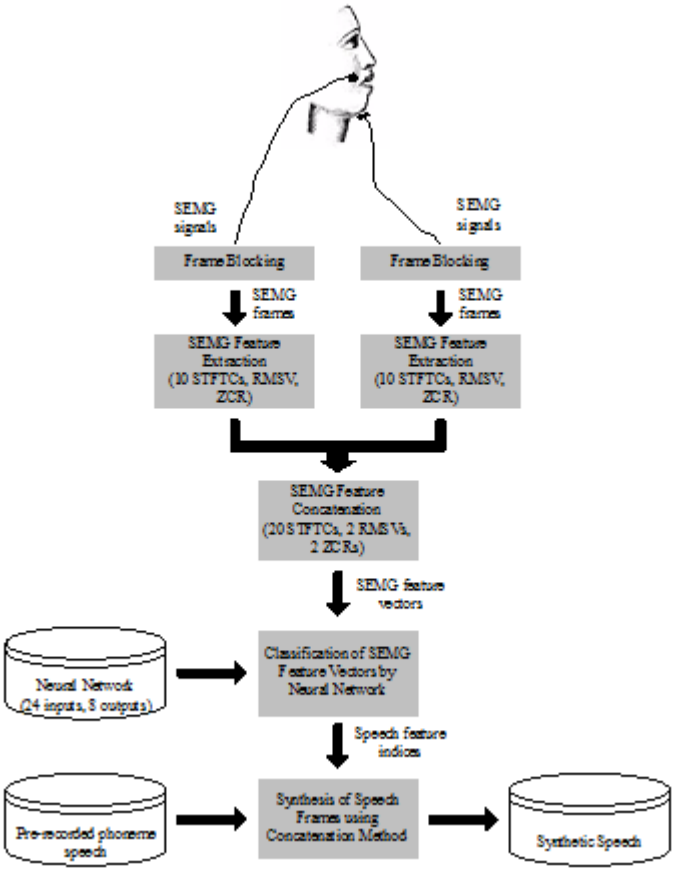

Figure 3: Speech synthesis process: SEMG signals recorded from both channels are blocked into frames, and SEMG feature vectors are extracted. The trained neural network is then used to classify these vectors into a sequence of speech feature indices. A concatenation synthesis method is applied to reconstruct the target speech

2.5 Speech Synthesis and Glitch Removal

In addition to synthesizing phonemes, the SEMG-based synthesis method proposed in this paper can also be applied to synthesize words as shown in Figure 3. To this end, SEMG signals recorded are blocked into frames, the features from the cheek and chin channels are concatenated to form SEMG feature vectors. Then the neural network is used to classify the concatenated SEMG feature vector into one of the seven phonemes or silence, which results in a sequence of speech feature indices for each word to be synthesized.

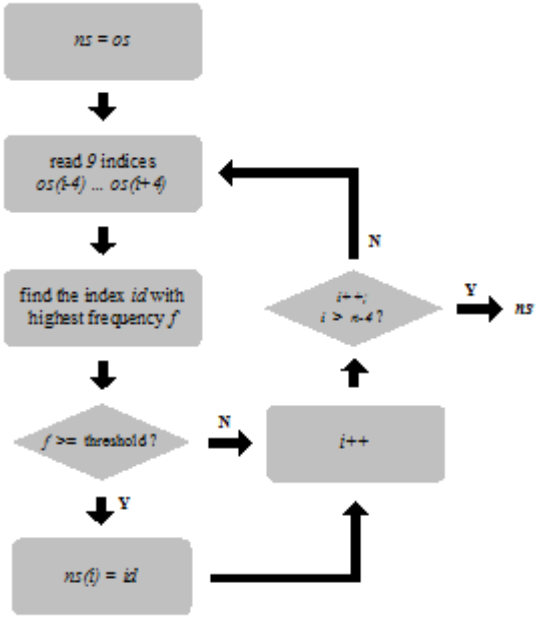

Figure 4: The glitch removal technique used to smooth the sequence of phonetic indices, where *os* is the sequence produced by the neural network, *ns* is the smoothed sequence, *n* is the sequence length of *os*. The threshold used in this work is 7.

The error rate of the produced sequence of speech feature indices can be improved by using a majority-filtering glitch removal technique. The idea is based on the observation that voiced speech signals are fairly stationary over a short period of time; in contrast, characteristics of the signal change over long periods of time, i.e. on the order of 200ms or more [25]. This technique scans the produced sequence of speech feature indices over a window of 9 indices (i.e. 202.5ms) with step 1, the index with the highest frequency within the window is found, and a new index equals the index found is produced if the frequency exceeds a threshold (Figure 4). In this work, a threshold of 7 is used.

After performing glitch removal on the sequence of speech feature indices, a concatenation synthesis method [26] is applied to reconstruct the target speech in a frame-by-frame basis. Based on the speech feature indices, target phoneme frames are loaded from the pre-recorded set and concatenated to form the complete speech. The transition between phonemes is smoothed using overlap and add method.

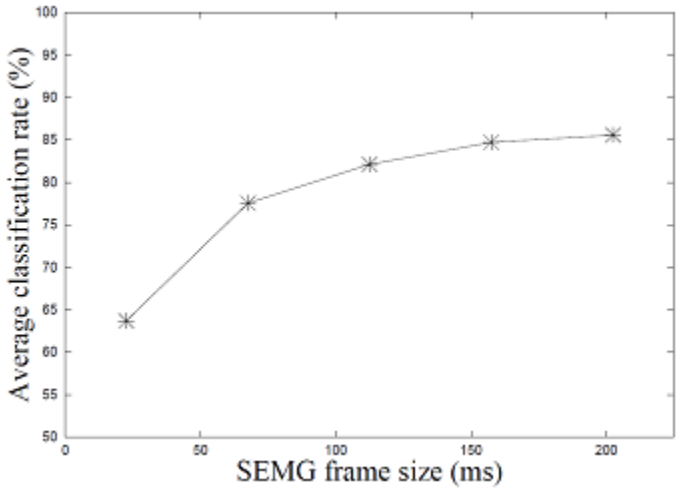

Figure 5: Average classification rate for SEMG feature vectors (20 STFTCs, 2 RMSVs, 2 ZCRs) for different SEMG frame sizes

Table II: Confusion matrix showing the classification performance based on 10 STFTCs, RMSV, and ZCR extracted from the cheek channel. The SEMG frame size is 112.5ms.

Classified phoneme label	True phoneme label							
	Silence	*ae*	*iy*	*ao*	*uw*	*sh*	*f*	*s*
Silence	82.7%	3.5%	11.7%	4.4%	2.4%	13.9%	3.3%	25.6%
ae	1.2%	91.2%	0.0%	4.6%	6.5%	1.5%	4.0%	0.0%
iy	4.3%	0.3%	66.8%	1.0%	1.4%	0.2%	2.1%	2.6%
ao	2.9%	0.9%	4.8%	52.1%	4.1%	1.1%	11.3%	0.0%
uw	1.4%	3.8%	0.0%	6.0%	64.5%	35.2%	8.9%	0.0%
sh	0.8%	0.0%	0.0%	0.5%	12.6%	45.7%	3.3%	0.0%
f	1.7%	0.3%	0.0%	31.4%	8.5%	2.1%	66.9%	0.0%
s	5.0%	0.0%	16.7%	0.0%	0.0%	0.3%	0.2%	71.8%

Table III: Confusion matrix showing the classification performance based on 10 STFTCs, RMSV, and ZCR extracted from the chin channel. The SEMG frame size is 112.5ms.

Classified phoneme label	True phoneme label							
	Silence	*ae*	*iy*	*ao*	*uw*	*sh*	*f*	*s*
Silence	80.5%	4.7%	4.7%	3.9%	5.3%	14.2%	9.4%	21.6%
ae	3.2%	54.0%	17.1%	1.0%	21.1%	1.1%	24.6%	8.9%
iy	1.0%	2.4%	48.0%	24.4%	4.1%	22.0%	0.2%	0.0%
ao	1.1%	2.6%	8.6%	52.6%	2.4%	19.6%	0.2%	0.0%
uw	2.8%	22.6%	4.5%	0.9%	22.1%	6.0%	9.6%	2.5%
sh	1.3%	5.2%	13.1%	16.4%	2.0%	36.2%	0.0%	0.0%
f	5.0%	1.6%	2.6%	0.3%	9.7%	0.6%	16.3%	16.6%
s	5.1%	6.9%	1.4%	0.5%	33.3%	0.3%	39.7%	50.4%

Table IV: Confusion matrix showing the classification performance based on 20 STFTCs, 2 RMSVs, and 2 ZCRs extracted from both channels. The SEMG frame size is 112.5ms.

Classified phoneme label	True phoneme label							
	Silence	*ae*	*iy*	*ao*	*uw*	*sh*	*f*	*s*
Silence	87.7%	3.5%	7.4%	4.9%	3.2%	14.6%	9.2%	12.2%
ae	1.7%	91.6%	0.0%	0.0%	6.5%	0.6%	4.2%	0.0%
iy	1.4%	0.0%	78.3%	0.0%	0.0%	0.0%	0.0%	1.4%
ao	1.5%	0.0%	3.6%	91.5%	0.0%	1.1%	0.2%	0.2%
uw	1.2%	4.0%	0.0%	0.2%	73.5%	6.0%	14.1%	0.0%
sh	0.9%	0.2%	0.0%	3.4%	2.4%	76.3%	0.5%	0.0%
f	1.7%	0.7%	0.0%	0.0%	14.0%	1.1%	71.6%	0.0%
s	3.9%	0.0%	10.7%	0.0%	0.4%	0.3%	0.2%	86.2%

Table V: Confusion matrix showing the classification performance after feature reduction, which uses 10 STFTCs, RMSV, and ZCR extracted from the cheek channel and STFTC 1 to 5, RMSV, and ZCR extracted from the chin channel.

Classified phoneme label	True phoneme label							
	Silence	*ae*	*iy*	*ao*	*uw*	*sh*	*f*	*s*
Silence	88.6%	3.8%	7.1%	4.8%	3.7%	15.1%	6.8%	14.6%
ae	1.3%	91.5%	0.0%	0.0%	5.7%	0.6%	4.9%	0.0%
iy	1.2%	0.0%	75.7%	0.0%	0.0%	0.0%	0.2%	2.0%
ao	1.5%	0.0%	4.3%	92.3%	0.0%	1.7%	0.2%	0.0%
uw	1.1%	3.6%	0.0%	0.0%	73.6%	6.0%	13.1%	0.0%
sh	1.0%	0.2%	0.0%	2.9%	2.0%	75.6%	0.0%	0.0%
f	1.4%	0.7%	0.0%	0.0%	14.8%	0.8%	74.6%	0.0%
s	3.9%	0.0%	12.9%	0.0%	0.2%	0.2%	0.2%	83.4%

3.Experiments and Results

3.1 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 is analyzed. Classification is done using SEMG feature vectors contains 20 STFTCs, 2 RMSVs, and 2 ZCRs, the neural network is trained using the training data set and performance is evaluated using the phonemes in the testing data set. The average classification rates for SEMG frame sizes from 22.5ms to 202.5ms are shown in Figure 5. 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. Because smaller frame size gives better time resolution, frame size of 112.5ms is chosen for further experiments despite larger frame size gives slightly higher classification rate.

3.2 Sensor Positioning

To further analyze the correlation between sensor position and performance, the neural network is trained using SEMG feature vectors extracted from a single channel, and classification performance is evaluated using single-channel SEMG signals of the testing phoneme set. Table II shows the confusion matrix for classification using 10 STFTCs, RMSV, and ZCR extracted from the cheek, and the results obtained for chin channel are shown in Table III. In these tables, the rows show the classified labels found by the neural network and the columns represent the true labels. The average classification rates of the cheek and chin channels are 67.7% and 45.0% respectively. Some phonemes, such as *s* and silence have similar characteristics, as we can see that nearly 30% SEMG frames of *s* are misclassified as silence in both tables. And 35.2% of *sh* are misclassified as *uw* when using the cheek channel; but it is only 6.0% when the chin channel is used. On average, using cheek channel is better. therefore, we can conclude that the cheek channel provides more discriminative information for phoneme classification.

The classification results based on 20 STFTCs, 2 RMSVs, and 2 ZCRs extracted from both channels are shown in Table IV. The average classification rate is 82.1%, which is better than using a single channel.

3.3 Assessment of STFTCs and Feature Reduction

Figure 6 shows the amplitude distribution of STFTCs obtained from the cheek and chin channels. One can see from Figure 6 that some phonemes such as *s* and silence, *sh* and *uw* have similar characteristics. On the other hand, some phonemes are more separable, e.g. *ae* and *iy*, *uw* and silence.

To measure the separability of phonemes, divergence test is employed. Figure 7 shows the scaled average divergence scores for STFTCs obtained from the cheek and chin channels. The average divergence scores are calculated using Equation 3. We can see from Figure 7 that STFTCs are more distinguishable in low frequency regions, and using the cheek channel is better than the chin channel in distinguishing these STFTCs. This explains the classification results obtained in Section 3.2. Similar characteristic can also be observed in Figure 8, where the scaled average divergence score for sub-STFTC vectors obtained from the cheek and chin channels is shown. The sub-STFTC vectors are formed by selecting STFTCs in Table I. This figure shows that the divergence scores saturate when the number of STFTCs increase, this phenomenon is especially

noticeable for the chin channel, which becomes saturated when number of STFTCs is larger than 5. This analysis suggests that STFTC 6 (226-270Hz) to STFTC 10 (406-450Hz) of the chin channel can be removed.

Classification is done using the reduced features consisting of all STFTCs of the cheek channel, STFTCs 1 to 5 of the chin channel, and all RMSVs and ZCRs from both channels. The results are show in Table V, and the average classification rate is 81.9%. It is almost the same as the average classification rate using all SEMG features extracted from both channels, which is 82.1%.

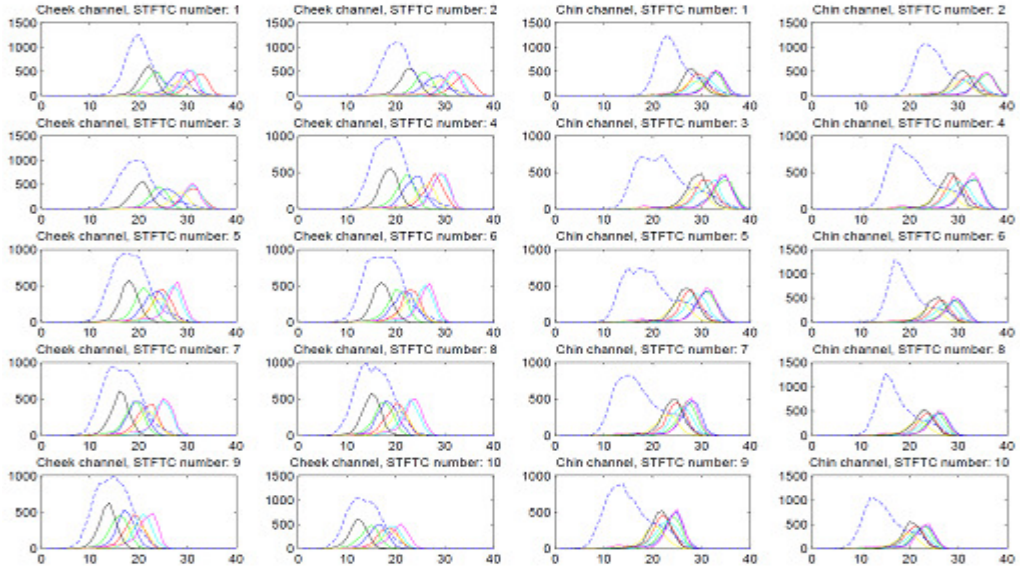

Figure 6: Distribution of STFTCs obtained from cheek and chin channels. The horizontal axis is the scaled STFTC amplitude, and the vertical axis is the number of occurrences. Red solid line: *ae*; green solid line: *iy*; blue solid line: *ao*; cyan solid line: *uw*; magenta solid line: *sh*; yellow solid line: *f*; black solid line: *s*; blue dash line: silence.

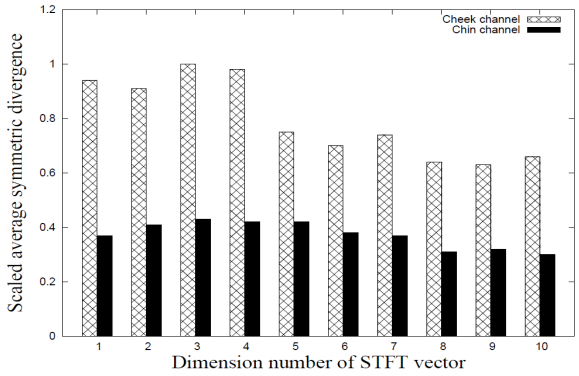

Figure 7: Scaled average divergence scores for STFTCs 1 to 10.

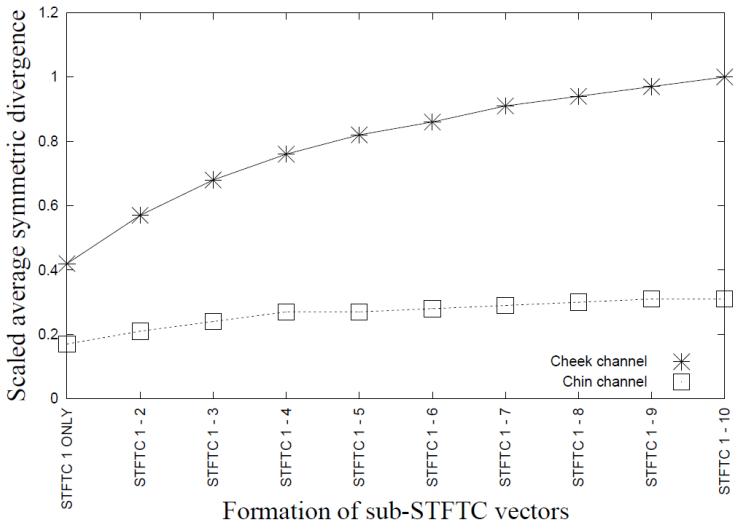

Figure 8: Scaled average divergence scores for sub-STFTC vectors obtained from the cheek and chin channels. The vertical axis is the average divergence score. The horizontal axis is the number of STFTCs involved to obtain a sub-STFTC vector, e.g. STFTC 1 - 3 means that the sub-STFTC vector consists of STFTC 1 (1-45Hz) to STFTC 3 (91-135Hz).

Table VI: Confusion matrix after applying the glitch removal technique to the produced speech feature indices based on reduced SEMG features.

Classified Classification	True phoneme label							
	Silence	*ae*	*iy*	*ao*	*uw*	*sh*	*f*	*s*
Silence	93.8%	3.6%	9.3%	7.3%	4.5%	19.0%	21.1%	5.5%
ae	1.0%	96.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
iy	0.9%	0.0%	83.3%	0.0%	0.0%	0.0%	0.0%	1.3%
ao	0.7%	0.0%	0.0%	92.7%	0.0%	2.1%	0.0%	0.0%
uw	0.6%	0.0%	0.0%	0.0%	91.5%	0.0%	0.0%	0.0%
sh	0.7%	0.0%	0.0%	0.0%	0.0%	78.9%	0.0%	0.0%
f	0.8%	0.0%	0.0%	0.0%	4.0%	0.0%	78.9%	0.0%
s	1.5%	0.0%	7.4%	0.0%	0.0%	0.0%	0.0%	93.2%

3.4 Glitch removal

The average classification accuracy based on reduced SEMG feature is 81.9% (Table V). The glitch removal process is then applied to correct misclassification errors. The results after applying the glitch removal are shown in Table VI. Although more voiced SEMG feature vectors are misclassified as silence vectors in some phonemes, e.g. *uw* and *sh*, the overall classification rates for all phonemes are improved and the average classification rate is improved to 88.6%. A summary of average classification rates from Table II to VI is shown in Table VII.

Table VII: A summary of average classification rates.

Table Number	Description	Average Classification Rate
II	Classification using 10 STFTC, RMSV, ZCR from the Cheek channel	67.7%
III	Classification using 10 STFTC, RMSV, ZCR from the Chin channel	45.0%
IV	Classification using 20 STFTCs, 2 RMSVs, 2 ZCRs from the both channels	82.1%
V	Classification using both channels after feature reduction	81.9%
VI	Classification using both channels with feature reduction and glitch removal	88.6%

Table VIII: Synthesis result for words.

Words	Number of words involved	Number of words synthesized correctly
she	8	6
ash	7	7
shaw	9	8
see	8	6
saw	9	6
shoe	8	4
fee	8	5
off	7	5
Total	64	47

3.5 Speech synthesis

Words are synthesized using the reduced features obtained from both channels and the glitch removal technique. In particular, SEMG feature vectors formed by concatenating 10 STFTCs, RMSV, ZCR from the cheek channel and STFTC 1 to STFTC 5, RMSV, ZCR from the chin channel, and presented to the neural network to produce the speech feature indices. Error correction is applied to the resulting sequence of indices and words are synthesized by the concatenation method. One twenty-second sample of each word is used in the experiment. Table VIII is the results obtained, which shows that the percentage of words correctly synthesized is 73.4%. A word is regarded as synthesized correctly if the phonetic transcriptions of the synthesized word match the reference word, e.g. a synthesized word *she* is regarded as synthesized correctly if its phonetic

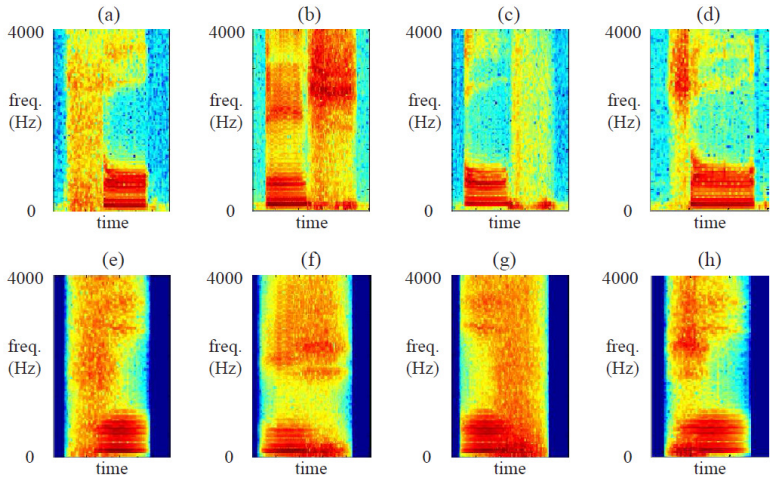

Figure 9: Spectrograms of four synthesized instances: (a) - (d) are the reference speech of *saw*, *ash*, *off* and *shaw*, respectively, and (e) - (h) are the corresponding synthesized speech.

transcriptions is a phoneme *sh* followed by a phoneme *iy*. Figure 9 shows the spectrograms of four synthesized instances. One can see that the synthesized instances and the reference speech have similar characteristics despite the words are not involved in the training process. Although the synthesized instances and the reference speech may not align perfectly, e.g. some silence frames before the reference word *off* are synthesized as phoneme *ao* as shown in sub-figures (c) and (g) of Figure 9, it looks like the synthesized word *off* is left shifted with several frames, the intelligibility of the synthesized instances is not affected. Currently, a simple speech synthesis model is used, and this paper is focusing on how to produce correct phonetic transcriptions, because it is directly correlated to the quality and intelligibility of the resulting speech. It is believed that the quality can be improved if a more sophisticated synthesis model is used.

4 Conclusions

A frame based speech synthesis technique using SEMG signals is presented. It is found that a frame size of 112.5ms can provide a good balance between time and frequency resolution. The quantitative assessment shows that the spectral features of SEMG signals are more distinguishable in the low frequency regions. It is also found that cheek channel provides more useful information for classifying SEMG signals and the features can be reduced with slight performance degration. The performance can be further improved by removing glitches in the produced index sequences. Experimental results show that words can be synthesized from SEMG signals using the proposed frame-based feature extraction and conversion methodology.

References

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

[2] 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, 2001, vol. 2, pp. 1727–1739.

[3] 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, 2003, vol. 4, pp. 3128–3133.

[4] H. Manabe and Z. Zhang, "Multi-stream HMM for EMG-Based Speech Recognition," in Proceedings of the 26rd Annual International Engineering in Medicine and Biology Society, 2004, vol. 2, pp. 4389–4392.

[5] K.S. Lee, "EMG-Based Speech Recognition Using Hidden Markov Models With Global Control Variables," IEEE Transactions on Biomedical Engineering, vol. 55, no. 3, pp. 930–940, 2008.

[6] G. Colby, J. T. Heaton, L. D. Gilmore, J. Sroka, Y. Deng, J. Cabrera, S. Roy, C. J. De Luca, and G. S. Meltzner, "Sensor Subset Selection for Surface Electromyograpy Based Speech Recognition," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 473–476.

[7] M. Wand and T. Schultz, "Speaker-Adaptive Speech Recognition Based on Surface Electromyography," Communications in Computer and Information Science, vol. 52, no. 3, pp. 271–285, 2010.

[8] B.J. Betts and C. Jorgensen, "Small Vocabulary Recognition Using Surface Electromyography in an Acoustically Harsh Environment," Interacting with Computers, vol. 18, no. 6, pp. 1242–1259, 2006.

[9] 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.

[10] S.P. Arjunan, H. Weghorn, D.K. Kumar, and W.C. Yau, "Vowel Recognition of English and German Language Using Facial Movement(SEMG) for Speech Control Based HCI," in Proceedings of the HCSNet Workshop on Use of Vision in Human-Computer Interaction, 2006, pp. 13–18.

[11] H. Liu, H. Ding, Z. Xiong, and X. Zhu, "Multi-modality: EMG and Visual Based Hands-Free Control of an Intelligent Wheelchair," in Proceedings of the Third international conference on Intelligent Robotics and Applications, 2010, pp. 659–670.

[12] S. Kumar, D.K. Kumar, M. Alemu, and M. Burry, "EMG Based Voice Recognition," in Proceedings of the Intelligent Sensors, Sensor Networks and Information Processing Conference, 2004, pp. 593–597.

[13] M. Walliczek, F. Kraft, S. C. Jou, T. Schultz, and A. Waibel, "Sub-Word Unit based Non-Audible Speech Recognition using Surface Electromyography," in Proceedings of the International Conference on Spoken Language Processing, 2006, pp. 1596–1599.

[14] E. J. Scheme, B. Hudgins, and P.A. Parker, "Myoelectric Signal Classification for Phoneme-Based Speech Recognition," IEEE Transactions on Biomedical Engineering, vol. 54, no. 4, pp. 694–699, 2007.

[15] J.A.G. Mendes, R.R. Robson, S. Labidi, and A.K. Barros, "Subvocal Speech Recognition Based on EMG Signal Using Independent Component Analysis and Neural Network MLP," in Proceedings of the Congress on Image and Signal Processing, 2008, pp. 221–224.

[16] E. Lopez-Larraz, O.M. Mozos, J.M. Antelis, and J. Minguez, "Syllable-Based Speech Recognition Using EMG," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2010, pp. 4699–4702.

[17] C. Jorgensen and S. Dusan, "Speech Interfaces Based Upon Surface Electromyography," Speech Communication, vol. 52, no. 4, pp. 354–366, 2010.

[18] N. Sugie and K. Tsunoda, "A Speech Prosthesis Employing a Speech Synthesizer - Vowel Discrimination from Perioral Muscle Activities and Vowel Production," IEEE Transactions on Biomedical Engineering, vol. 32, no. 7, pp. 485–490, July 1985.

[19] National Instruments Inc., 6023E/6024E/6025E User Manual, December 2000 Edition.

[20] T.E. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC-10," Speech Technology, pp. 40–49, April 1982.

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

[22] S. Theodoridis and K. Koutroumbas, Pattern Recognition, 2nd Edition, Academic Press, 2003.

[23] J.S. Karlsson, B. Gerdle, and M. Akay, "Analyzing Surface Myoelectric Signals Recorded During Isokinetic Contractions," IEEE Engineering in Medicine and Biology Magazine, vol. 20, no. 6, pp. 97–105, Nov-Dec 2001.

[24] M. Minsky and S. Papert, Eds., Perceptrons: An Introduction to Computational Geometry, Cambridge. MA: The MIT Press, 1969.

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

[26] A. Breen, "Speech Synthesis Models: A Review," Electronics & Communication Engineering Journal, vol. 4, pp. 19–31, February 1992.