IMPROVED PHONEME-BASED MYOELECTRIC SPEECH RECOGNITION

Quan Zhou, Ning Jiang, Englehart Kevin and Hudgins Bernie
University Of New Brunswick, Fredericton, N.B., Canada

***Abstract*— This paper introduces an adaptive phoneme-based multi-expert speech recognition system using the myoelectric signal (MES). The MES produced by the speaker's facial muscles can be used as another expert system to enhance recognition accuracy in noisy situations. In previous work, ten words are recognized by phoneme-based classifier. In the current study, an expanded set of words has been classified phonemically by an HMM classifier trained at the phoneme level using a subset of all the words. The raw MES signals are rotated by class-specific rotation matrices to spatially decorrelate the measured data prior to feature extraction. In a post-processing stage, an uncorrelated linear discriminant analysis (ULDA) is used for dimensionality reduction. The resulting data are classified through an HMM classifier to obtain the phonemic log likelihoods, which are mapped to corresponding words using an artificial neural network. It is shown that these methods provide a recognition accuracy of 89% when classifying an expanded lexicon containing the same phonemes as the ones used by the training set. As a result, the new words are recognized from the phoneme structure without retraining the HMM classifier.**

INTRODUCTION

The performance of traditional Acoustic Speech Recognition (ASR) system tends to deteriorate with high ambient noise levels. Therefore, a single expert ASR system cannot be a practically functional alternative control technology in an aircraft flight cockpit. Since the 1980s, research has been done to extract speech information from myoelectric signals (MESs) during speech. Initial studies showed that the classification accuracy of MES systems always stayed above the a priori classification accuracy, demonstrating the presence of speech information in the MES [1] [2]. To enhance the performance of the ASR system, specifically in noisy environments, Chan et al. [3] proposed a multi-expert automatic speech recognition system combining both traditional ASR and MES classification. Classification accuracies remained above 78.8% across the 18-dB range of acoustic noise of a ten-word vocabulary. The improvement encouraged further research in the MES-ASR system. Scheme et al. [4] modified the structure of Chan's classification system and built a phoneme-based MES speech recognition system. The work improved the classification accuracy and more importantly allows the system to easily expand its lexicon. To explore this advantage further, the current study collected both acoustic and MESs from one subject while 20 words were spoken. The HMMs were trained using the data from 11 words, and tested using all data. It was shown that the performances of the MES expert deteriorated as more words were added in the testing set. This may due to increasing variance of the MESs and the limited training database. To increase the performance, preprocessing, dimensionality reduction, and post processing have been added to the original system. A schematic diagram of the whole system is presented in Figure 1. In the methodology section, a general introduction will be made for the three new components. The improvement of classification accuracy will be presented in the results section, and future improvements will be discussed in the conclusion section.

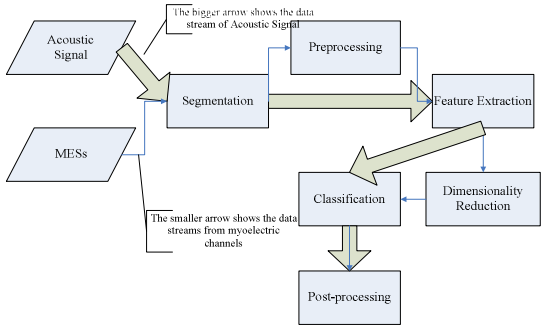

Figure 1: The block diagram of the whole system.

METHODOLOGY

Data Collection

One subject's data were collected for both an acoustic channel and MES channels. Five facial muscles were used for MES signal collection [3] [4]. Twenty words with a reasonably even distribution of 20 phonemes (listed in Table I) were chosen as the test set. A subset of 11 words containing all the 20 phonemes were picked as the training set. In the data collection experiment, a graphical user interface [4] was developed to show the words randomly and display the collected multi-channel signals in both the time and

frequency domains. Each word in the training set was repeated 50 times randomly, 60% of which were used in the training set while the remaining ones were used for the testing set.

Table II: Word list and phonemic breakdowns. The first 11 words in the table also appeared in the training set

Word	Phonemic Breakdowns			
zero	/z/	/i/	/r/	/ou/
one	/w/	/A/	/n/	
two	/t/	/u/		
three	/th/	/r/	/i/	
four	/f/	/o/		
five	/f/	/ai/	/v/	
six	/s/	/I/	/x/	
seven	/s/	/E/	/v/	/n/
eight	/ei/	/t/		
nine	/n/	/ai/		
east	/i/	/st/		
zoo	/z/	/u/		
very	/v/	/E/	/r/	/i/
throw	/th/	/r/	/ou/	
sweet	/s/	/w/	/i/	/t/
wait	/w/	/ei/	/t/	
sight	/s/	/ai/	/t/	
north	/n/	/o/	/th/	
west	/w/	/E/	/st/	
fox	/f/	/o/	/x/	

Preprocessing

In the preprocessing stage, a Principle Components Analysis (PCA)-based method was applied to sharpen the patterns of the MES data. “Principal Components Analysis (PCA) is a linear transformation that decorrelates multivariate data and projects it onto a new coordinate system so that the greatest variance in the data lies on the first coordinate; while the least variance in the data comes to lie on the last coordinate” [8]. However the classical PCA depends on the statistical properties of the common data distribution, while the class-conditional statistics are ignored [9]. Besides, the MES is engendered by a muscular contraction, which can be detected from the surface of skin. The surface EMG signal usually is the contribution of multiple muscles, especially when the movement is achieved by several small and closely spaced muscles like facial muscles. Occasionally subtle changes in some small or deep muscles can be masked by large or surface muscles and it is possible for these slight changes, associated with varying movements, to go undetected [7]. However, these understated changes may be key differences among the MESs from different phoneme-utterances. It is believed that this PCA-based data preprocessing methodology can decrease the effect of muscle crosstalk [7]. Therefore, to observe the subtle changes and to investigate further about within-class data distribution, the PCA-based preprocessing method was applied. Equivalent algorisms also have been employed in the facial image recognition and neural data processing for brain computer interface [9] [10]. Assume there are n_m samples in sum for all the training-utterances in m_{th} class. N represents the number of channels while M stands for the total number of classes.

$$X_m(t) = [x_{m1}(t), x_{m2}(t), \ldots x_{mN}(t)] \quad t \in 1,2,3,\ldots n_m$$

$$X_m = \begin{bmatrix} X_m(1) \\ X_m(2) \\ \vdots \\ X_m(n_m) \end{bmatrix} \quad m \in 1,2,\ldots M \qquad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix}$$

For each class X_m , a transformation matrix T_m was calculated. To obtain the total transformation matrix T , parallel all the T_m .

$$T_m = \left[t_{ij} \right]_{N \times N} \quad m \in 1,2,\ldots M \qquad T = \left[T_1, T_2, \ldots T_M \right]$$

Now the data of each phoneme-utterance is projected through T . S symbolizes the transformed data. n indicates the number of samples in each utterance.

$$S = XT = \left[s_{11}(t), s_{12}(t), \ldots s_{1N}(t), \ldots s_{M1}(t), \ldots s_{MN}(t) \right]$$

$$t \in 1,2,3,\ldots n$$

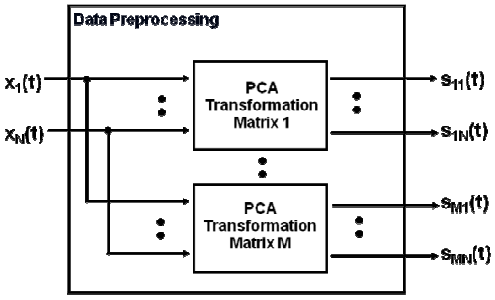

Figure 2: A block diagram showing the preprocessing block. Figure is reproduced with permission from [7].

The process can be clearly shown by Figure 2 [7]. It is hypothesized that if the data is projected through the PC transformation matrix, which is generated from the training data in the same class, will enhance or ‘tune’ the data while projection down the remaining PC transformation matrices will result in less meaningful linear combinations of the measured multivariate data [7].

Dimensionality Reduction

The high dimensional feature space has the potential to be much harder to be discerned by the classifier than the low dimensional feature. Known as the curse of dimensionality [11], a large number of samples are required to solve the above problem. However, that consumes time and effort on the part of the user. As a result, dimensionality reduction is necessary in our application. Uncorrelated Linear Discriminant Analysis (ULDA) was applied as the dimension reduction method. ULDA projects the data into the domain where the within-class distance is minimized and the between-class distance is maximized. Moreover, the optimal discriminant vectors in the transformation matrix are S-orthogonal [14]. Jieping [14] proposed a modified ULDA algorithm that can avoid the singularity problem in classical LDA and earlier ULDA methods. Using the training data, the transformation matrix G was calculated with singular value decomposition. Then the featured data was transformed by G and the dimension of the new feature set was minimized. The study [14] also compared classification accuracies with multiple dimensionality reduction methods including PCA, subspace, Orthogonal Centroid Method (OCM), and ULDA. The average classification accuracies, when ULDA was employed, achieved the maximum while processing different types of data. In our application, ULDA also decreased the dimension from 1300 to 19 and improved the classification performance.

Post Processing

Scheme [4] used maximum likelihood to make the phoneme classification decision from the output of HMMs. However, this method ignored the information contained in the reasonable phoneme combinations. Instead of choosing the phoneme represented by the highest score, we would like to find much more information contained in the possible phoneme combinations. As a result, a post processor based on an artificial neural network (ANN) was built to realize the mapping from the phoneme log likelihoods to the corresponding words. When a total neural network model was built for all the samples, some words were mistakenly mapped into other words even with a different number of phonemes. To increase the performance, separate neural network classifiers were built for words with different numbers of phonemes.

A three layer feedforward back propagation network was constructed for each classifier. Different numbers of neurons, non linear functions and training algorithms were tested to achieve the best performance. It was determined that for the two- or three-phoneme words the classifiers with 20 neurons in their hidden layers generated the best result. For the four-phoneme words the number of neurons in the hidden layer became 25, probably due to the larger dimension of the input. A sigmoid function was used in the hidden layer while the linear function was applied in the output layer. For each word there are only 10 samples in total, therefore, to avoid oscillatory outputs, leave-one-out cross validation (LOOCV) was combined with the ANN classifiers and an average performance was calculated.

RESULTS

Table II displays the result of phoneme-recognition accuracy and word-recognition accuracy of MES channels. It is obvious that the recognition accuracy of words is always much higher than that of phonemes. As a word constitutes several phonemes; even if some phonemes are mistakenly identified, the system can still find the correct word from the mapping classifiers. The first row shows the MESs-processing results of the original phoneme-based speech recognition system. The second row shows the MESs-processing results when the preprocessing and feature reduction are applied. The third row shows the results of MESs recognition using the post processor only. The fourth row displays the results of the combined improved phoneme-based speech recognition system.

Table II: results comparison

	Phoneme Accuracy	**Word Accuracy**
Original	43.93%	70.50%
Preprocessing, feature reduction	53.93%	75.00%
Post-processing	43.93%	73.00%
Improved system	53.93%	89.00%

The MES-recognition results of the original system decreases compared with Scheme's results mainly because of two reasons: first, the results were collected for the whole word set while the HMM classifier was trained with a subset of the 20 words. The variance of MESs of the same phoneme when it appears in different words and the limited training set caused performance decreasing. Second, since the system is speaker-dependent, the result will vary from person to person. The preprocessing and the dimensionality reduction improved the recognition results some degree. However, after applying the ANN mapping classifier, the accuracy was improved by 14%. This indicates the potential valuable information contained in the mapping process which is not used by the maximum

log likelihood-based mapping algorithm.

CONCLUSION

MES phoneme recognition improves upon previous word-based MES systems by enabling the addition of new words without further training. To demonstrate this advantage, 20 words were tested while only 11 words were used in training. The improved phoneme-based MES recognition achieved 89% accuracy when 9 additional words are included in testing, which implies that database expansion is not the only way to increase the performance of phoneme-based MES recognition. Applying the improved phoneme-recognition system also decreases the work load of database expansion, since the system can keep a good and robust performance under a certain ratio between the number of training words and testing words. For the original system this ratio will be higher, since more words need to be added into training to keep a reasonable accuracy. In the future, more words are expected to be tested determine how the performance is affected by changing the ratio between the number of trained words and tested words. For example, to keep the performance drop within 5%, while keeping the training set the same, how many words at most can be tested? Another important thing concerns the post-classifier. From the increase of performance after applying the ANN based post-classifier, we have shown that there is important information contained in possible phoneme combinations, which was missed by the original rule based mapping algorithm. However the ANN- based classifier needs to be trained when a new word is added. As more words are tested, the time of ANN-classifier training would be longer, which is not feasible in practical situations. Therefore, a more time-saving post-processor is needed in the future. Gaussian mixture model (GMM) may be applicable, since one GMM can be built for each word. In this way, whenever a new word is tested, only the new built model need to be trained.

ACKNOWLEDGEMENTS

This work is supported by grants from NSERC and the New Brunswick Innovation Foundation (NBIF).

REFERENCES

[1] M. S. Morse and E. M. O'Brien. (1986, Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes. Computers in Biology and Medicine, 16(6), pp. 399-410.

[2] N. 1. Sugie and K. 1. Tsunoda. (1985, 07). A speech prosthesis employing a speech synthesizer-vowel discrimination from perioral muscle activities and vowel production. IEEE Transactions on Biomedical Engineering BME-32(7), pp. 485-90.

[3] A. D. C. Chan. (2003, Multi-expert automatic speech recognition system using myoelectric signals. Available: http://proquest.umi.com/pqdweb?did=765339191&Fmt=7&clientId=65345&RQT=309&VName=PQD

[4] E. J. Scheme, B. Hudgins and P. A. Parker. (2007, 04). Myoelectric signal classification for phoneme-based speech recognition. IEEE Transactions on Biomedical Engineering 54(4), pp. 694-9. Available: http://dx.doi.org/10.1109/TBME.2006.889175

[5] C. 1. Furuichi, K. 1. Aizawa and K. 1. Inoue. (2000, Speech recognition using stochastic phonemic segment model based on phoneme segmentation. Syst. Comput. Jpn. 31(10), pp. 89-98. Available: http://dx.doi.org/10.1002/1520-684X(200009)31:10<89::AID-SCJ9>3.0.CO;2-7

[6] B. Ziolko, S. Manandhar and R. C. Wilson. Phoneme segmentation of speech. Presented at 18th International Conference on Pattern Recognition, ICPR 2006. Available: http://dx.doi.org/10.1109/ICPR.2006.931

[7] L. Hargrove, E. Scheme, K. Englehart and B. Hudgins. (2007, Principal components analysis preprocessing to reduce controller delays in pattern recognition based myoelectric control. Conf. Proc. IEEE Eng. Med. Biol. Soc. 1pp. 6511-6514.

[8] A. Hyvarinen, E. Oja, and J. Karhunen, Independent Component Analysis. New York : J. Wiley, 2001, pp. 481.

[9] Koel Das, Sergey Osechinskiy, and Zoran Nenadic, "A classwise PCA-based recognition of neural data for brain-computer interfaces," in 2007, pp. 6519-6520,6521,6522.

[10] K. Venkataramani, S. Qidwai and B. V. K. Vijayakumar. (2005, Face authentication from cell phone camera images with illumination and temporal variations. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews 35(3), pp. 411-418. Available: http://dx.doi.org/10.1109/TSMCC.2005.848183

[11] R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification. ,2nd ed.Wiley-Interscience, 2000,

[12] I. T. Jolliffe, Principal Component Analysis. Springer Series in Statistics, Berlin: Springer, 1986.

[13] W. J. Krzanowski, P. Jonathan, W. V. McCarthy and M. R. Thomas. (1995, Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data. Applied Statistics 44(1), pp. 101-115. Available: http://links.jstor.org/sici?sici=0035-9254%281995%2944%3A1%3C101%3ADAWSCM%3E2.0.CO%3B2-P

[14] J. Ye, R. Janardan, Q. Li and H. Park. (2006, Feature reduction via generalized uncorrelated linear discriminant analysis. IEEE Trans. Knowled. Data Eng. 18(10), pp. 1312-1322. Available: http://dx.doi.org/10.1109/TKDE.2006.160

[15] Z. Jin, J. Yang, Z. Hu and Z. Lou. (2001, Face recognition based on the uncorrelated discriminant transformation. Pattern Recognition, 34(7), pp. 1405-1416.

[16] J. Ye, T. Li, T. Xiong and R. Janardan. (2004, Using uncorrelated discriminant analysis for tissue classification with gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics 1(4), pp. 181-190. Available: http://dx.doi.org/10.1109/TCBB.2004.45

[17] L. R. Rabiner. (1989, A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), pp. 257-286.