Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Approximated Mutual Information Training for Speech Recognition Using Myoelectric Signals

Hua J. Guo, A. D. C. Chan, *Senior Member, IEEE*
Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada

***Abstract*—A new training algorithm called the Approximated Maximum Mutual Information (AMMI) is proposed to improve the accuracy of myoelectric speech recognition using hidden Markov models (HMMs). Previous studies have demonstrated that automatic speech recognition can be performed using myoelectric signals from articulatory muscles of the face. Classification of facial myoelectric signals can be performed using HMMs that are trained using the *maximum likelihood* (ML) algorithm; however, this algorithm maximizes the likelihood of the observations in the training sequence, which is not directly associated with optimal classification accuracy. The AMMI training algorithm attempts to maximize the mutual information, thereby training the HMMs to optimize their parameters for discrimination. Our results show that AMMI training consistently reduces the error rates compared to these by the ML training, increasing the accuracy by approximately 3% on average.**

***Keywords* – Speech recognition, myoelectric signal, hidden Markov models, Maximum Likelihood, Approximated Maximum Mutual Information**

I. Introduction

ONE of the major difficulties in the implementation of automatic speech recognition (ASR) systems is recognizing speech signals without severe impairment in recognition rates within a variable environment (e.g. changing ambient noise). It is natural to consider other noise resistant alternatives as supplemental information sources for ASR. Myoelectric signals (MES) from articulatory muscles of the face have been proposed as a second source of speech information [1,2]. A multi-expert ASR system was built using the acoustic signal and MES from five facial muscle sites. Performance of the multi-expert ASR system was then compared with a traditional acoustic ASR system under different levels of noise [3]. Experimental results demonstrate that the MES expert is more resistant to the noise interference and that the multi-expert system can achieve classification accuracies that exceed either the acoustic or MES expert. Another example of the MES based ASR systems can be found in [4], where MES collected in an acoustically harsh environment has been used to extract the speech information. Both examples above demonstrated the promising potentials of MES as an alternative or additional information source for the application of ASR.

A MES based ASR system can also be used in the implementation of MES controlled voice prosthesis, which would be beneficial for people with permanent or temporary speech impairments. The advantage of such a system over other communication methods, which are often cumbersome (e.g. writing) or require a significant training time (e.g. sign language), a user could simply mouth the words as they would in normal speech, and have a computer interpret their MES and speak the words for them.

HMMs have been proposed as effective method for MES classification [2,5]. Although the MES is a time-varying signal, within short time intervals the MES can be viewed as a "quasi-stationary" stochastic process, which can be characterized simply by the statistical distributions associated with each state of a HMM. Furthermore, due to articulatory constraints, there are certain dependencies between the MES which occurs in sequence, and a HMM can be used to effectively model this sequential structure.

The maximum likelihood (ML) training of HMM parameters is the most prevalent training algorithm in the community of HMM learning. This ML algorithm maximizes the likelihood of the observations given the training sequences. Although it has been proven that the ML algorithm is computationally efficient and converges after a finite number of iterations, maximization of the objective probability function is not directly associated with the minimization of the recognition error rate. In addition, accuracy of the estimated parameters depends on the availability of a sufficiently large and representative training set of data. A large number of training sequences is often undesirable or impractical in the real world applications.

Many other training algorithms have been proposed with the aim of combating the shortcomings of the standard ML training algorithm mentioned above, such as *maximum mutual information* (MMI) training [6] and corrective training [7], which are based on the concept of discriminative training. The objective function being maximized is the recognition rate rather than the likelihood of the training observation sequences. It has been shown that these approaches improve the recognition rates in certain applications where the classification rate (CR) is critical [6,7]; however, they are not immune to any shortcomings as well. For an example, the MMI training uses the technique of *gradient descent* in the search of the optimal parameters, which has been shown to be very sensitive to the selection of step sizes. Also, the corrective training has been proposed for discrete HMMs and cannot be directly extended to continuous HMMs.

The new AMMI training was introduced by Assaf Ben-Yishai *et al.* [8] and has been applied successfully for an

isolated and connected digit recognition system in a noisy environment. A notable improvement over the baseline achieved by the ML training has been reported. In the AMMI training, parameters of each HMM can be evaluated separately in a manner, similar to the Baum-Welch algorithm for ML training (section II-B), which is more computationally efficient compared with the MMI training.

In this study, we will apply the AMMI training on the classification of a MES database for the task of isolated digit speech recognition, and evaluate the performance of the AMMI training with comparison to the ML training under various parameter settings. The reminder of the paper is organized as follows. HMMs and the ML training and the AMMI training of parameters are reviewed in section II. The MES data collection and the related HMM model are specified in section III. Performance of the AMMI training compared to the ML training in section IV. Finally, conclusions are provided in section V.

I. Hidden Markov Models

A. Classification Using Hidden Markov Models

Let $O=\{o_1, o_2 \ldots o_T\}$ be the sequence of feature vectors extracted from the MES during a word utterance, using a sliding observation window, where o_t denotes the feature vector within the observation window at time t. Consider a first-order Markov chain with N states, parameterized by a state transition probability matrix $A=[a_{ij}]$, and an initial state probability $\Pi=[\pi_i]$, where a_{ij} denotes the probability of making a transition from state i to state j, and π_i denotes the probability of the model being in state i at the initial time t=0. For a given state sequence $Q=\{q_1, q_2 \ldots q_T\}$, q_t =1,2…N, the probability of Q being generated by the Markov chain can be calculated as $P(Q \mid A,\Pi) = \pi_{q_0} a_{q_0 q_1} a_{q_1 q_2} \cdots a_{q_{T-1} q_T}$.

Assume further that when at state q_t, the model generates an observation o_t according to the probability distribution law $b_{q_t}(o_t) = P(o_t \mid q_t)$, then the probability of observing the sequence O can be expressed as,

$$P(O \mid \lambda) = \sum_{Q} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1} q_t} b_{q_t}(o_t) \tag{1}$$

where $\lambda=\{A,\Pi,\{b_i\}\}$, i=1,2…N denotes the parameters of the HMM λ. The probability density function (pdf) for continuous HMMs, denoted as $\{b_i\}$, i=1,2…N, define the statistical characteristics of the observation feature vectors in state i.

Classification of MES using HMMs goes through the following three procedures: training, evaluation, and decision making. First, parameters of an HMM are trained by optimizing an objective function using the training sequences for each word. Next, for an unknown MES, the probability of generating the MES by each HMM is evaluated using the Viterbi algorithm. Finally, the HMM with the highest probability of generating the MES is selected based on the *maximum likelihood* decision making rule.

B. Maximum Likelihood Training

Training of the HMM parameters can be performed using the ML training method. The prevalence of the ML training is partly contributed to the existence of the computationally efficient forward-backward algorithm. Given the training sequence, we want to find the parameters of the HMM which give the highest probability of observing the training sequences, which is equal to finding the optimal λ that maximizes Eq. (1). The optimization problem cannot be solved analytically; however, it can be solved iteratively using the Baum-Welch algorithm [9]. The Baum-Welch algorithm involves two steps. First, a new function $F(\underline{\lambda},\lambda)$ is defined based on an initial guess of HMM parameters $\underline{\lambda}$ and the objective function $P(O|\lambda)$.

$$F(\underline{\lambda},\lambda) = \sum_{Q} P(Q \mid O,\lambda) \log[P(O,Q \mid \underline{\lambda})] \tag{2}$$

It has been proven that maximization of $F(\underline{\lambda},\lambda)$ as a function of λ increases the objective function $P(O|\lambda)$. Thus, if we can find λ that maximizes Eq. (2), we also find improving parameter estimation in the right direction under the ML criterion. We iterate the above two steps until a maximum number of iterations is reached or the change in the likelihood in two sequential training is below a small threshold. The re-estimation formulas for λ are given in [9].

C. Approximate Maximum Mutual Information Training

When the assumptions about the distribution of the data are true and an infinite training set is available, the ML training can produce a consistent estimation of the HMM parameters; however, statistical assumptions are usually not true, nor is the size of training sequences unlimited. In addition, the objective function of the ML training is not directly associated with minimizing the classification error. For classification problems, a new category of training named the *Maximum Mutual Information* (MMI) training, based on the concept of discriminative training, becomes more appropriate.

The MMI training adjusts parameters of HMM so that the mutual information between observations and correct classes is maximized, thus producing a model which has more capability in distinguishing observations generated by different classes. MMI training has to estimate parameters jointly across the entire classes, and the *gradient descent* searching algorithm is sensitive to step sizes; small step sizes take long time for the training to converge, while large step sizes may lead to instability. The AMMI training optimizes the objective function called the *approximated MMI criterion*. Unlike the MMI training, the parameters for each HMM can be calculated separately in a method similar to the Baum-Welch algorithm.

Let γ be the parameter of the HMM to be estimated,

$$J(\gamma) = \sum_{O \in u} \log P(O \mid \gamma) - d \sum_{O \in v} \log P(O \mid \gamma) \tag{3}$$

where u is the set of indices of the training data that were labeled as the class specified by γ, v is the set of indices of the training data that were recognized as the class specified by γ after the *maximum a posteriori* (MAP) criterion is applied on the training data, and d is a parameter that can be used for adjusting the discrimination rate [8]. The

optimization for Eq. (3) can be implemented in a way similar to the Baum-Welch algorithm and the re-estimation formulae are given in [8].

II. Methodology

A. MES Data Collection

The MES database used in this study was from a previous study [3]. A 10-word vocabulary consisting of digits from "zero" to "nine" was used. Five Canadian-English speaking subjects participated in this study. MES from articulatory muscles of face were recorded from each subject when uttering randomized word sequences from the vocabulary. MES data were collected with six levels of acoustic white noise (0, 6, 9, 12, 15, and 18 dB). Each word in the vocabulary was repeated 12 times for the 0 dB level of noise and 8 times for the other levels of noise.

MES were collected using Ag-AgCl electrodes from five different positions of the face: *anterior belly of the digastric*, *platysma*, *depressor anguli oris*, *levator anguli oris*, and *zygomaticus major*. MES were sampled simultaneously with the acoustic speech with a sampling rate of 10 kHz. Anti-aliasing filters, with cutoff frequency of 500 Hz and 5000 Hz were used for the MES and acoustic speech, respectively. The acoustic signals were used to segment the MES, using a pretrigger value of 500 ms from the start of acoustic speech, which was expected to be near optimal [1,2].

B. Classifier

A left-right HMM with single mixture observation Gaussian densities was used to classify the MES. For simplicity, a diagonal covariance matrix for observation features was used. Unfortunately, there is no systematic method in the choice of state size [9]; therefore, the optimal number of states was found empirically.

C. Feature Extraction

MES were downsampled to 1000 Hz and segmented into 1024 ms records. A fixed size of overlapping observation window was used to extract features from the MES. The root mean square (RMS) value and the autoregressive (AR) coefficients were computed within each observation window. Optimal window size, window spacing, and AR order were obtained empirically, as described in the section IV.

D. Training and Testing

The MES database used in this study were recorded with six different levels of acoustic noise [3]. In order to evaluate the effects of different training sizes and mitigate the effect of insufficient training sequences, MES data from the different levels of noise were combined together. Such combination was reasonable because the MES remained relatively unaffected by acoustic noise [3]. A total number of 260 MES data sets were used for HMM training; the remaining 260 MES data sets were used for testing. When a smaller training sequence was required, we decimated the 260 training sequences. During the HMM training phase, we applied the AMMI training and the ML, training separately, Performance of the resultant HMMs was compared by the error rates in the recognition of the testing MES data.

III. Results and Discussion

A. Observation Window Size

Different size of observation window and window overlapping were applied to extract features from the MES data. The window sizes used were: 16 ms, 32 ms, 64 ms, 128 ms, and 256 ms. Observation windows were spaced one eighth of the window size and a 10-state HMM was used. The RMS and first two AR coefficients were used as features. The error rates with different observation window sizes for both the AMMI training and the ML training are shown in Fig. 1. The error rates decreased significantly with increasing window size until the optimal window size of 128 ms was reached, after which the error rates increased. As shown in Fig. 1, the HMM model by the AMMI training had a consistently lower error rate than the ML training. With 128 ms observation window, the error rate was 12.77% and 13.69% for the AMMI and ML training, respectively.

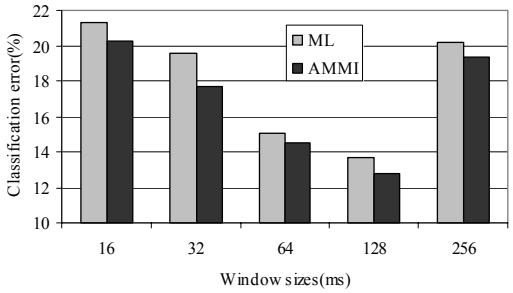

Fig. 1 Classification error as a function of window sizes

B. Number of States

In this experiment, a left-right HMM with the number of states varying from 3 to 15 was used. The observation window size was fixed at 128 ms, with 16 ms window spacing, and the RMS and first two AR coefficients were used as features. The error rates with the changing state sizes are plotted in Fig. 2. The error rates increased when we use a HMM with too few or too many numbers of states. The optimal state size was 10 for both the ML and AMMI training. The AMMI consistently outperformed ML.

C. Number of Features

In this section, we compared performance of the AMMI training to that of the ML training varying the number of features. A 10-state HMM was used with 128 ms observation window size, spaced 16 ms apart. The RMS and AR coefficients were used as features, with the AR order ranging from 0 to 12.

As shown in Fig. 3, when only the first AR coefficient was used as a part of feature vectors, the error rate obtained by the AMMI training was 16.92%, when the first two AR coefficients were applied, the error rate dropped significantly to 12.77%. The error rates from the ML training were 17.31% and 13.69% respectively; however using more than 2 AR coefficients does not further reduce the error rate. The

introduction of more features increases the number of free parameters in the HMM, which would require additional training data to properly estimate the HMM parameters. Unless the additional features are providing sufficient additional discerning information, they can increase the error rate. Again the AMMI consistently improved the error rate compared to the ML.

D. Training size

Using the empirically optimal HMM and feature settings found from the previous sections, the effect of the training size was evaluated. The training size was decreased from 260 to 100 by steps of 10. The error rates for the AMMI training and the ML training as a function of the training size is shown in Fig. 4. The classification errors tend to increase with the reduction of the training size, as expected. The classification error for the AMMI training was 12.77% (13.69% for the ML training) when the training size was 260, and rises to 23.31% (23.54% for the ML training) when the training size was reduced to 100.

IV. CONCLUSION

It has been shown that the AMMI training can be used as an alternative in the estimation of HMM parameters to reduce error rates of MES classification on ASR. The experimental results demonstrated that the AMMI training obtained a HMM model with a reduction in recognition error rates when compared to that with the ML training. This reduction in error rates was consistent when changing the window size, number of states, number of features, and training size. The AMMI training uses the misclassification information in the training sequences to re-estimate the parameters, which maximizes the approximated MMI between training sequences and correct classes, thus obtains a gain in recognition rates over the standard ML estimation.

ACKNOWLEDGMENT

We would like to thank Assaf Ben-Yishai and Antti Eronen for their insightful discussions on AMMI training. The work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

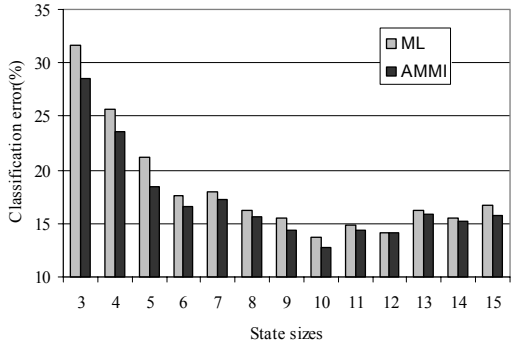

Fig. 2 Classification error rates as a function of state sizes

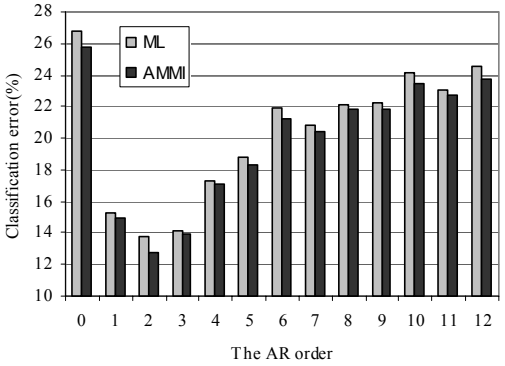

Fig. 3 Classification error as a function of the AR order

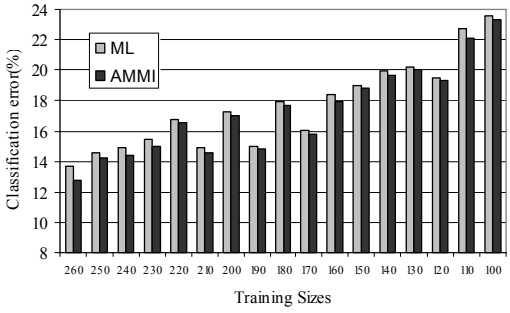

Fig. 4 Classification error rates as a function of training sizes

REFERENCES

[1] A. D. C. Chan, K. Englehart, B. Hudgins, D. F. Lovely, "Multi-expert automatic speech recognition using acoustic and myoelectric signals," in speech recognition," IEEE EMBS Magazine, vol. 21, no. 4, pp. 143-146, 2002.

[2] A. D. C. Chan, K. Englehart, B. Hudgins, D. F. Lovely, "Myoelectric signals to augment speech recognition," Med. Biol. Eng. Comp., 39(4):500-504, 2001.

[3] A. D. C. Chan, K. Englehart, B. Hudgins, D. F. Lovely, "Multi-expert automatic speech recognition using acoustic and myoelectric signals," IEEE Trans. Biomed. Eng., *accepted for publication*, 2005.

[4] B. J. Betts, C. Jorgensen, "Small Vocabulary Recognition Using Surface Electromyography in an Acoustically Harsh Environment," http://nel.arc.nasa.gov/Papers/NASA_TM_Betts05.pdf.

[5] A. D. C. Chan, K. Englehart, "Continuous myoelectric control for powered prostheses using hidden Markov models", IEEE Trans. Biomed. Eng., 52(1):121-124, 2005.

[6] L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," ICASSP 86, 49-52, Apr. 1986.

[7] L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer, "Estimating Hidden Markov Model Parameters so as to Maximize Speech Recognition Accuracy," IEEE Trans. Speech and Audio Proc., vol. 1, no.1, pp. 77-83,1993.

[8] A. Ben-Yishai, D. Burshtein, "A Discriminative Training Algorithm for Hidden Markov Models," IEEE Trans. Speech and Audio Proc., vol. 12, no.3, pp. 204-217, 2004.

[9] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, pp.257-285, 1989.

[10] L. R. Bahl, P. F. Brown, P.V. de Souza, and R. L. Mercer, "A New Algorithm for the Estimation of Hidden Markov Model Parameters," ICASSP, p. S11.2, 1988.

[11] J.-L. Gauvain and C.-H. Lee, "*Maximum a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," IEEE Trans. Speech and Audio Proc., vol. 2, no.2, pp. 291-298, 1994.