Connectionist Approach for Emission Probability Estimation in Malayalam Continuous Speech Recognition

Anuj Mohamed and K.N. Ramachandran Nair

Abstract Automatic speech recognition is one active research area which can exploit the pattern recognition capabilities of artificial neural networks. Several researchers have shown that the outputs of artificial neural networks trained in multi-class classification mode can be interpreted as estimates of a posteriori probabilities of output classes. These probabilities can be used by the state-of-the-art hidden Markov model for speech recognition in estimating the emission probabilities of the states of the hidden Markov model. In this paper, we explore a pairwise neural network system as an alternative approach to multi-class neural network systems to estimate the emission probabilities of the states of a hidden Markov model. Through experimental analysis it is shown that the pairwise recognition system outperforms the multiclass recognition system in terms of the recognition accuracy of spoken sentences.

Keywords Continuous speech recognition · Malayalam speech recognition · Multi-class pattern classification · Pairwise pattern classification

1 Introduction

Automatic speech recognition (ASR) is a classic pattern recognition problem that aims to produce a text transcription of spoken words automatically. Research in this area attained attention of researchers because of the fast growing demands of ASR systems in versatile applications. The dominant technology for ASR is hidden

A. Mohamed (✉)
School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India
e-mail: anujmohamed@mgu.ac.in

K.N. Ramachandran Nair
Department of Computer Science and Engineering, Viswa Jyothi College of Engineering and Technology, Vazhakulam, Muvattupuzha, Kerala, India
e-mail: knrn@hotmail.com

J.K. Mandal et al. (eds.), *Information Systems Design and Intelligent Applications*,
Advances in Intelligent Systems and Computing 339,
DOI 10.1007/978-81-322-2250-7_34

Markov model (HMM). An HMM is typically defined and represented as a stochastic finite state automation usually with a left-to-right topology as proposed in [1]. An HMM models an utterance as a succession of discrete stationary states with instantaneous transitions between the states and generates an observation [2]. The sequence of states, which represents the spoken utterance, is "hidden" and the parameters of the probability density functions of each HMM state are needed in order to associate a sequence of states to a sequence of observations. i.e., for any state sequence, the sequence of states can be observed only through the emission/ output probability distribution.

State-of-the-art continuous speech recognition (CSR) systems select basic sound units from a limited inventory, like sub-word units, syllables, phonemes, allophones etc. Each basic acoustic unit is modeled by one or more states of an HMM. Being a parametric model, HMM assumes that the underlying speech patterns have Gaussian distribution. Gaussian mixture models (GMM) are often used within each HMM state to model the emission probability of the acoustic patterns associated to that state. Several researchers [3, 4] have shown that the outputs of artificial neural networks (ANNs) trained in classification mode can be interpreted as estimates of a posteriori probabilities of output classes conditioned on the input. Using Bayes' rule, these state posteriors can be converted to likelihoods required by the HMM framework.

Use of ANN as emission probability estimator has shown good performance for ASR systems for various western languages [5, 6]. Application of multilayer perceptrons (MLP) trained in multi-class classification mode as emission probability estimator in an HMM based CSR system for Malayalam language is proposed in [7]. Many real applications translate into classification problems with a large number of classes and a huge amount of data. ASR falls into this category and the high number of classes to be separated makes the boundaries between classes complex. A multi-class classifier cannot perform at a high level in such cases as the classification algorithm has to learn to construct a large number of separation boundaries. Many studies have demonstrated that an adequate decomposition of such real world problems into sub problems can be favorable to the overall computational complexity. The pairwise classification approach has produced good results when applied to face recognition [8] and cursive handwriting recognition [9]. The efficiency of such pairwise ANN classifiers has not been explored sufficiently in ASR applications. A pairwise classifier for multi-class pattern classification in the context of vowel classification tasks is proposed in [10]. Motivated by this fact, this paper examines the prospect of using a pairwise classification approach for multi-class pattern classification, as a way to improve overall classifier performance. In the present paper, we report on the application of pairwise approach for improving the recognition accuracy of the CSR system for Malayalam language.

This paper is organized as follows. Section 2 gives an overview of ASR system. ANN as phonetic probability estimator of HMM states is explained in Sect. 3. Section 4 details the feature extraction, emission probability estimation using multi-class and pairwise ANN classifiers and the decoding process. Experiments conducted and the results obtained are given in Sect. 5. Conclusions and future work to enhance the performance of the system are given in Sect. 6.

2 Automatic Speech Recognition: Mathematical Formulation

The ASR problem in general can be viewed as a pattern recognition problem. Given an acoustic sequence O and a sequence of words W, the goal of the system is to find the most likely word sequence W′ such that

$$W' = \arg\max_{w} P(W|O). \tag{1}$$

Applying Bayes' rule to this fundamental equation of speech recognition returns

$$W' = \arg\max_{w} \frac{P(O|W)P(W)}{P(O)}. \tag{2}$$

As P(O) is equivalent for all complete decodings of O, this can be ignored, to give the equation

$$W' = \arg\max_{w} P(O|W)P(W). \tag{3}$$

P(W) is the priori probability of the word sequence W and is computed using the language model. P(O|W) is the conditional probability of the acoustic sequence O given the word sequence W, and is computed using the acoustic model. A continuous density hidden Markov model which models the real valued vector sequences resulting from acoustic parameter extraction is used as the acoustic model in state-of-the-art continuous speech recognition systems.

3 ANN as Phonetic Probability Estimator in HMM

The state-of-the-art HMM framework for ASR uses GMM to estimate the emission probabilities of the HMM states. An important stage in the development of an HMM based ASR system is the accurate estimation of the emission probabilities. The training algorithm for the estimation of the parameters of the GMM is based on likelihood maximization, which assumes correctness of the models and implies poor discrimination. Also, maximizing the likelihood of the training data is not closely related to the typical evaluation criteria of the recognizer: the error rate. These issues with maximum likelihood (ML) estimation motivate an alternative form of estimating model parameters called discriminative training. A model used for representing a phonetic class is said to be discriminant if it maximizes the probability of producing an associated set of features while minimizing the probability that they have been produced by rival models.

ANNs with its several features, such as their nonlinearity and high classification capability, is a promising field for solving real world problems. Neural network classifiers are easier to apply to real-world problems because they are less sensitive to assumed underlying distributional forms than more conventional probabilistic approaches. ANN by itself has not been shown to be effective for recognition of continuous speech. ASR systems can exploit the pattern classification capabilities of ANNs as it involves development of pattern classification models from speech data. When an ANN is used to solve a classification problem, the network can be trained either to provide the classification directly or to model the posterior probabilities of class membership. Using Bayes' rule, these probabilities can be converted to likelihoods required by the HMM. ANN follows discriminant-based learning. i.e., the network during training maximizes the probability of producing an associated set of features while minimizing the probability that they have been produced by rival models.

ASR falls into the category of classification problems with a large number of classes and a huge amount of data. The large number of classes to be separated makes the boundaries between classes complex. A multi-class classifier cannot perform at a high level in such cases as the classification algorithm has to learn to construct a large number of separation boundaries. Computationally expensive learning algorithms learn many small problems much faster than a few large problems. Therefore, the use of posteriors derived from pairwise classifiers may be helpful to improve the performance. Motivated by this fact, the prospect of pairwise modeling approach for multi-class pattern classification is explored in this paper. The pairwise approach to pattern classification is simple since the binary decision is learned on fewer training examples. It also helps to improve the generalization ability of the network because of the redundancy in the training data.

4 Experimental Setup

4.1 Speech Corpora

Malayalam is a Dravidian language spoken mainly in the South West of India. Research in Malayalam speech recognition has started recently and the area of CSR is relatively less investigated. The speech corpora developed for training and testing the CSR system described in this paper contains naturally and continuously read Malayalam sentences from both male and female speakers who speak various dialects of Malayalam. Two databases (Dataset1 and Dataset2) for training and one (Dataset3) for testing were developed. The entire corpus consists of 255 sentences with 1,275 words and a total of 7,225 phonemes by 20 (9 male & 11 female) different speakers.

4.2 Acoustic Pre-processing

The acoustic pre-processor extracts a compact set of features required for speech recognition. The feature vector used in this work is mel-frequncy cepstral coefficients (MFCC). The feature vectors are extracted by the following procedure. First the speech signal is digitized at 8 kHz sampling rate with A/D conversion precision of 16 bits. An FFT is performed on every 8 ms on a window of 256 samples which produced 12 dimensional vectors representing the energy in 28 triangular filters spaced according to the mel-frequency scale.

4.3 HMM Emission Probability Estimation Using Multi-Class MLP Classifiers

In the multi-class classification approach, a single network is used to generate posterior probabilities of all the selected classes. When an input vector, x, is presented to the network, the activation of each output unit represents the corresponding posterior probability, provided the system has enough parameters and the training does not get stuck at a local minimum. HMMs use likelihoods P(O|W) where O is the acoustic observation and W is the acoustic model. But the neural network estimates the posteriors P(W|O*)* and this can be converted to likelihoods by applying the Bayes' rule. This can be achieved by dividing the posterior estimates from the MLP outputs by estimates of corresponding class priors. Supervised training using ANN requires the temporal segmentation of the training sentences and frame labeling. As hand segmentation of the speech corpus is tedious, the baseline CDHMM system described in [11] is used to generate segmented training data by Viterbi-aligning the training reference transcription to the acoustic data.

The multi-class classifier used in this work is a two-layer MLP with a single hidden layer. The input feature vector given to the network is computed from a window of seven speech frames i.e., acoustic vectors from the frame to be classified along with three surrounding frames for the left and right contexts, respectively. A frame is represented by a spectral vector of 12 acoustic features computed every 10 ms. So the input layer of the network has 84 nodes. The hidden layer has 95 nodes all with sigmoid activation function. The output layer has number of nodes corresponding to the context independent phonetic units. The output unit corresponding to the input vectors class has a value "1" while all other outputs have value "0". Network output should sum to one for each input value if outputs accurately estimate the posterior probabilities. This is achieved by using the soft-max activation function at the output layer. The cross-entropy error function for multi-class classification problem is used to compute the error between the output of the ANN and the target vector. Back-propagation with batch mode of updating the weights is used for training the network.

4.4 HMM Emission Probability Estimation Using Pairwise MLP Classifiers

In the pairwise approach to K-class pattern classification there are K (K−1)/2 binary neural networks. The outputs of the pairwise neural network classifiers provide the probabilities p_{ij} *f*or all pairs of classes (i, j) with i ≠ j. Let $p_{ij}(x)$ be the pairwise class probability estimate of the input vector x to belong to class i, with i ≠ j. Then the pairwise class probability estimate of the input vector x to belong to class j, is given by $p_{ji}(x) = 1-p_{ij}(x)$. Now the multi-class posteriori probabilities are estimated by combining the outputs of the K (K−1)/2 pairwise classifiers using the method described in [9] i.e.,

$$P(c_i|x) = \frac{1}{\sum\limits_{j=1;j\neq i}^{K} \frac{1}{p_{ij}} - (K-2)}. \tag{4}$$

For each pair of classes a two layered MLP with a single output unit is created as it is trained on the data of the two classes. The first 12 MFCCs extracted from the input speech signal is used as the input to the network and the MLP has input nodes corresponding to the number of acoustic features. As each of the K pattern classes is trained against every one of the remaining pattern classes, the redundancy in the training data improves the generalization ability of the network and there is no need to incorporate contextual information. This reduces the dimension of the input feature vector. The output layer consists of a single neuron with sigmoid activation function which has an output range [0, 1] representing the output phoneme class. The activation function used by the hidden units is sigmoid and the optimal number of hidden units is determined empirically.

The back-propagation algorithm with cross-entropy error criteria is used for training the network. The weights and biases of the network are initialized randomly. In order to make the network training more efficient, the inputs and outputs are normalized to have zero mean and unit standard deviation. In order to improve the generalization capability of the network, performance obtained on a cross-validation set is used as the stopping criterion. The cross-validation set contains the speech utterances that are not used for training. Training continued as long as the performance on the validation set is improved and the training is stopped when the network ceased to improve.

4.5 Decoding and Model Evaluation

Decoding refers to the process of searching for the sub word sequence that may have generated an observation sequence. In CSR, this is achieved by inferring the actual sequence of states that generated the given observation sequence and then recovering the word sequence from the state sequence. The Viterbi [12] algorithm is widely used for decoding in ASR systems. This algorithm returns a single best state sequence for an unknown input sequence. Trace back through this sequence gives the most likely phone and word sequence. The hypothesized transcription and the reference transcription can then be compared by evaluating their distance according to the Levenshtein metric, in terms of insertions, deletions and substitutions. Using these measures, the word error rate (WER), a common evaluation measure for ASR systems, is computed.

5 Results and Discussions

This section compares the recognition performance of the Malayalam CSR system, when the posteriors were estimated using the MLP trained with multi-class and pairwise classification approaches. Table 1 shows the WER of both the systems and the percentage reduction in WER obtained when the pairwise classification approach was used for the emission probability estimation. A strong improvement in recognition performance was observed when the HMM emission probabilities were computed using the pairwise classification approach i.e., a relative improvement of 66.67 % and 57.23 % were obtained when the system was trained with Dataset1 and Dataset2 respectively. This result supports the use of MLPs trained in pairwise classification mode for the estimation of emission probabilities of HMM states.

The pairwise approach yields a more flexible class of models than a multi-class approach, as only one decision boundary requires attention. Number of free parameters to be trained in the case of the two approaches, for the best recognition performance obtained, is shown in Table 2.

Table 1 Performance of the system with multilayer perceptrons trained with multi-class and pairwise classification approaches for HMM emission probability estimation

Training datasets	WER		% Reduction in WER Multi-class → pairwise
	Multi-class MLP classifier	Pairwise MLP classifier	
Dataset1	2.67	0.89	66.67
Dataset2	3.11	1.33	57.23

Table 2 Comparison of trainable parameters: HMM with GMM emission probabilities versus HMM/ANN hybrid approaches with MLPs trained in multi-class and pairwise classification modes for emission probability estimation

No. of trainable parameters		% Reduction in Trainable parameters
Multi-class MLP classifier	Pairwise MLP classifier	Multi-class MLP classifier → pairwise MLP classifier
10,475	8,700	16.95

6 Conclusions

Recognition accuracy is an important aspect of ASR systems. They must achieve a very high level of performance to be of general interest as a man-machine interface. Enhancing recognition accuracy by way of improving discrimination between pattern classes is one of the fundamentally important research areas for speech recognition. This paper has examined the integration of MLPs trained in pairwise classification mode as emission probability estimators in an HMM based continuous Malayalam speech recognition system. The system with MLPs trained in pairwise classification mode outperforms the system with multiclass classification approach in terms of recognition accuracy of the spoken sentences. The proposed pairwise approach is simple since the binary decision is learned on fewer training examples. It also improves the generalization ability of the network because of the redundancy in the training data and is very useful in the case of very limited training material. Also, adding a new class or modifying the training set of an existing one can be done without retraining all two-class classifiers.

In this work, the a posteriori probabilities obtained from the MLPs trained in multiclass and pairwise classification modes were used as the estimates of emission probabilities of HMM states in an HMM based Malayalam CSR system. Auto associative neural networks and deep neural networks are two emerging alternatives to MLPs as emission probability estimators. The use of these neural network architectures may further improve the performance of the system.

References

1. Bakis, R.: Continuous speech recognition via centiseconds acoustic states. J. Acoust. Soc. Am. **59**. (S1) (1976)
2. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proc. IEEE **77**, 257–285 (1989)
3. Richard, M.D., Lippmann, R.P.: Neural network classifiers estimate bayesian a posteriori probabilities. Neural Comput. **3**, 461–483 (1991)
4. Bourlard, H., Morgan, N.: Continuous speech recognition by connectionist statistical methods. IEEE Trans. Neural Networks **4**, 893–909 (1993)
5. Seid, H., Gamback, B.: A speaker independent continuous speech recognizer for amharic, INTERSPEECH, 9th European Conference on Speech Communication and Technology, Lisbon (2005)
6. Meinnedo, H., Neto, J.P.: Combination of acoustic models in continuous speech recognition hybrid systems. In: Proceedings of ICSLP, Beijing (2000)
7. Anuj, M., Nair, K.N.R.: HMM/ANN hybrid model for continuous malayalam speech recognition. Procedia Eng. **30**, 616–622 (2012)
8. Uglov J., Jakaite L., Maple C.: Comparing robustness of pairwise and multiclass neural network systems for face recognition. EURASIP J. Adv. Signal Proc. (2008)
9. Price, D., Knerr, S., Personnaz, L., Dreyfus, G.: Pairwise neural network classifiers with probabilistic outputs. In: Advances in Neural Information Processing Systems, Tesauro, G., Touretzky, D., Leen, T. (eds.), vol. 7, The MIT Press, Cambridge (1995)

10. Klautau, A., Jevtic, N., Orlitsky, A.: Combined binary classifiers with applications to speech recognition. In: Proceedings of International Conference on Spoken Language Processing, pp. 2469–2472 (2002)
11. Anuj, M., K.N. R. Nair.: Continuous malayalam speech recognition using hidden markov models. In: Proceedings of 1st Amrita-ACM-W Celebration on Women in Computing in India (2010)
12. Viterbi, A.J.: Error bounds for convolutional codes and asymptotically optimum decoding algorithms. IEEE Trans. Inf. Theory **13**, 260–269 (1982)