NEURAL NETWORK ACOUSTIC MODEL FOR RECOGNITION OF CZECH SPEECH

Tomáš Pavelka and Kamil Ekštein

Laboratory of Intelligent Communication Systems,
Dept. of Computer Science and Engineering,
University of West Bohemia in Plzeň, Czech Republic

Abstract: The paper compares two methods for acoustic modeling: artificial neural networks and Gaussian mixtures. The former is used by the current version of the LASER recognizer which is under development by our research team. The achieved results are compared to those of a recognizer based on continuous density hidden Markov models with Gaussian mixtures. From the experimental results we can conclude that the use of neural networks for acoustic modeling can lead to a reduction in the number of trainable parameters.

Keywords: Automatic speech recognition, neural network, acoustic models

1. INTRODUCTION

Today's state-of-the-art automatic speech recognition (ASR) systems are mostly based on continuous density hidden Markov models (CDHMMs). CDHMMs can be automatically trained (given a large number of speech recordings and their phonetic transcription) and are very effective in terms of computational costs. Despite their obvious success there are some well known limitations which may decrease their performance on the task of speech recognition.

One of the most notable alternatives to hidden Markov models is the so called hybrid approach which combines the advantages of HMM systems with the universality of artificial neural networks (ANNs). A typical hybrid system uses HMMs with state emission probabilities computed from output neuron activations of a neural network (such as the multi layer perceptron).

At the Laboratory of Intelligent Communication Systems we are currently developing a hybrid ASR system for the recognition of continuous Czech speech called LASER (LICS Automatic Speech Extraction/Recognition, see Ekštein et al. 2004). The goal is to develop a set of tools that would allow training of general acoustic models and the recognition with user supplied dictionary and grammar (or language model).

In order to evaluate the system's performance a set of experiments has been carried out with the current version of the LASER system and a CDHMM ASR system built with the HTK software (see Young et al. 2002). This paper will describe the architecture of both recognizers as well as discuss the achieved experimental results.

2. CONTINUOUS DENSITY HIDDEN MARKOV MODELS

A hidden Markov model is a probabilistic finite state automaton which changes state every discrete time unit and generates an *observation* (i.e. a feature vector in speech recognition). An HMM is defined as a set of states, a state transition probability matrix and a set of emission probability distributions.

The emission probability distribution function (PDF) estimates the probability with which a given observation has been generated (emitted). In ASR the most commonly used PDF is the mixture of multivariate Gaussian functions (see e.g. Rabiner 89 for details). In order to reduce the number of trainable parameters and lower the computational costs Gaussians with diagonal covariance matrix are often used. This simplification requires that the individual components of the observation vector are statistically independent.

For the application to ASR there are three basic problems of interest (an efficient algorithm for the solution of each of these problems exists):

1. Given an observation sequence and a HMM what is the probability that the sequence has been generated by the HMM?
2. Given an observation sequence and a HMM which sequence of model states is most likely to have generated the observation sequence?
3. How to adjust the model parameters in order to maximize the probability that the HMM has generated the observation sequence?

The solution to problem 1 also solves the problem of isolated word recognition: If there is an HMM for each word in the dictionary the recognized word is the one with the highest probability. Having a set of observation sequences belonging to each word the solution to problem 3 can be used to train the word models.

Continuous speech recognition is more difficult since there cannot be a HMM for every possible utterance. A sentence is rather modeled as a sequence of words and those words as sequences of some sub-word units (often referred to as *phonetic units*). One or more HMM states represent such phonetic unit. This reduces the number of training parameters and thus lowers the computational requirements and the number of training data.

A composite HMM representing all the recognized utterances is constructed and the Viterbi algorithm (which solves problem 2) is applied to find the best state sequence. A word sequence can then be easily derived from the state sequence.

3. NEURAL NETWORK ACOUSTIC MODEL

The purpose of an *acoustic model* is to provide scores for each phonetic unit based on the input feature vector (observation). The Gaussian mixture discussed in the previous chapter is an example of an acoustic model where the score is an emission probability. A neural network, such as the multi layer perceptron can be used as an acoustic model. It has been proven (see e.g. Bourlard et al. 97) that when trained by sufficient data the output neuron activations can be interpreted as *a posteriori* probabilities of classes (i.e. phonetic units).

If such acoustic model is to be used in an ANN/HMM hybrid the state emission probabilities can be computed from a posteriori probabilities using Bayes rule:

$$P(o \mid S_j) = \frac{P(S_j \mid o) \cdot P(o)}{P(S_j)} \tag{1}$$

According to many researchers (e.g. Bourlartd et al.) the usage of neural networks has many potential advantages over the conventional Gaussian mixtures, namely:

- **Model accuracy**. The use of Gaussian mixtures requires some assumptions on the form of the probability distribution being modeled such as that the features are statistically independent (if the covariance matrix is diagonal). On the other hand a multi layer perceptron does not place any such assumptions on the data and can approximate any function. This even allows the joining of different feature types on the input of the classifier, e.g. binary and real numbers.
- **Context sensitivity**. If several adjacent feature vectors are appended to form the input of the neural network $X_{t-c}^{t+d} = \{x_{t-c}, \ldots, x_{t,} \ldots, x_{t+d}\}$ or if recurrent neural networks are used then time correlation of the feature vectors can be taken into account when estimating the probability distribution. Gaussian mixture based systems also try to incorporate contextual information by adding first and second order derivatives to the feature vector.
- **Economy**. Gaussian mixtures use their parameters to model the surface of the density function in acoustic space, in terms of likelihoods, while neural networks use their parameters to model class boundaries, in terms of posteriors. Boundaries require less parameters and thus can make a better use of limited training data. Empirically ANN/HMM hybrids require less trainable parameters to obtain the same accuracy as conventional HMM systems.
- **Discrimination**. The standard training criterion of HMMs (maximum likelihood) does not guarantee discrimination between models. Neural networks, if trained with SSE or similar error criterion, easily provide discrimination although on local (frame) level only.

3. RECOGNIZER ARCHITECTURE

Both recognizers work with three state HMMs of phonetic units. There are 36 context-independent phonetic units (including one for silence) which are roughly equal to Czech phonemes. In the following experiments a deterministic language model is employed, the user supplies a pronunciation dictionary and a grammar which defines all the utterances to be recognized. From this it is possible to construct a composite HMM which is then searched by the Viterbi algorithm for the most likely word sequence.

While this architecture is not suitable for more complicated problems such as large vocabulary continuous speech recognition it can be used for testing of the acoustic model performance.

Both recognizers have 13 MFCCs (Mel-scale Frequency Cepstral Coefficients) for features normalized to have zero mean.

3.1 HTK Recognizer

This recognition system was built with the HTK toolkit (Young et al. 2002). Each HMM state has 32 Gaussian mixtures with diagonal covariance matrix. The 13 MFCC coefficients are augmented by their first and second time derivatives leading to total 39 components of a feature vector. A significant advantage of the HTK recognizer is that it only needs phonetic transcriptions of the data for training (i.e. the exact phoneme boundaries are not necessary).

3.2 LASER Recognizer

This architecture is also based on three state HMMs but the unlike in the former case the emission probabilities of all three states are *tied* i.e. are computed from the activation of the same output neuron. Also the HMM transition probabilities are not trained but are considered to have uniform distribution. We have conducted an experiment with trained transition probabilities which have been replaced by uniform distribution and the decrease of the systems performance was negligible.

A multi layer perceptron trained by the backpropagation algorithm serves as an acoustic model. Features from nine consecutive frames are used as the input of the neural network (altogether there are 117 input neurons). There are 36 output neurons (equals the number of phonetic units) and the number of hidden neurons has been empirically determined to be 400 (further increase in number does not increase recognition accuracy with the available training data). All neurons have nonlinear sigmoidal activation functions.

For the training of multi layer perceptron exact locations of phoneme boundaries must be known. The data can either be labeled manually (which is quite time consuming task) or automatically. The automatic procedure is called *forced Viterbi alignment* and an already trained recognizer is necessary for this task.

4. EXPERIMENTAL RESULTS

The recognizers were trained by a studio-quality (i.e. very quiet environment with no reverberation) 16 KHz audio corpus consisting of two kinds of utterances:

1. Utterances for voice control of a chess game (Such as “Move the king to D3”, “Start a new game” etc.)
2. Artificially constructed nonsense utterances compensating for infrequent phonemes.

By the time of writing of this paper recordings from 81 speakers (31 male, 50 female) were available. Each speaker reads 40 utterances. The total time is 3:09 hours which results in approx. 700 thousand training vectors.

The first word error rate tests were conducted on random sentences from the chess corpus, but were soon found to be inappropriate. The most errors were made on chess moves (the chess column names A, B, C,… sound very similar) while general commands like “Start a new game” were recognized almost perfectly, so the error rate on the test set was largely dependent on the command/move ratio. In order to overcome this problem another testing corpus consisting of 400 move utterances from 4 speakers was recorded. All the following experiments were conducted on this test set.

Table 1. The percentage of correctly recognized words for both architectures

Architecture	No. of parameters	H	N	%Correct
HTK	269568	2942	3018	97.48
LASER	61636	2944	3018	97.55

Table 1 shows the achieved results for both recognizers. The column “H” shows the total number of correctly recognized words, “N” is the total number of words in referential transcription. The percentage of correctly recognized words is computed as H/N*100 %.

The number of trainable parameters for both recognizers was iteratively increased (for HTK by adding mixtures, for LASER by adding hidden neurons) until the performance ceased to increase.

5. CONCLUSIONS

It can be seen from the experimental results that the neural network recognizer can achieve the same recognition accuracy as a CDHMM based one (given the available training data). The claim (stated in section 3) that neural network acoustic model is more economical in terms of trainable parameter numbers has been verified by this experiment since it needs only approx. 23% of the parameters needed by the Gaussian mixtures.

We have tried to further increase the number of mixtures (up to 128) and hidden neurons (up to 2000) but it resulted only in very slight increase of recognition accuracy. We believe that this is due to the size of our training corpus which is, for the purpose of speaker independent ASR very small.

REFERENCES

Bourlard, H. and Morgan, N. (1997). Hybrid HMM/ANN Systems for Speech Recognition: Overview and New Research Directions, *Summer School on Neural Networks*, 1997.

Ekštein, K., Pavelka, T. (2004). LINGVO/LASER: Prototyping Concept of Dialogue Information System with Spreading Knowledge, *Proc. of NLUCS 2004,* INSTICC PRESS, Porto, Portugal.

Young, S. et al. (2002), *The HTK Book (for HTK v. 3.1)*, Cambridge University Engineering Dept.