ANALYSIS OF DEEP NEUTRAL NETWORKS IN ACOUSTIC MODELING

Vibhute Pritish Mahendra
Regd No: 221120067
Research Scholar
Department Of Electronics And Communication Engineering
SJJT University, Rajasthan.

Dr. Mohammad Iliyas
Prof & Hod - Ece
Shadan College Of Engineering & Technology, Hyderabad.

Dr. Bharat Gupta
Guide
Department Of Electronics And Communication Engineering
SJJT University, Rajasthan.

Abstract

In this work, we propose a modular combination of two popular applications of neural networks to large-vocabulary continuous speech recognition. First, a deep neural network is trained to extract bottleneck features from frames of mel scale filterbank coefficients. In a similar way as is usually done for GMM/HMM systems, this network is then applied as a nonlinear discriminative feature-space transformation for a hybrid setup where acoustic modeling is performed by a deep belief network Most current speech recognition systems use hidden Markov models (HMMs) to deal with the temporal variability of speech and Gaussian mixture models to determine how well each state of each HMM fits a frame or a short window of frames of coefficients that represents the acoustic input. An alternative way to evaluate the fit is to use a feedforward neural network that takes several frames of coefficients as input and produces posterior probabilities over HMM states as output.

Keywords: *Acoustic Modeling, Deep Belief Networks.*

Introduction

Recently, multiple works have demonstrated that the performance of automatic speech recognition systems can be heavily improved by using deep neural networks (DNNs) for acoustic modeling. The key advantages over much earlier approaches to this hybrid setup combining neural networks and hidden Markov models are improved learning algorithms that can leverage the high modeling capacity of deep networks and the usage of a large number of context-dependent phonetic target states during network training Work is still done in determining which speech features are most useful when training neural network acoustic models. For generative models like restricted Boltzmann machines, it has been argued that raw mel scale spectral coefficients are more suitable than further preprocessed features with reduced covariances like MFCCs. In practice, though, it appears that deep networks can been trained with similar performance on a variety of acoustic data, including windows of features reduced with linear discriminant analysis or speaker-adapted features.

The Neural Network

The enrollment/training phase and the recognition phase are the two main components of any given speaker recognition system. In all stages, the input for modelling comes from the feature extraction component. The enrollment step involves training and storing a speaker model in a database; the recognition phase involves comparing an input sample with the trained and stored speaker models. The match score is compared to a cutoff score to get a verdict. The identification result is often chosen as the speaker model with the highest score. In this chapter, we explore two distinct modelling strategies neural network (NN) and support vector machine (SVM).

The concept of a "neural network" was first developed in the field of biology. Because the human brain is capable of such remarkable recognition tasks, scientists have sought to create a mathematical model that mimics its architecture. It is believed that there are 10 billion neurons (nerve cells) in the human brain, with these interconnected by 60 trillion synapses. Our whole body's worth of data is processed and judgments are made by this network.

Deep Belief Network Acoustic Modeling

In this work, we use deep belief networks (DBNs), a particular type of deep neural networks, for acoustic modeling. A DBN consists of multiple stacked restricted Boltzmann machines (RBMs), each being pre-trained in an unsupervised manner on the actual input features or the hidden representation of the previous one. RBMs are bipartite graphical models in which hidden units learn a representation of visible units. In the standard configuration, both visible and hidden units are binary units that are sampled from a Bernoulli distribution. The probability of being active is computed using weighted connections to the hidden and visible units, respectively.

RBMs are energy-based models, and each configuration of visible units υ and h is assigned an energy term E:

$$E(\boldsymbol{v},\boldsymbol{h}) = -\sum_{i=1}^{V}\sum_{j=1}^{H} v_i h_j w_{ij} - \sum_{i=1}^{V} v_i c_i - \sum_{j=1}^{H} h_j b_j$$

where w_{ij} is the weight assigned to the connection between a visible unit vi and h_j, and c_i and b_j are their bias terms. For modeling real-valued data, which is the usual case for acoustic features, the binary visible units can be replaced with Gaussian units The energy of a configuration becomes:

$$E(\boldsymbol{v},\boldsymbol{h}) = \sum_{i=1}^{V} \frac{(v_i - c_i)^2}{2\sigma^2} - \sum_{j=1}^{H} b_j h_j - \sum_{i=1}^{V}\sum_{j=1}^{H} \frac{v_i}{\sigma} h_j w_{ij}$$

with σ representing the variance of the normal distribution from which the visible units are sampled. Unsupervised learning of a model is done by maximizing the log-likelihood for known configurations (i.e., the training data) as determined by the energy term. The contrastive divergence algorithm provides a fast approximation of this objective by computing the difference of correlations between two configurations obtained by alternating Gibbs sampling.

After pre-training a stack of RBMs, the weights and biases of the hidden units can be used to initialize the hidden layers of a deep belief neural network. When used for discriminative training, an additional classification layer is connected to the last hidden layer, and the resulting network is fine-tuned with standard backpropagation.

Acoustic Modeling

When employing neural networks as acoustic models in combination with hidden Markov models, they are used to compute a posteriori emission probabilities of phone states. If the network is trained to estimate probabilities $p(q_t|x_t)$ of states q_t given observations as input feature vectors x_t using a cross-entropy criterion, the emission probabilities can be obtained with Bayes' rule:

$$p(\boldsymbol{x}_t|q_t) = \frac{p(q_t|\boldsymbol{x}_t)p(\boldsymbol{x}_t)}{p(q_t)}$$

where p(q_t) denotes the prior probability of a phone state, which is estimated using the available training data. During decoding, the most likely sequence of states is computed by the HMM. Since the observation x is independent of the state sequence, its probability p(x_t) can be ignored.

METHODOLOGY

Various methodologies and stages of speaker recognition systems are discussed in this chapter. The main objective of a noise cancellation system is to enhance the speech signal to obtain a clean signal with higher quality. Such system has been widely used in long distance telephony applications. However, the presence of noise in speech signals will contribute to a high degree of inaccuracy in systems that require speech processing such as speech recognition, synthesis and speaker identification systems. If filter of noise has not been carried out appropriately, feature of the noise signal will be extracted together with the feature of the actual speech signal during the feature extraction process. Thus, the desired parametric representation will carry a high amount of inaccuracy. In this work a novel filter is proposed by applying Conventional Neural Network (CNN) ensemble where the noisy signal and the reference one are the same in a learning process. This Neural Network (NN) ensemble filter not only well reduces additive and multiplicative white noise inside signals, but also preserves signals‟ characteristics. It is proved that the reduction of noise using NN ensemble filter is better than the Conventional ε nonlinear filter and single NN filter while signal to noise ratio is low. The performance of the NN ensemble filter is demonstrated in the audio signals processing.

Deep Bottleneck Features Training

The neural network for extracting deep bottleneck features (DBNFs) was trained as described. 30 filterbank coefficients were obtained as described above, normalized on a per-speaker level and concatenated with past and future samples to feature frames consisting of 330 elements. Five auto-encoder layers containing 1000 units each were stacked and pre-trained individually, and a bottleneck layer with 42 units as well as one additional hidden layer and a classification layer were added. The network was then trained to predict context-independent monophone states. A random subset containing 5% of the available training examples was used as a held-out validation set to perform early stopping. The GMM/HMM systems trained on bottleneck features used the phonetic decision tree from the MFCC baseline and therefore ended up with the same number of tied states. As for the baseline, features extracted from 11 adjacent positions were reduced to 42 dimensions with LDA. Speaker-adaptive training was performed using fMLLR. The DBNF system was used to generate new alignments for training the deep belief network acoustic models.

RESULTS AND DISCUSSION

Table 4.1: Statistical Data for Back Propagation Neural model

Neurons in input layer	4
Number of hidden layer and number of neurons In hidden layer	1,4
Neurons in output	1

layer	
Transfer function (input, hidden and output)	Tran sigmoid, Transigmoid, linear
Epochs	1400

Graph of mean square error (MSE) with epochs is shown in Figure 2. It can be concluded from the graph that error reduces with epochs.

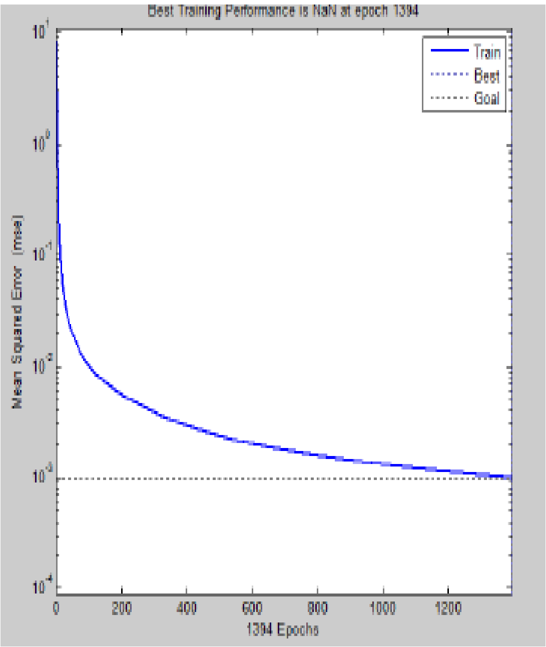
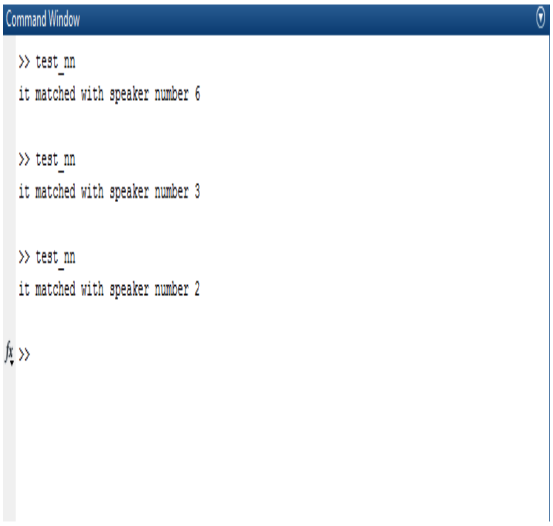


Figure 1: Graph of mean square error (MSE) with epochs

Table: Speaker Identification Rate using NN

NOISE TYPE/ SNR	30dB	20dB	10dB
Babble Noise	92.6	60.5	24.5
Car Noise	90.4	58.6	19.8
Exhibition Hall Noise	92.0	59.4	20.6
Train Noise	89.6	57.8	19.0
Airport Noise	89.4	58.2	19.4

Screenshot of MATLAB Comm and window which show results of Speaker Recognition system is shown in Figure 2. It shows the identity of the speaker with which the unknown speaker's voice matches.



```
>> test_nn
it matched with speaker number 6

>> test_nn
it matched with speaker number 3

>> test_nn
it matched with speaker number 2

fx >>
```

Figure 2: Screenshot of MATLAB Command window

CONCLUSION

These days, mood detection, tension detection, attention detection, and speaker involvement detection are just some of the many uses for speech recognition systems outside of biometric verification. For character, word, and phrase data sets, the speaker recognition systems may be defined in a number of ways. These voice samples were recorded in various settings and hence include environmental, technological, and interference artefacts. Both the accuracy and speed of speech/speaker recognition systems suffer when they are subjected to this sort of distorted capture. For this reason, the suggested speaker identification system is built to efficiently manage a wide range of such interference and speech noise. With the results obtained above, we have demonstrated that bottleneck features are useful input features for DBN/HMM speech recognition setups. It could be shown that the modular combination proposed enables the acoustic model to use an increased temporal context of acoustic features more efficiently than an identical network trained directly on the input features. The performance improvements achieved by using deep bottleneck features for the hybrid DBN/HMM systems are significant, though not as large as for the GMM/HMM baseline system. However, deep neural networks have a much higher modeling capacity then GMMs, and it is to be expected that a good part of the modeling performed in the bottleneck network can be learned in a standalone DBN as well. Nevertheless, we regard our approach to combining neural networks for acoustic modeling as promising and the general principles of modularity as an important paradigm that is applicable to deep neural networks as it is to shallow ones.

REFERENCES

1. *G. Dahl, D. Yu, L. Deng, and A. Acero, (2012) "Context-dependent pretrained deep neural networks for large-vocabulary speech recognition," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 1, pp. 30–42,.*
2. *F. Seide, G. Li, X. Chen, and D. Yu, (2011) "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on. IEEE, pp. 24–29.*
3. *Leandro D. Vignolo, S.R. Mahadeva Prasanna, Samarendra Dandapat, H. Leonardo Rufiner, Diego H. Milone, (2016) Feature optimization for stress recognition in speech, Pattern Recognition Letters, Vol. 84, Pages 1-7.*
4. *Y. Zong, W. Zheng, Z. Cui and Q. Li, (2016) Double sparse learning model for speech emotion recognition, Electronics Letters, Vol. 52, Issue 16, pages 1410-1412,.*
5. *SonayKammi, Mohammad Reza KaramiMollaei, (2017) Noisy speech enhancement with sparsity regularization, Speech Communication, Vol. 87, Pages 58-69.*
6. *154. Qingyang Hong, Lin Li, Jun Zhang, Lihong Wan, Huiyang Guo, (2017) Transfer learning for PLDA-based speaker verification, Speech Communication, Vol. 92, Pages 90-99.*
7. *T. Thiruvaran, V. Sethu, E. Ambikairajah and H. Li, (2015) Spectral shifting of speaker-specific information for narrow band telephonic speaker recognition, Electronics Letters, Vol. 51, Issue 25, Pages 2149-2151.*
8. *Yong-Sun Choi, Soo-Young Lee,(2013) Nonlinear spectro-temporal features based on a cochlear model for automatic speech recognition in a noisy situation, Neural Networks, Vol. 45, Pages 62-69.*