International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 9 (2015) pp. 22991-22998

http://www.ripublication.com

Compare The Perform of Voice Recognition Using Deep Neural Network and Fuzzy Logic

J.Seethalakshmi[1]
[1] *PhD Research Scholar, Bharathiar University, Coimbatore, India*
[1]*Assistant Professor, S.S.K.V College of Arts and Science for Women, Kanchipuram*
[1]*Email : seetha2207@gmail.com*
C. Jayakumar [2]
[2] *PhD Research Supervisors, Bharathiar University, Coimbatore , India*
[2] *Professor, CSE Department, RMK Engineering College, Kavaraipetai*
[2] *Email : cjayakumar2007@gmail.com*

Abstract

In this paper, we compare the performance of voice recognition using Hidden Markov models (HMM) in Deep Neural Networks (DNN) and Fuzzy Logic. The data sets used are speech from The DARPA TIMIT Acoustic-Phonetic Speech Corpus. Currently, most speech recognition systems are based on Hidden Markov Models, a statistical framework that supports both acoustic and temporal modeling. Despite their state-of-the-art performance, HMMs make a number of suboptimal modeling assumptions that limit their potential effectiveness. The recognition process consists of the Training phase and the Testing (Recognition) phase. The audio files from the speech corpus are preprocessed and features like Short Time Average Zero Crossing Rate, Pitch Period, Mel Frequency Cepstral Coefficients (MFCC), Formants and Modulation Index are extracted. The model database is created from the feature vector using HMM/DNN and is trained with deep belief network (DBN) to initialize the parameters of a deep Neural network (DNN) algorithm. During recognition the test set model is obtained which is compared with the database model. The same sets of audio files are trained for the speech recognition using HMM/Fuzzy and the fuzzy knowledge base is created using a fuzzy controller. During the recognition phase, the feature vector is compared with the knowledge base and the recognition is made. From the recognized outputs, the recognition accuracy (%) is compared and the best performing model is identified. Recognition accuracy (%) using Deep Neural Networks were found to be superior to recognition using Fuzzy.

Keywords: Voice Recognition, Hidden Markov Model, Deep Neural Networks, deep belief network, Fuzzy, Sentence Recognition, Recognition Accuracy.

Introduction

In Automatic Speech Recognition (ASR) systems, the computer must adapt to different voices used as input. Human listeners are more flexible in adapting to a machine's accent than a computer is in deciphering human accents. Many systems adapt via learning procedures. The input speech gets modified into patterns and is stored in memory as models. The complexity of the search increases with the length of the utterance and ASR cannot accept long words as input. Segmenting utterances into smaller units for ASR simplifies computation and aids accuracy by reducing the search space. Clear pronunciation of each word reduces effects of co-articulation and renders each word in a test utterance closer in form to word utterances in the reference patterns. Most ASR systems use accuracy or error rates to measure performance. We mostly employ accuracy as the evaluation criteria as it is the one most often cited. Evaluation of ASR performance requires databases of speech labeled with textual transcriptions. The DARPA TIMIT Acoustic-Phonetic Speech Corpus contains a total of speech. It includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16 kHz speech file for each utterance.

Speaker recognition, which can be classified into identification and verification, is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers [10]

The most common ASR parameters are mel-based spectral coefficients, zero crossing rate, formants etc. Enough spectral information is captured to identify spoken phones. Features represent an intermediate step between parameters and phonetic features. Feature extraction is most common in expert systems. Each feature pattern for a frame of speech can be viewed as an N-dimensional vector having N parameters per frame. If parameters are well chosen, then separate regions can be established in the N-space for each segment. A memory of reference model is characterized by an N-dimensional feature vector during training.

Deep-neural-network HMMs, or DNN- HMMs [5, 6], apply the classical ANN-HMMs of the 90's to traditional tied-state triphones directly, exploiting Hinton's deep-belief-network (DBN) pre-training procedure. This was shown to lead to a very promising and possibly disruptive acoustic model as indicated by a 16% relative recognition error reduction over discriminatively trained GMM-HMMs on a business search task [5, 6], which features short query utterances, tens of hours of training data, and hundreds of tied states. For simple ASR tasks, DNN's can provide high accuracy

HMM's, but they are used in efficient training of HMM probabilities. Hybrid HMM/DNN systems keep the basic structure and recognition process, but estimate the

Probability Density Functions via ANN's in training[1,2]. In speech recognition using Fuzzy logic, a model variable is described in terms of fuzzy space. This space is generally composed of multiple overlapping sets, each fuzzy set describing a semantic partition of the variables. The audio files from the database are trained for the speech recognition system using HMM/Fuzzy and the fuzzy knowledge base is created using a fuzzy controller. During the recognition phase, the feature vector is compared with the knowledge base and the recognition is made. From the recognized outputs, the recognition accuracy is considered as a parameter for comparison. The performance of speech recognition using HMM/DNN and HMM/Fuzzy are compared and the best performing classifier is reported.

Preprocessing and Feature Extraction

Human speech can be represented as an analog wave that varies over time. The height of the wave represents intensity (loudness), and the shape of the wave represents frequency (pitch). The properties of the speech signal changes relatively slowly with time. This allows examination of a Short-time window of speech to extract parameters presumed to remain fixed for the duration of the window. The signal must be divided into successive windows or analysis frames so that the parameters can be calculated often enough to follow the relevant changes. The result of signal analysis is a sequence of speech frames. To extract the features from the speech signal, the signal must be preprocessed and divided into successive windows or analysis frames.

Each sentence was taken through different stages of preprocessing which included Preemphasis, Frame Processing and Windowing [5, 8]. The higher frequencies of the speech signal are generally weak. As a result there may not be high frequency energy present to extract features at the upper end of the frequency range. Pre-emphasis is used to boost the energy of the high frequency signals. Frame blocking is a process adopted to split the speech signal into frames The speech samples are segmented into 32 ms frames with each frame having 50% overlap with the adjacent frames. The next step in preprocessing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of the frame. To minimize the signal discontinuities Hamming window is used which has the form.

$$W(n) = 0.54 - 0.46 \cos\left(\frac{2\Pi n}{N-1}\right) 0 \le n \le N-1$$

The result of windowing is

$$X(n) = x_1(n)\, w(n),\ 0 \le n \le N-1$$

Analysis Frame

The purpose of feature extraction is to represent the speech signal by a finite number of measures of the signal. It gives the invariant representations in the signal. The features selected are the Short Time Average Zero Crossing Rate [7], Pitch Period Computation, Mel Frequency Cepstral Coefficients (MFCC), Formants and Modulation Index. The more features we use, the better the representation.

Deep Neural Network HMM

A deep neural network (DNN) is a conventional multi-layer perceptron (MLP, [8]) with many hidden layers, optionally initialized using the DBN pre-training algorithm. In the following, we want to recap the DNN from a statistical viewpoint and describe its integration with HMMs for speech recognition. For a more detailed description, please refer to [6].

DBN Pre-Training

The deep belief network (DBN), proposed by Hinton [11], provides a new way to train deep generative models. The layerwise greedy pre-training algorithm developed in DBN was later found to be also effective in training DNNs. The DBN pre-training procedure treats each consecutive pair of layers in the MLP as a *restricted Boltzmann machine* (RBM) [11] whose joint probability is defined as

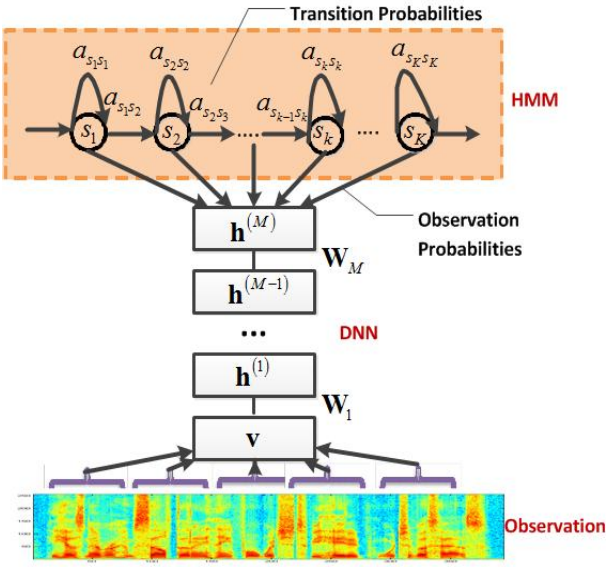
$$P_{h,v}(h, v) = \frac{1}{Z_{h,v}} \cdot e^{v^T W h + V^T b + a^T h}$$

Integrating DNNs with HMMs

Following the traditional ANN-HMMs of the 90's [1], we replace the acoustic model's Gaussian mixtures with an MLP and compute the HMM's state emission likelihoods $P_{o|s}(o|s)$ by converting state posteriors obtained from the MLP to likelihoods:

$$P_{o|s}(o|s) = \frac{P_{s|o}(s|0)}{P_s(s)} \cdot \text{const}(s). \quad (2)$$

Here, classes *s* correspond to HMM states, and observation vectors *o* are regular acoustic feature vectors augmented with neighbor frames (5 on each side in our case). ***P*s**(*s*) is the prior probability of states. This is a critical factor in achieving the unusual accuracy improvements in this paper.



Training DNN-HMMS

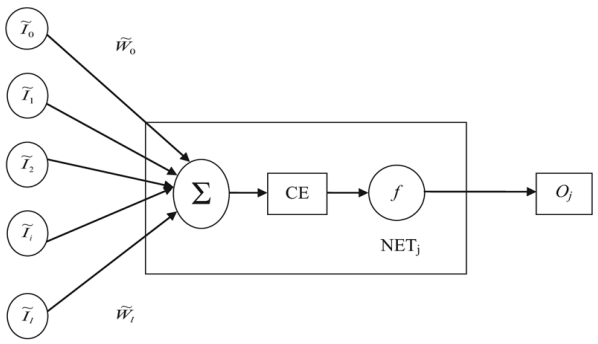
In this section, we will describe the process and some practical considerations in training DNN-HMMs.

Basic Training Process

DNN model learning begins with the DBN pre-training (section 2.2), using one full sweep through the 309 hours of training data for all hidden layers but the first, where we use two full sweeps. Slight gains may be obtained if the pre-training procedure sweeps the data additional times. However, this seems to be not critical [5]. RBMs are not scale-invariant, so the training corpus is normalized to zero mean and unit variance [13]. The alignment is updated once during training.

Fuzzy Based Speech Recognition

Rule based Fuzzy Systems are fundamental methodologies to represent and process linguistic information, with mechanisms to deal with uncertainty and imprecision. One of the most important tasks in the development of Fuzzy systems is the design of its knowledge base. Fuzzy control directly uses fuzzy rules in the fuzzy theory and creates a fuzzy controlled machine which has a fuzzifier, rule evaluator and then a defuzzifier. The Fuzzifier converts the input into a linguistic variable using the membership functions stored in the fuzzy knowledge base. The Rule evaluator converts the fuzzy input into a fuzzy output using If-Then fuzzy rules. The defuzzifier converts the fuzzy output of the inference engine to crisp membership functions used by the fuzzifier. The fuzzy knowledge base is the information storage for Linguistic variables definitions and fuzzy rules. The knowledge base containing the clusters of features is created as a Fuzzy knowledge Base [3,9].



Fuzzy Neuron *J*

Training and Recognition Process Using Classifiers

The HMM/DNN is trained with feature vectors [4,10] grouped as a model database. In a finite amount of time, it successfully learns to distinguish between the feature vectors. The training process involves calculation of weight functions (hidden layer), getting the response and calculation of error. It updates the weights based on error and learning rate. In the testing phase (Fig.3) the same network is tested with new feature vectors using MATLAB codes. It gives an output '1' (full activation) for the

identified vectors and '0' for others. The error rate decreased and accuracy improved with more learning iterations.

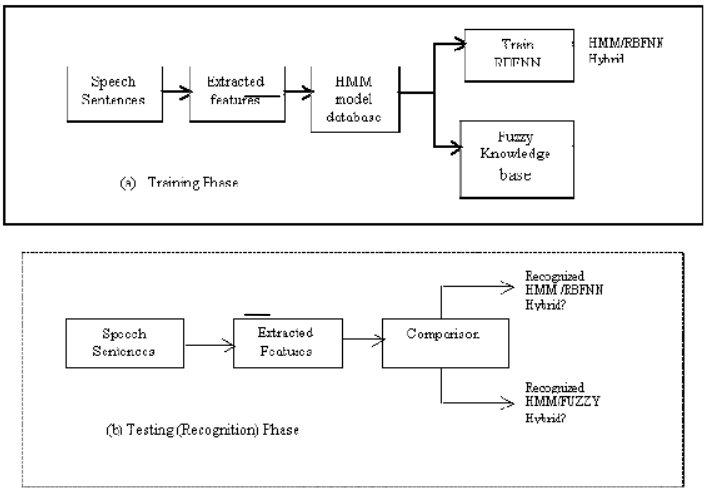


Figure 3: Training and Recognition Process

In speech recognition using Fuzzy logic, a model variable is described in terms of fuzzy space. This space is generally composed of multiple overlapping sets, each fuzzy set describing a semantic partition of the variables. The audio files from the database are trained for the speech recognition system using HMM/Fuzzy and the fuzzy knowledge base is created using a fuzzy controller [6]. MATLAB codes were used. During the recognition phase, (Fig.3) the feature vector is compared with the knowledge base and the recognition is made. From the recognized outputs, the accuracy% is considered as a parameter for comparison, the performance of the speech recognition system using HMM/ANN and HMM/Fuzzy are compared in terms of recognition accuracy rate (%) and the best performing classifier is found.

The CD-GMM-HMM Baseline Result

Criterion Test Accuracy	
DNN/HMM	FUZZY/HMM
99%	97.7%.

In the training phase, 800 words were taken from the database "The DARPA TIMIT Acoustic- Phonetic Continuous Speech Corpus". These words are spoken by different speakers. Irrespective of the gender, speaker independent datasets were taken for the study. Speaker independent task of recognition is more difficult than the speaker dependent one. The features extracted were Zero crossing Performance of Speech Recognition using Artificial Neural Network and Fuzzy Logic 46 rate, MFCC, Pitch Period, Formants and Modulation Index and corresponding feature vectors were obtained. The HMM model database is computed using HTK toolkit. Model database was created for the training. Two test data sets were chosen with an overlap of 400

swords. Minimum error identifies and recognizes each of the tested words. The performance for DNN/HMM classifier was found to be 99% recognition accuracy and Fuzzy classifier had an accuracy of 97.7%. The DNN/HMM had a superior performance [12] when compared to the Fuzzy classifier.

Conclusion

The results show that by taking into account the interaction among subsets of information sources, the DNN/HMM based technique amply outperforms simpler methods such as Fuzzy. The Deep Neural Network classifier has been found to have a better performance than the Fuzzy classifier. The performance of recognition of isolated words by DNNFNN is very high (99%). But it still performs better when compared to the performance of Fuzzy classifier. Achieving up to 33% relative WER reduction over a discriminatively trained Fuzzy-HMM. Our results suggest that the three critical factors that contribute to the remarkable accuracy gains from the DNNHMMs. We have considered increasing the database of words, and with the modular approach we have been able to achieve about 99% recognition rate on over 800 words. We still have to make more tests with different words and levels of noise.

References

[1] Hongbin SUO1 , 2 , Ming LI1, Ping LU1 and Yonghong YAN, " Automatic Language Identification with Discriminative Language Characterization Based on SVM," IEICE-Transactions on Info and

[2] Proceedings. (ICASSP '02). IEEE International Conference on , vol.1, no., pp. I757-I-760 vol.1, 2002.

[3] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker Verification Using Adapted Gaussian, Mixture Models, Digital Signal Processing," Vol. 10, 19–41 (2000).

[4] Melin, and O. Castillo, "A New Method for Adaptive Control of Non-Linear Plants Using Type-2 Fuzzy Logic and Neural Networks", *International Journal of General Systems*, Taylor and Francis, Vol. 33, 2004, pp. 289-304

[5] D. Rumelhart, G. Hinton, and R. Williams, "Learning Representations By Back-Propagating Errors," Nature, vol. 323, Oct. 1986.

[6] G. Hinton, S. Osindero, and Y. Teh, "A Fast Learning Algorithm for Deep Belief Nets", Neural Computation, vol. 18, pp. 1527–1554, 2006.

[7] G. Hinton, "A Practical Guide to Training Restricted Boltzmann Machines", Technical Report UTML TR 2010–003, University of Toronto, 2010.

[8] A. Mohamed, G. Dahl, and G. Hinton, "Deep Belief Networks for Phone Recognition," in Proc. NIPS Workshop Deep Learning for P. Melin, A. Mancilla, C. Gonzalez, and D. Bravo, "Modular Neural Networks with

Fuzzy Sugeno Integral Response for Face and Fingerprint Recognition", Proceedings of IC-AI'04, Las Vegas, USA, 2004, pp. 91-97.

[9] Yoon, T., Zhuang, X., Cole, J., Hasegawa-Johnson, M., "Voice quality dependent speech recognition", In Tseng, S. (Ed.), Linguistic Patterns of Spontaneous Speech, Special Issue of Language and Linguistics, Academica Sinica, 2007.

[10] Anil K. Jain, Multimodal User Interfaces: Who's the User? International Conference on Multimodal Interfaces, Documents in Computing and Information Science, 2010.

[11] L. Deng and D. Yu, "Deep convex network: A scalable architecture for speech pattern classification," in Proc. Interspeech, 2011. IEEE SIGNAL PROCESSING MAGAZINE [16] November 2012

[12] V. Vanhoucke, A. Senior, and M. Z. Mao, "Improving the speed of neural networks on CPUs," in Proc. Deep Learning and Unsupervised Feature Learning NIPS Workshop, 2011.

[13] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Improvements in using deep belief networks for large vocabulary continuous speech recognition," Speech and Language Algorithm Group, IBM, Tech. Rep. UTML TR 2010-003, Feb. 2011.

[14] http://www.eie.polyu.edu.hk/~mwmak/Book