A Literature Survey on Recurrent Neural Network and Various Techniques for Speech Recognition

Amitkumar Panchal[1], **Ompriya Kale**[2]

[1, 2]Department of Computer Engineering, Gujarat Technological University, Gujarat, India

Abstract: ***Speech is the vocalized form of human communication. Each spoken word is created out of the phonetic combination of a limited set of vowel and consonant speech sound units. Speech Recognition (SR) is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format. SR has evolved quite a bit over the past few years and the initially it used to work in discrete dictation mode where you had to pause and stop between every spoken word. Today, however it uses continuous dictation; It is also become smarter with its own set of grammar rules to make out the meaning of what is being said. In this work, different method is used for Speech Recognition (SR) in Recurrent Neural Network (RNN) which improve the performance SR.***

Keywords: Speech Recognition, Recurrent Neural Network, Deep Neural Network, Mel-frequency cepstrum, Principal Component Analysis

1. Introduction

Speech Recognition (SR) is the process of converting a speech signal to a sequence of words [7]. The standard approach to large vocabulary continuous SR is to assume a simple probabilistic model of speech production whereby specified the word sequences, W produces an acoustic observation sequence Y with probability P(W,Y).The goal is then decode word string based on the Y [7].

1) Types of speech
   There are two types of speech:
   Discrete speech consists of isolated words that are separated by silences. Continuous speech consists of words will be spoken without silences [6].
2) Speaker dependence
   There are three types of Speaker dependence. One is Speaker dependent system, Second is Speaker independent system, and Third is Speaker adaptable system [6].
3) Size of vocabulary
   The vocabulary is the set of words that have to be recognized. A small vocabulary contains less than about 30 words. A 500 word is vocabulary is average size. A vocabulary with more than 25000 words generally will be seen as very big [6].
4) Types of Speech Recognition (SR)
   - **Isolated words recognition**
     It accepts single words or single utterance at a time. Usually doing processing during the pauses [7]. One key problem with this technology was the sensitivity to background noises, and extraneous speech which was inadvertently spoken along wih the command word. The Various types of' keyword spotting' algorithms evolved to solve these types of problems [9].
   - **Connected words recognition**
     It is similar to isolated words, but allows separate utterances to be run-together with a minimal pause between them [7]. This technology was built on top of word recognition technology, choosing to exploit the word models that were successful in isolated word recognition and extend the modeling to recognize a concatenated sequence of such word models as a word string [9].
   - **Continuous speech recognition**
     Recognizers with continuous speech capabilities are some of most difficult to create because they utilize special methods to determine utterance boundaries [7]. This technology led to the first large vocabulary recognition systems which were used to access databases (the DARPA Resource Management Task) [9].
5) Relevant issue of Automatic Speech Recognition (ASR) design are :
   Environment, Transducer, Channel, Speakers, Speech styles, Vocabulary [7] are main issues on which recognition accuracy depends have been presented it.
6) SR techniques
   - Acoustic phonetic approach [7] which represent spectral analysis with feature detection phonemes/ segmentation and labeling.
   - Pattern recognition approach [7] which represent speech, pixels and curves, samples, set of spectral vectors, features, set of sequence of spectral vector .
   - Neural network approach [7] which represent speech feature/rules/perceptron/units/procedures.
   - Support Vector Machine (SVM) [7] which represents Kernel based feature.
   - Artificial intelligence [7] which represents Knowledge based.

Various methods for Feature extraction method [7] in speech recognition in table 1:

Table 1: Feature extraction method [7]

Method	*Property*
Principal Component Analysis (PCA) [7]	Non linear feature extraction method, linear map, eigenvector-based
Linear Discriminant Analysis (LDA) [7]	Non linear feature extraction method, Supervised linear map, eigenvector-based
Mel-frequency cepstrum (MFCCs) [7, 10]	Power spectrum is computed by performing fourier analysis

Neural Networks (NN) have found their way into Voice Activity Detection in Recurrent NN [1].

The whole statistical Automatic Speech Recognition (ASR), system consists of four major components. During the search all these four sources are combined to obtain the optimal word sequence.

The signal analysis extracts acoustic features from the input speech signal. Afterwards, the sequence of acoustic features is passed on to the speech recognizer.

The acoustic model consists of statistical models for the smallest sub-word units, to be distinguished by the speech recognizer, e.g. phonemes, syllables or whole words, and a pronunciation lexicon which defines the composition of an acoustic model for a given word from the sub-word units.

The language model provides the a priori probability of a hypothesized word sequence, based on the syntax, semantics and pragmatics of the language to be recognized.

The search algorithm combines the acoustic model, and the language model. The final hypothesis of the search is the word sequence which maximizes Equation. The full search space for continuous speech recognition, for optical character recognition, and for automatic sign language recognition consists of all possible word sequences, which can be produced by a (finite) vocabulary.

2. Related work

The following are the survey done on various methods for speech recognition:

2.1 Recurrent Neural Networks for Voice Activity Detection [1]

In this paper, present a novel Recurrent Neural Network (RNN) model for voice activity detection. Multi-layer RNN model [6], in which nodes compute a quadratic polynomials, the outperforms a much larger baseline system composed of Gaussian Mixture Models (GMMs) and the hand-tuned State Machine (SM) for temporal smoothing.

In RNN Voice Acitvity Detection (VAD) architecture, is different from MLP like RNNs others have applied to speech de-noising [11]. VAD is a Feed-forward Neural Network (FFNN) [12] with recurrence added at various points. Nodes marked with S have tanh non-linearities and others achieve non-linearity by evaluation quadratic polynomials [5] function of its inputs.

An arrow pointing towards a node represents that a node receives as inputs the values of the nodes at the arrow's tail.Then, compute the RNN's error at each timestep by running it on training data. It compute the difference of the RNN's output node $N_{output}[T]$ with a slightly delayed target output:

$$\mathrm{Error}[T] = N_{output}[T] - \mathrm{Target}[T - \Delta] \qquad [1]$$

Where Δ is fixed delay which allows the RNN to process input frames.

In Training procedure, a Ceres solver which is used for optimization. Then an Automatic differentiation which is used recurrent because node outputs are fed back as input. Last apply, Two-stage training, first stage, it fix all recurrent parameter and only train the feed forward parameters, and second stage it optimize all the parameter together, including W_L weights controlling the tapped delay line.

Here, it training the speech "hello world" for voice activity detection. Experiments are shown demonstrating the feasibility of the approach for VAD.

Using this method, reducing per-frame false alarm (FA) rate by 26% and increasing overall recognition speed by 17% with a modes 1% relative decrease in word error rate (WER).Also it can be apply using Gradient-based optimization for better speech recognition.

2.2 Multilingual Acoustic Models Using Distributed Deep Neural Networks [2]

In this paper, present an experimental results for cross and multi lingual network training of eleven Romance languages on 10k hours of data in total.

In multilingual architecture, represents for speech recognition with language-independent feature extraction, and language specific classifiers on top of it.

In cross and multi-lingual approaches to learning based on knowledge transfer. First, feature learning uses the weights of bottom three hidden layer of source network in the target network, and keeps these weights fixed during training of target networks. Second, transfer learning uses the situation where the network of target language is initialized with an existing source network and last multitask learning, multiple tasks are learned in parallel, and use a shared representation.

Using distributed deep neural network (DNN), find out optimization procedure for training DNNs.

In experiments results, provide data and setup for multilingual training for a set of Romance language. Data is taken from Voice search and test on it.

Using this method, conclude that corresponding multilingual neural network is better than monolingual neural network for training.

2.3 Multiframe Deep Neural Networks For Acoustic Modeling [3]

In this paper, present a method of tying neural network parameter over time which achieves comparable performance to typical frame-synchronization model.

DNN have a more parameters than typical GMMs. First, an Error rate against complexity of neural network acoustic models for US English, trained on thousands of hours of data, and Iberian Portuguese, trained on 100 hours of data.

Second, Frame synchronous baseline approach is used for training data. Third, Frame asynchronous, acoustic model running at half the frame rate of feature stream and predictions are simply copied every other frame. Fourth, mutiframe prediction, a neural network is trained to issue jointly for multiple consecutive frames. And last, found decoding speed. Using this method, 4X reduction in the computational cost of a neural network activations.

2.4 ASR Error Detection Using Recurrent Neural Network Language Model and Complementary ASR [4]

In this paper, train a neural network predictor of errors using a variety of features.

Given an ASR system, decode training utterances {X} to generate confusion networks and ASR hypotheses. It uses a binary y = 0/1 label on each hypothesized word to indicate error or correct. Given a training examples {(xi, yi)} where xi denotes a feature vector of the i-th training example, it employ a feed-forward neural network (FFNN) classifier to predict the word confidence p(yi|xi). A neural network classifier is trained using back propagation.

First step is baseline feature, extract contextual features from a confusion network at eachword position j of the ASR hypothesis W = w1w2...wj ...wN.

Second step is RNNLM features, a RNNLM p(wj |wj−1, hj−1) has a recursive structure that predicts a current word wj given the previous word wj−1 and previous hidden state vector hj−1.

Third step is Incremental unsupervised RNNLM adaption. Fourth step is combining complementary ASR.

In experiments results, Data is taken from DARPA english TRANSTAC dataset using MFCC (Mel-Frequency Cepstrum) feature. This result have shown significant improvement in Automatic Speech Recognition (ASR) error prediction using state-of-the art DNN ASR with proposed approaches.

2.5 Extensions of Recurrent Neural Network Language Model [5]

In this paper, present language model approaches that lead to more than 15 times speedup for both training and testing phases. Then show importance of using a backpropagation through time (BPTT) algorithm.

First, present a simple recurrent neural network. The vector w(t) is current word size. Vector s(t-1) represents output values in hidden layer from previous time step. The network is trained by using BPTT. Second, BPTT algorithm is used. Then speed up techniques of time complexity one training step is represented.

$$O = (1+H) \times H \times \tau + H \times V \qquad [5]$$

Where, H is the size of the hidden layer, V size of the vocabulary and τ the amount of steps we backpropagate the error back in time. After that it uses Factorization of output layer and compression layer. In experiments results, Data is taken from Penn corpus. In conclusion, found that RNN model can thus be smaller, faster both during training and testing and more accurate than the basic one.

3. Conclusion

The main contribution of this study is that it presents the idea of Speech Recognition using Recurrent Neural Network (RNN). The study shows that these approaches can be used to increase accuracy of noisy test data for Speech Recognition. Our effort will be directed toward developing the more appropriate and convenient method.

References

[1] Thad Hughes and Keir Mierle, “RECURRENT NEURAL NETWORKS FOR VOICE ACTIVITY DETECTION,” Google inc., USA. IEEE, 2013, pp. 7378-7382.

[2] G.Heigold, V. Vanhoucke, A. Senior, P.Nguyen, M. Ranzato, M. Devin and J. Dean, “MULTILINGUAL ACOUSTIC MODELS USING DISTRIBUTED DEEP NEURAL NETWORKS,” *Google Inc*., USA. IEEE, 2013, pp. 8619-8623.

[3] Vincent Vanhoucke, Matthieu Devin, Georg Heigold, “MULTIFRAME DEEP NEURAL NETWORKS FOR ACOUSTIC MODELING,” Google, Inc., USA. IEEE, 2013, pp. 7582-7585.

[4] Yik-Cheung Tam, Yun Lei, Jing Zheng and Wen Wang, “ASR ERROR DETECTION USING RECURRENT NEURAL NETWORK LANGUAGE MODEL AND COMPLEMENTARY ASR,” 2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014, pp. 2331-2335.

[5] Tom´aˇs Mikolov, Stefan Kombrink, Luk´aˇs Burget, Jan “Honza” ˇCernock´y, Sanjeev Khudanpur2, “EXTENSIONS OF RECURRENT NEURAL NETWORK LANGUAGE MODEL, ” 2011 IEEE ICASSP, 2011, pp. 5528-5531.

[6] Ms. Vrinda, Mr. Chander Shekhar, “SPEECH RECOGNITION SYSTEM FOR ENGLISH LANGUAGE,” International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 1, January 2013.

[7] Sanjivani S. Bhabad, Gajanan K. Kharate, “An Overview of Technical Progress in Speech Recognition, ”International Journal of Advanced Research in Computer Science and Software Engineering Vol. 3, Issue 3, March 2013.

[8] Ilya Sutskever, James Martens, Geoffrey Hinton, “*Generating Text with Recurrent Neural Network,*” Proceedings of the 28 th International Conference on Machine Learning, Bellevue, WA, USA, 2011.

[9] Lawrence R. Rabiner, “APPLICATIONS OF SPEECH RECOGNITION IN THE AREA **OF** TELECOMMUNICATIONS,” AT&T Labs IEEE 1997, pp. 501-510.

[10] Abhishek Thakur, Rajesh Kumar, Naveen Kumar, “Automatic Speech Recognition System for Hindi Utterances with Regional Indian Accents: A Review,” IJECT Vol. 4, Issue Spl – 3, April – June 2013.

[11] Andrew L. Maas, Quoc V. le, Tyler M. O’Neil, Oriol Vinyals, Patrick Nguyen, Andrew Y. Ng, “Recurrent Neural Networks for Noise Reduction in Robust ASR, ” Google, Inc., USA

[12] M. Sundermeyer, I. Oparin, J.-L. Gauvain, B. Freiberg, R. Schluter, H. Ney, “COMPARISON OF FEEDFORWARD AND RECURRENT NEURAL NETWORK LANGUAGE MODELS, ” 2013 IEEE ICASSP, 2013, pp. 8430-8434