ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Speech Processing Of Tamil Language With Back Propagation Neural Network And Semi-Supervised Traning

N.Pushpa[1], R.Revathi[2], C.Ramya[3], S.Shahul Hameed[4]

ME (CSE), Karpagam University, Coimbatore[1]

ME (CSE), Narasu's Sarathy Institute of Technology, Salem[2]

Assistant Professor, Narasu's Sarathy Institute of Technology, Salem[3]

Assistant Professor, Department of Computer Science and Engineering, Karpagam University, Coimbatore[4]

ABSTRACT: Speech recognition has been an active research topic for more than 50 years. Interacting with the computer through speech is one of the active scientific research fields particularly for the disable community who face variety of difficulties to use the computer. Such research in Automatic Speech Recognition (ASR) is investigated for different languages because each language has its specific features. Especially the need for ASR system in Tamil language has been increased widely in the last few years. In this paper, a speech recognition system for individually spoken word in Tamil language using multilayer feed forward network is presented. To implement the above system, initially the input signal is preprocessed using four types of filters namely preemphasis, median, average and Butterworth bandstop filter in order to remove the background noise and to enhance the signal. The performance of these filters are measured based on MSE and PSNR values. The best filtered signal is taken as the input for the further process of ASR system. The speech features being the major part of speech recognition system, are analyzed and extracted via Linear Predictive Cepstral Coefficients (LPCC). These feature vectors are given as the input to the Feed-Forward Neural Network for classifying and recognizing Tamil spoken word. We propose a technique for training deep neural networks (DNNs) as data-driven feature front-ends for large vocabulary continuous speech recognition (LVCSR) in low resource settings. To circumvent the lack of sufficient training data for acoustic modelling in these scenarios, we use transcribed multilingual data and semi-supervised training to build the proposed feature front-ends.

I. INTRODUCTION

Automatic Speech Recognition (ASR) deals with automatic conversion of acoustic signals of a speech utterance into text transcription. Even after years of extensive research and development, accuracy in ASR remains a challenge to researchers. There are number of well known factors which determine accuracy. The prominent factors are those that include variations in context, speakers and noise in the environment. Therefore research in ASR has many open issues with respect to small or large vocabulary, isolated or continuous speech, speaker dependent or independent and environmental robustness.

ASR for western languages like English and Asian languages like Chinese is well matured. But similar research in Indian languages is still in its infancy stage. Another major hurdle in ASR for Indian language is resource deficiency. Annotated speech corpora for training and testing the acoustic models are scarce. Recently there is a growing interest in ASR for Tamil and other Indian languages. There are speech recognition works for Tamil language which are targeted towards low and restricted vocabulary task [1]. There are some funded research works in spoken digit recognition [2]. Others have

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

attempted to speech recognition for isolated word recognition in Tamil using Artificial Neural Network (ANN) [3]. Some research has been reported to tackle the issue of resource deficiency by means of cross language transfer and cross language adaptation techniques [4].

There are other literatures concerning large vocabulary continuous speech recognition (LVCSR) in Indian languages including Tamil. These works are mainly concentrated and targeted towards exploiting the syllabic nature of Tamil and other Indian languages. Basically there are three approaches to speech recognition with respect to the choice of sub-word units namely, word based, phone based and syllable based recognition.

II. PROBLEM FORMULATION

Fundamentally, the problem of speech recognition can be stated as follows. When given with acoustic observation X = X1X2…Xn, the goal is to find out the corresponding word sequence W = w1w2…wm that has the maximum posterior probability P (W|X) expressed using Bayes Theorem as shown in equation (1).

$$W=\arg\max_{w} P(W|X)=\arg\max_{w} \frac{P(W)P(X|W)}{P(X)} \quad (1)$$

P (X|W) is the probability of acoustic observation X when word W is uttered. P (X|W) is also known as class conditioned probability distribution. P(X) is the average probability that the observation X will occur. Since the maximization of equation (1) is done with variable X fixed, to find the word W it is enough to maximize the numerator alone.

$$W= \arg\max_{w} (P(W)P(X|W)) \quad (2)$$

The first term in equation (2), P (W), is computed with the help of a language model. It describes the probability associated with a hypothesized sequence of words. The language model incorporates both the syntactic and semantic constraints of the language and the recognition task. Generally the language model may be of the form of a formal parser, syntax analyzer, N-gram model or a hybrid model [5].

The second term in equation (2), P (X|W), is computed using an acoustic model which estimates the probability of a sequence of acoustic observations conditioned on the word W. The recognizer needs to know the class conditioned probability P (X|W) from the acoustic model in order to compute the posteriori probability P (W|X). HMM has become the common structure of acoustic models because HMM can normalize speech signal's time-variation and characterize speech signal statistically thus helping to parameterize the class conditioned probabilities. Thus the acoustic model forms the core knowledge base representing various parameters of speech in optimal sense. Even though speech decoding with other classification models like neural networks and support vector machines are also reported in the literature[6], at present, all state-of-the-art commercial and most laboratory speech recognition systems are based on HMM that give very low WER when tested on standard speech databases [7][8].

III. THE TAMIL LANGUAGE

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Tamil is a Dravidian language spoken predominantly in the state of Tamilnadu in India and Sri Lanka. It is the official language of the Indian state of Tamilnadu and also has official status in Sri Lanka and Singapore. With more than 77 million speakers, Tamil is one of the widely spoken languages of the world.

3.1 Pronunciation in Tamil

Tamil has its unique letter to sound rules. There are very restricted numbers of consonant clusters. Tamil has neither aspirated nor voiced stops. Unlike most other Indian languages, Tamil does not have aspirated consonants. In addition, the voicing of plosives is governed by strict rules. Plosives are unvoiced if they occur word-initially or doubled. The Tamil script does not have distinct letters for voiced and unvoiced plosives, although both are present in the spoken language as allophones.

Generally languages structure the utterance of words by giving greater prominence to some constituents than others. This is true in the case of English: one or more phones standout as more prominent than the rest. This is typically described as word stress. The same is true for higher level prosody in a sentence where one or more constituent may bear stress or accent. As far as Tamil language is concerned, it is assumed that there is no stress or accent in Tamil at word level and all syllables are pronounced with the same emphasis. However there are other opinions that the position of stress in the word is by no means fixed to any syllable of individual word. In connected speech the stress is found more often in the initial syllable. Detailed study on pronunciation in Tamil can be found in [9] [10]. In our experiment, stress on syllable is ignored because we are dealing with read speech.

IV. SYSTEM OVERVIEW

There are variety of speech recognition [11][12] approaches available such as Neural Networks, Hidden Markov Models, Bayesian networks and Dynamic Time Warping etc. Among these approaches Neural Networks (NNs) [13] have proven to be a powerful tool for solving problems of prediction, classification and pattern recognition. Rather than being used in general-purpose speech recognition applications it can handle low quality, noisy data and speaker independence applications. Such systems can achieve greater accuracy than HMM based systems, as long as there is training data and the vocabulary is limited.

One of the most commonly used networks based on supervised learning algorithm is multilayer feed forward network which is implemented in this paper for classifying and recognizing Tamil spoken words [14][15]. In Tamil language, the pronunciation of independent letters and group of letters forming words are not different. Tamil speech recognizing system [15] does not require the support of a dictionary. Thus the recognizing process in Tamil speech [15] is fairly simple compared.

To implement the system, initially the speech data is preprocessed using filtering, framing [16] and windowing techniques. Subsequent to that, the enhanced signal is given as the input to the LPCC algorithm to extract features. These feature vectors also called cepstral coefficients are given as the input to the network. After that the network is trained with these input vectors and the target vectors. Finally classification and recognition is done based on pattern matching. The above figure 1 demonstrates the overall structure of the system.

4.1 FEED FORWARD NEURAL NETWORK FOR TAMIL WORD RECOGNITION

A feed forward neural network [15] is a biologically inspired classification algorithm which falls under the category, "Networks for Classification and Prediction" and has widespread interesting applications and functions related to speech processing. It consists of a (possibly large) number of simple neuron-like processing units, organized in layers. Every unit in a layer is connected with all the units in the previous layer. These connections are not all equal and may have

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

a different strength or weight. The weights on these connections encode the knowledge of a network. They usually consist of three to four layers in which the neurons are logically arranged. The first and last layers are the input and output layers respectively and there are usually one or more hidden layers in between the other layers. The term feed-forward means that the information is only allowed to "travel" in one direction. This means that the output of one layer becomes the input of the next layer, and so forward. Feed-forward networks are advantageous as they have the fastest models to execute.

In the Network, the input layer does have a part in calculation, rather than just receiving the inputs. The raw data is computed, and activation functions are applied in all layers. This process occurs until the data reaches the output neurons, where it is classified into a certain category. The operation of this network can be divided into two phases:

i. The learning phase
ii. The classification phase

During the learning phase the weights in the FFNet will be modified. All weights are modified in such a way that when a pattern is presented, the output unit with the correct category, hopefully, will have the largest output value. In the classification phase the weights of the network are fixed. A pattern, presented as the input will be transformed from layer to layer until it reaches the output layer. Now classification can occur by selecting the category associated with the output unit that has the largest output value. In contrast to the learning phase classification is very fast. [40]

V. TRAINING

Acoustic models for state-of-the-art speech recognition systems are typically trained on several hundred hours of task specific training data, but in low resource scenarios, one often has to make do with much less training data. Annotated training data can be especially hard to come by. In these settings, it is possible to take advantage of transcribed data from other languages to build multilingual acoustic models [20,21]. Multilingual training with Subspace Gaussian Mix-ture Models [19] have also been proposed to train acoustic models [22, 23].

An alternative approach moves the focus to data-driven feature front-ends. The key element in this data-driven approach is a multi-layer perceptron (MLP) trained on large amounts of task independent data, i.e. multilingual data or data from the same language but collected under different settings [25, 26]. Features corresponding to limited task specific data are then derived using the trained MLP for ASR [37, 28, 29, 30, 31]. We build on this front-end-based approach since features produced using these front-ends can further improve performance in low-resource settings when combined with other ASR modeling techniques.

While this work is related to several recent approaches [27, 28, 29, 30, 31], we use two different techniques to derive better features and improve acoustic model training in low resource settings: data driven features extracted using deep neural networks (DNN)

5.1 SEMI-SUPERVISED TRAINING

Semi-supervised training has been effectively used to train acoustic models in several languages and conditions [32, 33, 33, 35, 36]. This section discusses the application of these approaches to low-resource settings. We start by using a baseline decoder (the best front-end and acoustic model we have so far) to generate recognition hypotheses for any available untranscribed training data. The most reliable of these estimated transcriptions are then combined with the limited existing ranscribed training data to train both of the DNN front-end and GMM-HMM acoustic models in a semi-supervised fashion.

5.1.1. Selective semi-supervised training of DNNs

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

The initial multilingual DNN training experiments described earlier were based on only 1 hour of transcribed data. For semi-supervised training of DNNs we include additional data with noisy transcripts. These utterances are selected from the untranscribed data based on their utterance level confidences. To avoid detrimental effects from noisy semi-supervised data during discriminative training of neural networks, we make the following design choices

(a) During back-propagation training, the semi-supervised data is de-weighted. This is done by multiplying the cross-entropy er-ror with a small multiplicative factor during training.
(b) The semi-supervised data is used only in the final pre-training stage after all the layers of the DNN have been created.
(c) Only a limited amount of selected semi-supervised data is added.

5.1.2. Semi-supervised training of acoustic models
Features from the DNN front-end with semi-supervised data are used to extract data-driven features for semi-supervised training of the ASR system. Similar to the weighing of semi-supervised data dur-ing the DNN training, we also use a simple corpus weighing while training the ASR systems. This is done by adding the 1 hour of fully supervised data with accurate transcripts twice.

VI. CONCLUSION

In recent years, neural network has become an enhanced technique for tackling complex problems and tedious tasks such as speech recognition. Speech is a natural and simple communication method for human beings. However, it is an extremely complex and difficult job to make a computer respond to spoken commands. Recently there is a momentous need for ASR system to be developed in Tamil and other Indian languages. In this paper such an important effort is carried out for recognizing Tamil spoken words. To accomplish this task, feature extraction is done after employing required preprocessing techniques. The most widely used LPCC method is used to extract the significant feature vectors from the enhanced speech signal and they are given as the input to the feed forward neural network. The adopted network is trained with these input and target vectors. Semi-supervised training is used for training both neural network front-ends as well as acoustic models.

The results with the specified parameters were found to be satisfactory considering less number of training data. More number of isolated and continuous words to be trained and tested with this network in future. This preliminary experiment will helps to develop ASR system for Tamil language using different approaches like Hidden Markov Models or with other hybrid techniques.

REFERENCES

[1] A. Nayeemulla Khan and B.Yegnanarayana, Development of Speech Recognition System for Tamil for Small Restricted Task, Proceedings of National Conference on Communication, India, 2001.
[2] M. Plauche, N. Udhyakummar, C. Wooters, J.Pal, and D. Ramachadran, Speech Recognition for Illiterate Access to Information and Technology, Proceedings of First International Conference on ICT and Development, 2006.
[3] S. Saraswathi and T. V. Geetha, Implementation of Tamil Speech Recognition System Using Neural Networks, Lecture Notes in Computer Science, Vol. 3285, 2004.
[4] C. S. Kumar and Foo Say Wei, A Bilingual Speech Recognition System for English and Tamil, Proceedings of Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia, vol. 3, 2003, pp. 1641 – 1644.
[5] Frederick Jelinek, Statistical Methods for Speech Recognition, MIT Press, (ISBN 0262100665), 1997
[6] Mari Ostendorf, Vassilios V. Digalakis, and Owen A. Kimball, From HMM To Segment Models: A Unified View of Stochastic Modeling for Speech Recognition, IEEE Transactions on Speech and Audio Processing, Volume 4, No. 5, 1996, pp 360-378.
[7] Xuedong Huang, Alex Acero and Hsiao-Wuen Hon, Spoken Language Processing – A Guide to Theory, Algorithm and System Development, Prentice Hall PTR (ISBN 0-13-022616-5),2001
[8] Daniel Jurafsky, James H. Martin, Speech and Language Processing, Pearson Education, (ISBN 8178085941), 2002

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

[21] Shrikant Narayanan, Dani Byrd and Abigail Kaun, Geometry, Kinematics and Acoustics of Tamil Liquid Consonants, Journal of Acoustical Society of America, 106(4), Pt. 1, 1999, pp 1993-2007.
[9] Harold F. Schiffman, A Reference Grammar of Spoken Tamil, Cambridge University Press (ISBN-10: 0521027527), 2006
[10] Balasubramanian T, Timing in Tamil, Journal of Phonetics, Volume 8, 1980, pp 449 – 467.
[11] Amin Ashouri Saheli, Gholam Ali Abdali, Amir Abolfazl suratgar,(2009) "Speech Recognition from PSD using Neural Network", Proceedings of the International MultiConference of Engineers and Computer Scientists 2009, Vol I,IMECS 2009, March 18 - 20, 2009, Hong Kong.
[12] S K Hasnain, Azam Beg(2008), "A Speech Recognition System for Urdu Language", in International Multi-Topic Conference (IMTIC'08), Jamshoro, Pakistan, 2008, pp. 74-78.
[13] Abdul Manan Ahmad', Saliza Ismail+, Den Fairol SamaonL(2004) "Recurrent Neural Network with Backpropagation through Time for Speech Recognition', Intemational Symposium on Communications and Information Technologies 2004 (ISClT2004), Sapporo, Japan, October 26- 29.
[14] Muthanantha murugavel, (2007) "Speech Recognition Model for Tamil Stops, Proceedings of the World Congress on Engineering 2007 Vol I, WCE 2007, July 2 - 4, 2007, London, U.K.
[15] N.Uma Maheswari, A.P.Kabilan, R.Venkatesh, (2010) "A Hybrid model of Neural Network Approach for Speaker independent Word Recognition", International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010,1793-8201.
[16] S K Hasnain, (2008) "Recognizing Spoken Urdu Numbers Using Fourier Descriptor and Neural Networks with Matlab", Second International Conference on Electrical Engineering,25-26 March 2008,University of Engineering and Technology, Lahore (Pakistan) (2008)
[16] H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, and C.H. Lee, "A study on multilingual acoustic modeling for large vocabulary ASR," in IEEE ICASSP, 2009.
[17] D. Imseng, J. Dines, P. Motlicek, P.N. Garner, and H. Bourlard, "Comparing different acoustic modeling techniques for multilingual boosting," in ISCA Interspeech, 2012.
[18] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafi´at, D. Povey, et al., "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in IEEE ICASSP, 2010.
[19] L. Lu, A. Ghoshal, and S. Renals, "Regularized subspace Gaussian mixture models for cross-lingual speech recognition," in IEEE ASRU, 2011.
[23] Y. Qian, D. Povey, and J. Liu, "State-level data borrowing for low-resource speech recognition based on subspace GMMs," in ISCA Interspeech, 2011.
[24] S. Sivadas and H. Hermansky, "On use of task independent training data in tandem feature extraction," in IEEE ICASSP, 2004.
[25] A. Stolcke, F. Gr´ezl, M.Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in IEEE ICASSP, 2006.
[26] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and multistream posterior features for low resource LVCSR systems," in ISCA Interspeech, 2010.
[27] Y. Qian and J. Liu, "Cross-lingual and ensemble MLPs - Strategies for low-resource speech recognition," in ISCA Interspeech, 2012.
[28] N. Thang, B. Wojtek, F. Metze, and T. Schultz, "Initialization schemes for multilayer perceptron training and their impact on ASR performance using multilingual data," in ISCA Interspeech, 2012.
[29] S. Thomas, S. Ganapathy, and H. Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in IEEE ICASSP, 2012.
[30] S. Thomas, S. Ganapathy, A. Jansen, and H. Hermansky, "Data-driven posterior features for low resource speech recognition applications," in ISCA Interspeech, 2012.
[31] S K Hasnain, (2008) "Recognizing Spoken Urdu Numbers Using Fourier Descriptor and Neural Networks with Matlab", Second International Conference on Electrical Engineering,25-26 March 2008,University of Engineering and Technology, Lahore (Pakistan) (2008)
[32] G. Zavaliagkos, M. Siu, T. Colthurst, and J. Billa, "Using untranscribed training data to improve performance," in ISCA ICSLP, 1998.
[32] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in ISCA Eurospeech, 1999.
[34] L. Lamel, J.L. Gauvain, and G. Adda, "Unsupervised acoustic model training," in IEEE ICASSP, 2002.
[35] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised training on large amounts of broadcast news data," in IEEE ICASSP, 2006.
[36] S. Novotney and R. Schwartz, "Analysis of low-resource acoustic model self-training," in ISCA Interspeech, 2009.
[37] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," in ISCA Eurospeech, 1997.
[38] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky, and J. Cernocky, "Combination of strongly and weakly constrained recognizers for reliable detection of OOVs," in IEEE ICASSP, 2008.
[39] A.M. Natarajan, R. Thangarajan, "Word And Triphone Based Approaches In Continuous Speech Recognition For Tamil Language" WSEAS Transactions On Signal Processing, Issn: 1790-5022 Issue 3, Volume 4, March 2008
[40] Computer Science & Engineering: An International Journal (Cseij), Vol.1, No.4, October 2011 DOI : 10.5121/CSEIJ.2011.1401 1 " Isolated Word Recognition System For Tamil Spoken Language Using Back Propagation Neural Network Based On LPCC Features" Dr.V.Radha, Vimala.C, M.Krishnaveni
[41] Samuel Thomas, Michael L. Seltzer, Kenneth Church And Hynek Hermansky "Deep Neural Network Features And Semi-Supervised Training For Low Resource Speech Recognition"