Multilingual Speech to Speech Translation for Telugu to English and Hindi to English Speech

Smt. P. Prithvi
Assistant Professor
Department of Electronics and Communication Engineering
National Institute of Technology, Warangal-506004, Telangana

Dr. T. Kishore Kumar
Associate Professor
Department of Electronics and Communication Engineering
National Institute of Technology, Warangal-506004, Telangana

Abstract

Speech processing is emerged as an essential task in the context of modern communication system. The concept of speech translation deals with the speech signals in a source language A to the target language B. In my work mainly speech to speech translation aiming our local language Telugu to English translation and our national language Hindi to English. In this process, first extraction of features and then reducing noise which can further be used to transfer into text form is done. In this paper, deep learning based modelling technique is employed for speech recognition. After conversion of the text, the data is compared with dictionary data as per the transcriptions for language identification. Mapping is used to generate the signals for transcription. The index value of the recognition is used for language identification. After the language is identified the phonetic approach is used for generate the corresponding text to speech signal.

Keywords: Deep Neural Networks, Phonetic approach, Speech Recognition, Mapping, Language Identification, ASR, Transcriptions

I. INTRODUCTION

Speech is the communication tool between people all around the world. For several decades attempts to build systems for recognizing speech signals and to artificially synthesize speech signals have been made. Speech systems find wide range of applications in medicine, marketing, education, military and cross border operations etc. Language of speech is an important factor to be dealt with in order to complete an effective communication link. Over the past few decades, the need to overcome this language barrier between people belonging to different linguistic backgrounds has been an area of interest in research. Language translation systems have served as a major breakthrough for this issue. As an expansion to this milestone, a speech-to speech translation system (S2ST) is an attempt being made over several institutions across the globe. On an overview, any S2ST system consists of three modules namely, speech recognition, machine translation and speech synthesis. Translation becomes a challenging task when the languages involved have completely different linguistic structures.

Sneha Tripathi et.al in [1] compared the various approaches that can be used for machine translation. Phillip Koehn in [2] gives a detailed description on statistical machine translation. Alon Lavie in [3] describes the various techniques which can be used to evaluate the translation system. A number of significant research works are conducted in the field of S2ST. Sakriani Sakti et.al in [4] came up with the first ever network based speech-to speech translation system for Asian languages, which was built by successfully combining the three subsystems using rule-based translation and networking. V. V. Babu formulated a system named ANUV AADHAK which is a two-way Indian language Speech-to-Speech Translation System for local travel information assistance. It was noticed that most of the speech-to-speech translation systems use a rule-based language processing approach by incorporating Natural Language Processing (NLP).

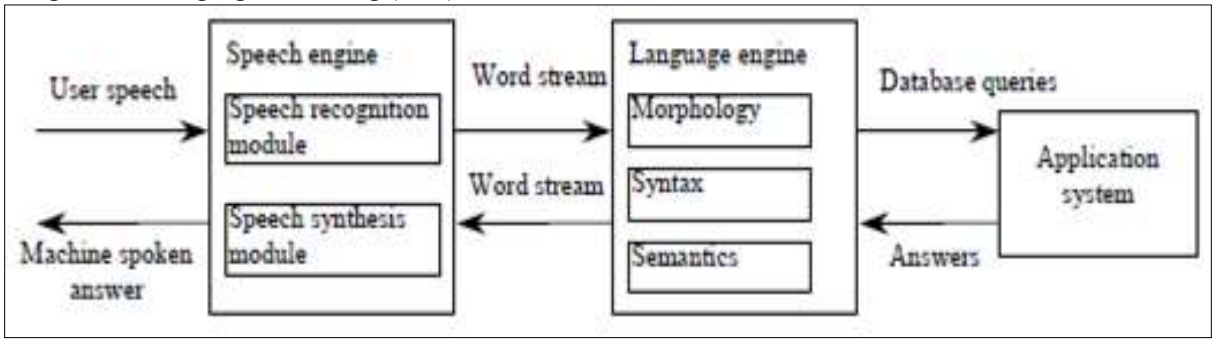

Fig. 1: Proposed Speech to speech translation system

The proposed work is an S2ST system for general expressions which can perform translation between Hindi to English and Telugu to English which is based on statistical machine translation approach.

A. Speech Recognition Module

Speech recognition systems hold the capability to identify words in a specific language and convert them into text form which can be presented in a machine readable format. It typically uses algorithm through acoustic and language modelling. Acoustic modelling accomplishes the task of bringing out a relationship between linguistic units of speech and the corresponding audio signal. Further, language modelling takes the responsibility of matching sounds with word sequence. This eventually helps to differentiate words with similar sounding.

The phenomenon of automatic speech recognition belongs to the class of machine learning principles. It refers to application of computational techniques especially in acoustic speech signals into words and further into machine readable format. There are many real-time applications of this automatic speech recognition system (ASRS) associated with latest wireless communication. Human and mobile device interaction often involves in retrieving data for automated call and other control purpose.

The ASRS system plays a vital role where communication is significantly required in natural language. Systems with such facility often were used as useful gadgets in prosthetic electronics to serve blind people for daily activities.

The efficiency of the system for the purpose of recognition is directly dependent on the performance of the algorithm employed. Statistical methods like Markov source or the hidden Markov model with versatile mathematical structure has the ability to model any form of signal system which is produced as a real time process in the Mother Nature. This capability rapidly fascinated the researchers to employ in speech recognition applications. The imperative feature of the HMM is that it is capable of characterising signals which are in discrete and continuous form. The HMM encompasses in evaluation of probability of a sequence, then determination of best sequence followed by the step of adjusting the model parameters.

A pole zero vocal tract model proposed typically involves in minimizing the log likelihood, ratio cost. Natural language consists of speech features and lexical words which are correlated to each other. As a result the correlation factor can be employed for classification. Hence, it is considered that model based on any one of these canner sufficiently provide performance which is possible in the case where both of them are utilized. This technique is referred as joint acoustic and linguistic modelling [4]. This is based on the maximum entropy principle.

In its simple form the speech recognition can be referred as the process of search for a word sequence which has higher likelihood with the sequence from test speech. However, it is concluded in various literatures that are the maximum likelihood canner ensures better classification.

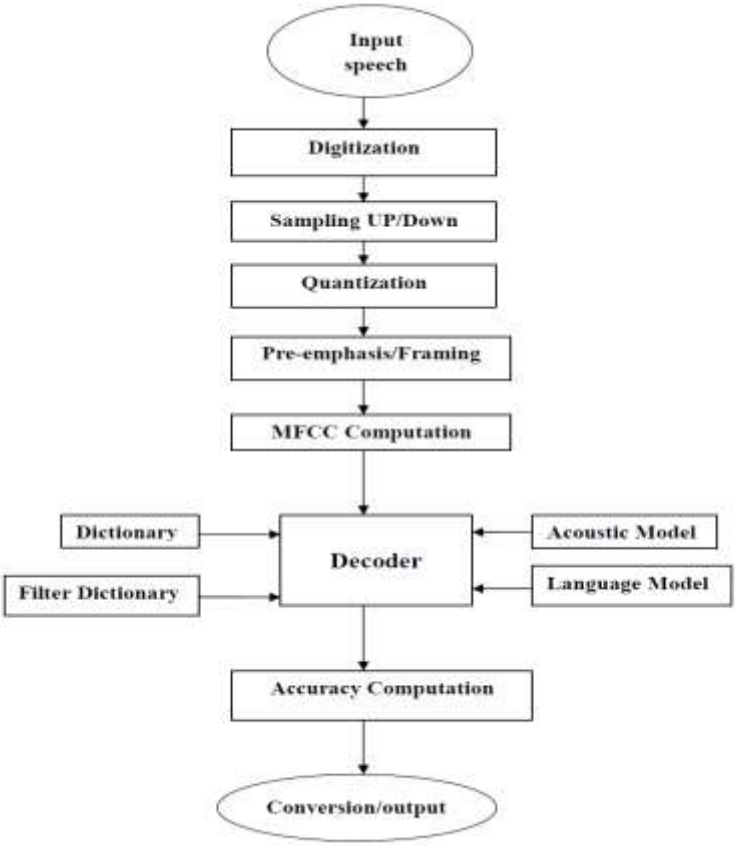

Fig. 1: Proposed Block Diagram of speech recognition

B. DNN

Deep Neural Networks (DNNs) are considered as a significant tool in machine learning. The tool can be applied to model more effectively. However, the literature states that the process of training the data in DNN is extremely time-consuming. The latest

computing techniques are not even at par to accelerate. DNN training is especially slow for tasks with large data sets. Existing approaches for speeding up the process involve parallelizing the Stochastic Gradient Descent (SGD) algorithm used to train DNNs. Even under such conditions it is possible to determine whether DNNs would still be able to produce state-of-the-art results using this low-precision arithmetic. To answer this question, we implement an approximate DNN that uses the low-precision arithmetic and evaluate it on the TIMIT phoneme recognition task and the Wall street Journal (WSJ) Based speech recognition task. For both tasks, we find that acoustic models based on approximate DNNs perform as well as ones based on conventional DNNs; both produce similar recognition error rates. The approximate DNN is able to match the conventional DNN only if it uses Kahan summations to preserve precision. These results show that DNNs can run on low-precision hardware without the arithmetic causing any loss in recognition ability. The low-precision hardware is therefore a suitable approach for speeding up DNN training.

Machine learning is being used to solve problems that are increasingly important in day-to-day life. Machine learning techniques have been applied in fields such as bioinformatics, computer vision, economics, fraud detection, robotics, and speech. Recently, one of the most successful machine learning techniques has been Deep Neural Networks (DNNs). For instance, DNNs have currently been used extensively in the speech recognition community. For Automatic Speech Recognition (ASR), DNN-based models result should relative improve in word error rates over traditional methods. In a speech task that uses 81 hours of speech as training data, the DNN takes more than half a day to train. It is general knowledge that speech recognizers benefit greatly from larger amounts of data; some systems use training sets with thousands of hours of speech. At the present rate, those DNNs would take weeks to train. Worse yet, selecting the best DNN architecture and tuning a DNN's hyper-parameters typically require training many DNNs. The slow training is a significant impediment to DNN researchers.

A DNN is a feed-forward neural network which has more than one hidden nonlinear layer. For an input vector Xt, each hidden layer transforms its input vector from the layer below to the layer above by applying an affine transform and nonlinear mapping as follows:

$$z^0 = x_t$$

$$y_i^{(l+1)} = \Sigma_{j=1}^{N^{(l)}} w_{ij}^{(l)} z_j^{(l)} + b_i^{(l)}$$

$$z_i^{(l+1)} = \sigma(y_i^{(l+1)}),$$

Different activation functions which can be employed are listed below as

$$\sigma(x) = \begin{cases} sigmoid(x) = \frac{1}{1+exp(-x)} \\ tanh(x) = \frac{1-exp(-2x)}{1+exp(-2x)} \\ ReLU(x) = max(0,x) \\ softplus(x) = log(1+e^x). \end{cases}$$

DNNs with many hidden layers are hard to optimize. Gradient descent from a random starting point near the origin is not the best way to find a good set of weights, and unless the initial scales of the weights are carefully chosen, the back propagated gradients will have very different magnitudes in different layers. In addition to the optimization issues, DNNs may generalize poorly to held-out test data. DNNs with many hidden layers and many units per layer are very flexible models with a very large number of parameters. This makes them capable of modelling very complex and highly nonlinear relationships between inputs and outputs. This ability is important for high quality acoustic modelling, but it also allows them to model spurious regularities that are an accidental property of the particular examples in the training set, which can lead to severe over fitting. Weight penalties or early stopping can reduce the over fitting but only by removing much of the modelling power. Very large training sets can reduce over fitting while preserving modelling power, but only by making training very computationally expensive. What we need is a better method of using the information in the training set to build multiple layers of nonlinear feature detectors.

C. Language Identification

Language identification is the task of automatically detecting the language(s) present in a document based on the content of the document. Language identification techniques commonly assume that every document is written in one of a closed set of known languages for which there is training data, and is thus formulated as the task of selecting the most likely language from the set of training languages. In this work, language identification in documents that may contain text for more than one language from the candidate set. This work used index based mapping method that concurrently detects that a document is multilingual, and estimates the proportion of the document that is written in each language.

D. Synthesis System: Phonetic Analysis – Generate Text Sound

The flowchart of phoneme based text to speech synthesis for words is shown in figure.3.In this part, the input text is considered only syllable word to produce speech as natural. Firstly, the input word is given from language identification. In the next step, it is

necessary to convert from word to phonetic transcription which is also called grapheme to phoneme conversion. Dictionary based approach, more exact than rule based approach, is applied in this step. Then, phoneme sounds are concatenated by depending on the phonetic transcriptions of word to produce speech.

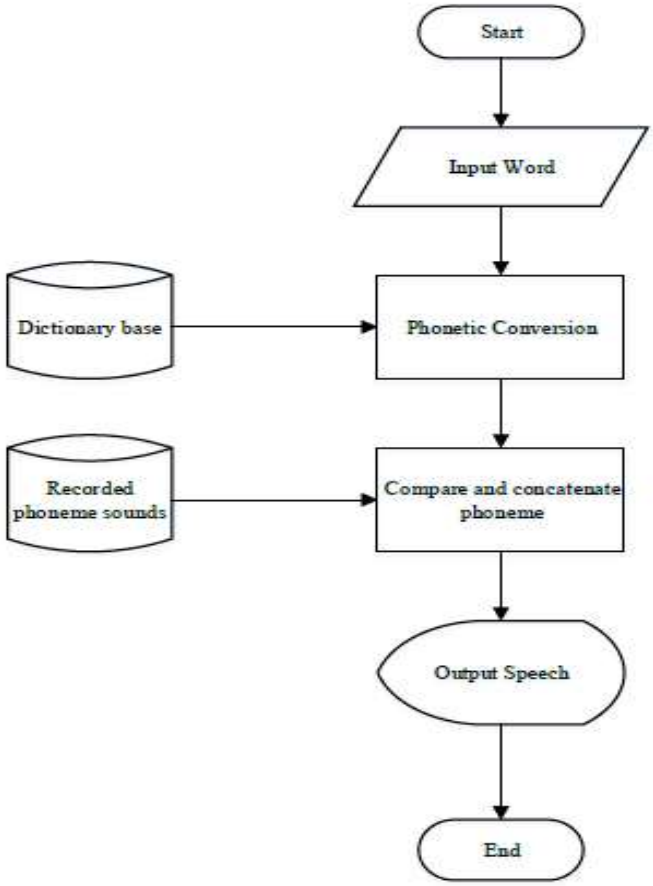

Fig. 3: flow chart of phoneme based text to speech approach

II. EXPERIMENT AND IMPLEMENTATION

For test data sequence, the available open source namely Sphinx III and pocket sphinx speech decoders are utilized. Their respective acoustic models are considered separately. The initial step involves in extracting the features of the speech signal. This feature vector efficiently takes a significant role in building an acoustic model. The procedure followed to speech recognition is explained in the flow chart.

The initial process test speech file is which is in digitalized form is considered for MFCC extraction. These features are fed to the decoder where the corresponding dictionary, filter dictionary, acoustic models and Language models for Telugu are available. This step is followed by process of computing recognition accuracy, which involves in calculating word error rate (WER).

The calculation of WER is as given below

$$WER = \frac{100(S + D + I)}{N}$$

Where N - total number of words in the test
S - Substitution
D - Deletion
I - Insertion

The stored speech audio and video file is given to ffmpeg tool. The ffmpeg is a cross platform solution to record, convert and stream Audio and Video. The ffmpeg reads the arbitrary number of inputs. The data entered is in the form of text. The corresponding output is saved into the folder as .wav file. Consecutively sphinx_fe.exe extracts the MFCC features from the input file. The input file is, output of ffmpeg tool. During the process of decoding various parameters are supplied to the frame work. The parameters are preserved in the form of text file.

The params.txt file has the information about the paths of file to be converted and model parameters and model architecture and the converted file storage location. Sphinx3_decode is an executable which decodes the speech file according the information in

params.txt. Further the extracted MFCC featured data are decoded with respect to the Acoustic model and language model. The output of the decoder is speech files converted into .text file.

III. SIMULATION RESULTS

The accuracy of the two models is computed to test the performance of the technique. Training the speech data of 800 sentences consists of 4064 unique words as a training data and the data with 30 sentences consists of total 702 words. The corresponding recognition accuracy using both Sphinx-3 and pocket Sphinx are evaluated. Automated speech recognition (ASR) accuracy is about 87.32% with Sphinx-3 model. While, it is about 92.59 % with pocket sphinx model. Total 702 words are used for this simulation based experiment.

The training and testing of timid database Plays a major role in the experimentation with different parameters. The corresponding parameters and their respective magnitudes are tested in Table I.

Table – 1
Speech Recognition system Accuracy

Total testing words	*ASR accuracy with Sphinx-3*	*ASR accuracy with Pocket Sphinx*
702	*87.32 %*	*92.59 %*

Table - 2
Training and testing accuracy of our technique with timid database

Parameter	**8-Times**	**12-Times**
Total Words	14550	14550
Correct Recognized Words	14107	14205
Errors	512	411
Total Percent Correct	96.96 %	97.63 %
Error Percentile	3.52 %	2.824 %
Accuracy	96.28 %	97.17 %

The identification results can be observed based on the verification of the total no of speakers. Representation of codebooks and its distribution is as shown in fig. 4.

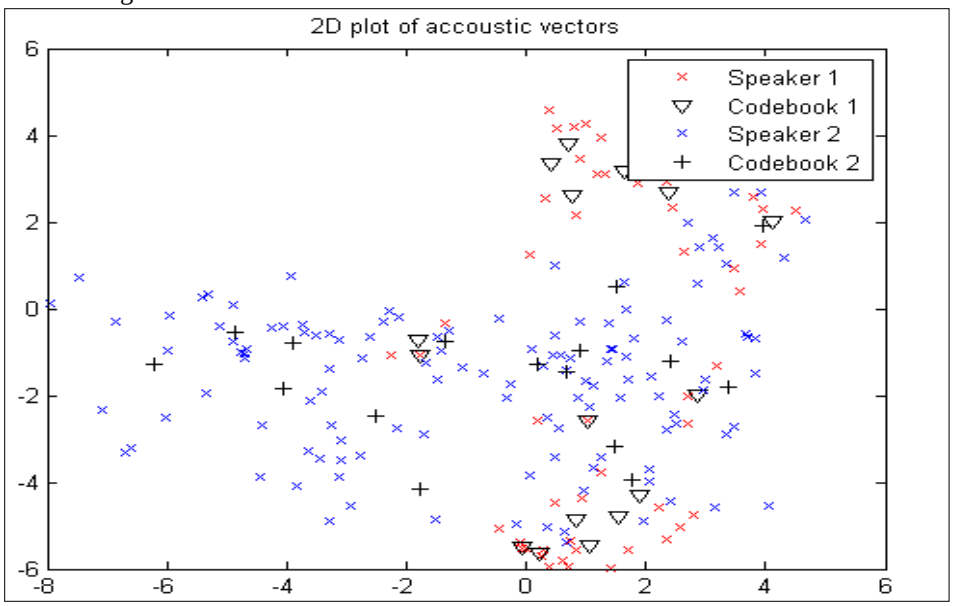

Fig. 4: Representation of codebooks of the speakers

IV. CONCLUSION

Novel DNN is successfully employed in modelling the extracted features of the given data set which is further used as training set for speech recognition. The test data features are efficiently compared with the training. Recognition and further translation into the data set defined language is successfully performed. The performance is evaluated using the ASR accuracy computation and further analysed for language identification and speech translation. The language identification generates each index value for the corresponding speech signal, when is compared with the database signals using indexed based mapping we can get the corresponding language. After language identified the successive phonetic sound is generated for the text at the speech synthesis. Finally we get the corresponding output language as the English for any given input speech as Hindi or Telugu.

REFERENCES

[1] Sneha Tripathi, Juran Krishna sakhel, "Approaches to Machine Translation" in: Annals of Library and Information in Studies, vol 57, December 2010, page 388-393.
[2] Philipp Koehn, Ondrej Bojar, Rajen Chatterjee, Federmann, Jimeno Yepes, "Findings of the 2016 Conference on Machine Translation", (WMT16) Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 131–198, Berlin, Germany, August 11-12, 2016. C 2016 Association for Computational Linguistics.
[3] Alon Lavie, Austin Matthews, Waleed Amma,r Archna Bhatia, Weston Feely, Greg Hanneman, Eva Schlinger, Swabha Swayamdipta, Yulia Tsvetkov, Chris Dyer, "The CMU Machine Translation Systems" at WMT 2014.
[4] Sakriani Sakti, Thang Tat Vu, Andrew Finch, Michael Paul, Ranniery Maia, Shinsuke Sakai, Teruaki Hayashi, Shigeki Matsuda, Noriyuki Kimura, Yutaka Ashikari, Eiichiro Sumita, Satoshi Nakamura, " NICT/ATR ","Asian Spoken Language Translation System for Multi-Party Travel Conversation" NICT Spoken Language Communication Research Group ∗ 2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan.
[5] G. Dahl, D. Yu, L. Deng, and A. Acero, "Large vocabulary continuous speech recognition with context-dependent DBN-HMMs," ICASSP, 2011.
[6] L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," Interspeech, 2010.
[7] J. Baker, L. Deng, J. Glass, S. Khudanpur, Chin hui Lee, N. Morgan, and D. O'Shaughnessy, "Developments and directions in speech recognition and understanding, part 1," Signal Processing Magazine, IEEE, vol. 26, no. 3, pp. 75–80, may 2009.
[8] Hinton, G., Osindero, S., and Teh, Y. "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, pp. 1527-1554, 2006.
[9] S.Young, "Large Vocabulary Continuous Speech Recognition: A Review," IEEE Signal Processing Magazine, vol. 13, no. 5, pp. 45–57, 1996.
[10] Mohamed, A., Dahl, G., and Hinton, G. "Deep belief networks for phone recognition," in Proc. NIPS Workshop Deep Learning for Speech Recognition and Related Applications, 2009.
[11] Bengio Y. "Learning deep architectures for AI," in Foundations and Trends in Machine Learning, Vol. 2, No. 1, 2009, pp. 1-127.
[12] Marc'Aurelio Ranzato, Christopher Poultney, Sumit Chopra and Yann LeCunEfficient Learning of Sparse Representations with an Energy-Based Model, in J. Platt et al. (Eds), Advances in Neural Information Processing Systems (NIPS 2006), MIT Press, 2007.
[13] Yoshua Bengio, Pascal Lamblin, Dan Popovici and Hugo Larochelle,Greedy Layer-Wise Training of Deep Networks, in J. Platt et al. (Eds), Advances in Neural Information Processing Systems 19 (NIPS 2006), pp. 153-160, MIT Press, 2007.
[14] Li Deng, "A Tutorial Survey of Architectures, Algorithms, and Applications for Deep Learning" to appear in APSIPA Transactions on Signal and Information Processing, Cambridge University Press, 2014.
[15] Bengio Y, "Deep learning of representations: looking forward," in: Statistical Language and Speech Processing, pp. 1--37, Springer, 2013.
[16] Bengio Y., Courville, A., and Vincent, P. "Representation learning: A review and new perspectives," IEEE Trans. PAMI, 2013a.
[17] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," IEEE Trans. Audio, Speech, Lang. Proc., vol. 20, pp. 30–42, 2012.