ANALYSIS OF FMLLR PERFORMANCE AT VARIOUS NOISES AND LEVELS FOR KANNADA

[1] Vishwas K Singh, [2] Dr. G. F. Ali Ahammed
[1]Masters Student, [2]Associate Professor and Course Coordinator
[1]Department of Digital Electronics and Communication Systems,
[1]VTU PG Center, Mysuru, India

Abstract : ***Automatic speech recognition systems are very useful in human to machine interface. A FMLLR technique is applied to clean and noisy conditions data in Kannada with connected digit utterances. This method can be applied to any network topology that delivers log-likelihood-like scores. This can be useful when it is not easy to train a DNN - HMM system in conditions that are similar to the testing ones. A connected digit utterances in Kannada speech database is created and noisy data is also created using artificial white noise, babble noise, traffic noise, train noise and factory noise. Analysis of performance of the system is made and it is found that system excelled for clean noise and an average WER of 30 percent for all the noises.***

IndexTerms - FMLLR, ASR, Kaldi, Kannada.

I. INTRODUCTION

Humans have from years interacted with machines with many kinds of interfaces such as text inputs through keyboard, gestures, images, voice etc. Though voice interfaces are being deployed and used from the past decade, it is still in an evolutionary stage and new. Research is performed in such area where the machine is made to understand humans through voice. Understanding the user by his voice presents with us a problem of recognizing words produced by him.

The main focus is on the speech recognition using telecommunication equipment such as IVR systems. But not limited to it as these are also hand held devices and can be used by many people. The acoustic modelling tells us about the features of speech sounds. Language modelling is a statistical approach where sounds of language are represented as output from an information source based on some discrete probability distribution. Sounds are produced in a certain manner or pattern which when interpreted by another entity is understood as the same. This is common for a language such as English, Kannada, Hindi, Telugu and Malayalam etc. Kannada is a native language of Karnataka state spoken by almost 6Crore people. A Feature Space Maximum Likelihood Regression technique for speech recognition is used in the project performing recognition in noisy environment.

II. LITERATURE SURVEY

The papers [1],[2],[3],[4] presented Isolated Digits Recognition in Kannada, it presents an ASR system based on HMM. It demonstrates an approach of pattern recognition. [5],[6] show the use of a binary support vector machine in isolated word recognition which also provides better results. A detailed study of HMM for speech recognition tasks and parameters to be considered for such a model was presented in [7]. A Neural Network approach for solving the problem of Speech Recognition was presented in [10] and suggested that a hybrid approach of HMM - ANN based system was more powerful. The speech processing in [11] showed the acoustic and phonetic basics needed to representation and modelling of systems and the domain of automatic speech recognition. In a more detailed approach was presented for the problem of speech recognition, various methods for addressing the issue. The comparison between the statistical Gaussian Mixture Model and Hidden Markov Model approaches was shown in [11] and suggested HMM approach over GMM. The Mel Frequency Cepstrum Coefficients (MFCCs) and Linear Predictive Coding (LPC) model parameters were used to model the speech signal in a digital domain approach. One method is of a DNN- HMM system in a multi noise/ multi condition environment which reduces the gap between the clean speech training and multi noise training, when an ideal environment is unavailable.

Literature proposed many methods for dealing with Robust speech recognition such as modified HMM and GMM techniques such as FMLLR, Uncertainty weighting and propagation, support Vector Machines etc, as the research is much more mature than for native languages. We also found that FMLLR would be more computationally efficient for smaller task of recognizing digits in noisy environments. This would be more suitable for low complexity systems and hardware.

III. FMLLR

Feature space Maximum Likelihood Linear Regression is a modified technique of DNN – HMM system. DNNs can be modelled as cascade systems of affine transforms (transforms between affine spaces which preserve the characteristics of the points, lines etc..) and non linear activation functions.

The activation functions are a sigmoid for hidden layers and soft-max for output layer. The output at the first hidden layer can be written as

□□□□□□□ϕ□□□□□X□□□□□□ □□□

where X is the input, A and b are bias for the first layer. Let A_i, b_i = 1: N be the DNN parameters. The parameters are speaker specific and can be re scaled and re trained to obtain more efficient DNNs. The speaker information may be transferred by passing them as i-vectors through DNNs and FMLLR transformed features may be obtained. These can be adapted as a bias in the first layer of DNN.

□□□□□□□□□□□□ϕ□□□□X□□□□□□□□□□□□□□□□ϕ□□□□X□□□□□□□□□□□□□ □□□

where b_1^{spkr} is the speaker side information. T is an affine transform learnt using back-propagation.

We can say that a feature transformation is an inverse of model transformation. Here feature space is transformed by normalization of feature vectors. Advantages of the FMLLR can be as follows:

1. Feature estimations are quick and require only a few iterations.

2. A few minutes of audio clip is sufficient for robust parameter estimation in DNN
3. MLLRs can compensate for mismatch in audio.
4. They are less sensitive to transcription errors.

IV. IMPLEMENTATION

4.1 Database preparation

Database consists of utterances of Kannada digits spoken by 7 native Kannada speakers(6-train,1-test). Each speaker has spoken 10 of 3 digit connected sentences. Hence a total of 210 digits utterances are recorded in a controlled environment with very low background noise. The recording is done through the microphone built within the laptop. Audacity tool is used for recording the sentences. The sampling rate used for recording the speech is 44.1kHz. Data is segregated into the train and test folders. Note that here the train data does not contain the test speaker data. Since our main application is concentrated on the telecommunication field, it was needed that the speech be down sampled to 8kHz, the data was filtered using an anti aliasing filter and down sampled. For noisy speech data, noise signals was then added to the clean speech data at desired SNRs 3dB, 5dB, 10dB, 15dB noise corrupted database was created.
The transliterated text was generated using the online interface to ITRANS. The Table 1. Shows the transliteration and Lexicon Mappings.

Table 1. Kannada transliteration and Lexicon Mappings

Number	English Text	Transliterated text	Lexicon mappings
1	One	Ondu	o .n du
2	Two	Eradu	e ra Du
3	Three	Muru	mU ru
4	Four	Naalku	naa l ku
5	Five	Aidhu	ai du
6	Six	Aaru	aa ru
7	Seven	Elu	E Lu
8	Eight	Entu	e .n Tu
9	Nine	Ombhattu	o .n bha ttu
0	Zero	Sonne	so .n ne

4.2 System Algorithm

1. Prepare acoustic data – generate the above files
2. Feature Extraction - MFCCs coefficients
3. Prepare language Data
4. Generate Language model
5. Perform Mono Training, Decoding
6. Perform Tri training, Decoding

V. RESULT ANALYSIS

The results obtained through above experiments are compared and analyzed using Word Error Rates calculated as follows

$$WER = (S+D+I)/N \quad (3)$$

Where S = No. of Substitutions, D = No. of Deletions, I = No. of Insertions N = No. of Sentences

Database	Training Type	3dB	5dB	10dB	15dB
Clean Speech	Mono	0.0	0.0	0.0	0.0
	Tri	0.0	0.0	0.0	0.0
Babble Noise	Mono	30.00	23.33	13.33	3.33
	Tri	40.00	26.67	13.33	6.67
Factory Noise	Mono	20.00	26.67	3.33	10.00
	Tri	16.67	33.33	3.33	13.33
White Noise	Mono	6.67	6.67	3.33	6.67
	Tri	16.67	6..67	3.33	6.67
Traffic Noise	Mono	50.00	43.33	33.33	20.00
	Tri	56.67	50.00	33.33	26.67
Train Noise	Mono	56.67	56.67	13.67	0.00
	Tri	50.00	46.67	66.67	13.33

Fig.1 Summary of WERs

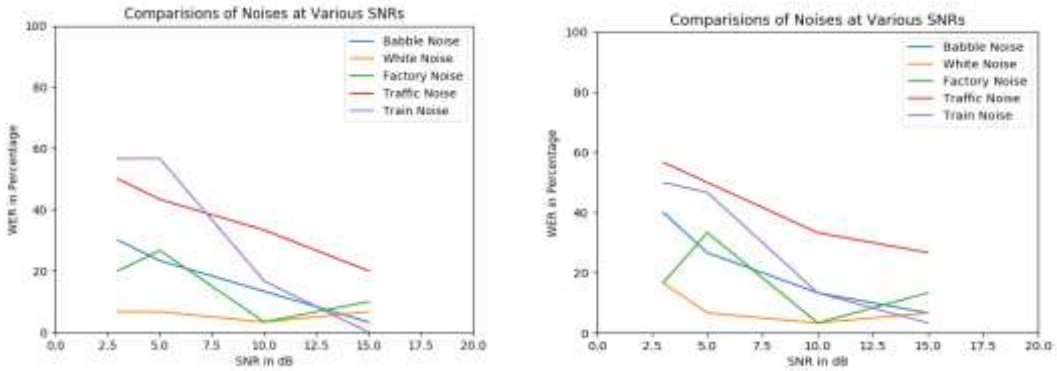

Fig.2 Comparison of effects of noises for Mono training and Tri training

VI. CONCLUSION

The creation of clean Kannada Speech database, noisy speech database with 6 different types of noises: White, Babble, Factory, Train, Traffic at 3dB, 5dB, 10dB, 15dB was performed. Analysis was done on these databases by a FMLLR front end in Kaldi ASR toolkit. Both the Mono training and Tri training was performed. Traffic noise, Train noise, Babble noise performance was almost linearly decreasing with increase in SNR. WER reduced to 0 at 15dB SNR for train noise. But the factory noise performance varied, at low SNR it was low but increased in the middle and then decreased at high SNRs. Overall performance of Mono training was found better comparative to Tri training, but a closer look revealed that the results were almost similar and WERs were near. It was found that the system performed well for most of the SNRs when the noise was White with WERs within 20 percent.

REFERENCES

[1] V Sneha, G Hardhika, K Jeeva Priya, 2018 Isolated Kannada Speech Recognition using HTK- A detailed approach, Advances in Intelligent Systems and Computing 564, 185 -194.

[2] Gurudath K P, Dr. D J Ravi, 2016 Isolated Digits Recognition in Kannada Language, International Journal of Computer Applications Vol. 140- No. 10, 23-29.

[3] Suma Swamy, K V Ramakrishnan, 2013 An Efficient Speech Recognition System, Computer Science and Engineering: An International Journal Vol. 3 No. 4, 21-27

[4] Murali Krishna H, Ananthakrishna T, Dr. Kumara shama, 2013 HMM Based Isolated Kannada Digit Recognition System using MFCC, International Conference on Advances in Computing, Communications and Informatics , 730-733

[5] M. A. Anusuya, S K Katti, 2012 Speaker Independent Kannada Speech Recognition using Vector Quantization, International Journal of Computer Applications, 32-35

[6] Sarika Hegde, Achary K K, Surendra Shetty, 2012 Isolated Word Recognition for Kannada Language using Support Vector Machine, ICIP 2012 CCIS 292,262-269

[7] Mark Gales, Steve Young, 2007The Application of Hidden Markov Models,

[8] Wendy Holmes, 2001 Speech synthesis and recognition

[9] Lawrence Rabiner, Biing-Hwang Juang, 1993 Fundamentals of speech recognition

[10] L. R. Rabiner, 1989 A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, IEEE Proceedings vol 77

[11] L. R. Rabiner, R. W. Schafer, 1978 Digital Processing of Speech