Large Vocabulary Isolated Word Recognition Using Syllable, HMM and Normal Fit

Hemakumar G., Punitha P.

Abstract— this paper addresses the problem of large vocabulary speaker dependent isolated Kannada words recognition using the syllables, Hidden Markov Model (HMM) and Normal fit method. This experiment has covered 5.5 million words among the 10 million words from Hampi text corpus. Here 3-state Baum–Welch algorithm is used for training. For the 2 successor outputted λ(A, B, pi) is combined and passed into normal fit, the outputted normal fit parameter is labelled has syllable or sub-word. Our model is compared with Gaussian Mixture Model and HMM (3-state Baum–Welch algorithm). This paper clearly shows that for normal fit applied for HMM will reduce the memory size while building the speech models and works with excellent recognition rate. The average WRR is 91.22% and average WER is 8.78%. All computations are done using mat lab.

Index Terms— ASR, Voice Detection, Speaker Dependent, Segmentation, LPC, Normal fit and Baum-Welch Algorithm.

1 Introduction

Automatic speech recognition (ASR) is the process by which a computer maps an acoustic speech signal to text. The goal of speech recognition is to develop techniques and systems that enable computers to accept speech input and translate spoken words into text and commands. The problem of speech recognition has been actively studied since 1950s and it is natural to ask why one should continue studying speech recognition. Speech recognition is the primary way for human beings to communicate. Therefore it is only natural to use speech as the primary method to input information into computational device or object needing manual input. Speech recognition is the branch of human-centric computing to make technology as user friendly as possible and to integrate it completely into human life by adapting to humans' specifications. Currently, computers force humans to adapt to computers, which is contrary to the spirit of human-centric computing. Speech recognition has the basic quality to help humans easily communicate with computers and reap maximum benefit from them. The performance of speech recognition has improved dramatically due to recent advances in speech service and computer technology with continually improving algorithms and faster computing.

The speech recognition system may be viewed as working in a four stages namely converting analog speech signal into Digitalization (Normalization part) form, Feature extraction part, Speech Model building part, and Testing. In the speech signal, feature extraction is a categorization problem about reducing the dimensionality of the input vector while maintaining the discriminating power of the signal. As we know from fundamental formation of speech recognition system, that the number of training sets and test vector needed for the classification problem grows with the dimension of the given input, so we need feature extraction techniques. In speech processing there are so many methods for feature extraction in speech signal, but still Linear-Predictive coding (LPC) coefficients and Mel-Frequency Cepstral Coefficient (MFCC) are most commonly used technique [1][5][6].

The objective of modelling technique is to generate speech models using speaker specific feature vector. The speech recognition is divided into two parts that means speaker dependent and speaker independent modes. In the speaker independent mode of the speech recognition the computer should ignore the speaker specific characteristics of the speech signal and extract the intended message. On the other hand in case of speaker dependent recognition machine should extract speaker characteristics in the acoustic signal. To developing speech models there are many techniques namely, Acoustic-Phonetic approach, Pattern Recognition approach, Template based approaches, Dynamic time warping, Knowledge based approaches, Statistical based approaches, Learning based approaches, The artificial intelligence approach, Stochastic Approach [5][6][7].

This paper discussing the large vocabulary speaker dependent isolated Kannada word recognition using Syllable, HMM and Normal fit technique and compared with HMM and GMM, for the memory size required in storing the speech model and accuracy of recognition.

The remaining part of the paper is organized into four different sections; Section 2 deals with the Text corpus and speech database creation. Section 3 deals with proposed model. Section 4 deals with Experimentation. Section 5 deals with discussion and conclusion.

2 Text Corpus and Speech Database Creation

Text corpus of 10 million words has collected from Dr. K. Naryana Murthy, Professor, Department of Computer and

- Hemakumar G. Research Scholar, Bharathiar University. Department of Computer Science, Government College for Womens', Mandya. hemakumar7@yahoo.com
- Dr. Punitha, Professor, Department of MCA, PESIT, Bangalore.

Information science, University of Hyderabad, Hyderabad, India in the year 2011. The top 10,000 most frequently occurred words are taken from this corpus. These 10,000 words have occurred 6 million times in Hampi text corpus. 10,000 words is record at sampling rate of 8 KHz, 16 bps, mono channel by one adult male speaker for 3 times each word for training and rerecorded each word for testing purpose. These signals were recorded at a little noisy environment, while Gold Wave Software was used to record with the help of mini microphone of frequency response 50 – 12500Hz.

3 Proposed Method

In this experiment we have designed algorithm in five stages for speaker dependent isolated Kannada word recognition. The proposed model works in offline mode. So all speech signals are pre-recorded and stored in speech database and then passed on to our algorithm for training or testing the unknown signal.

First stage is Pre-processing stage: In this stage analog speech signal is sampled and quantized at the rate of 8,000 samples/s. S(n) is the digitalized value. Then DC component is removed from digitalized sample value using the formula S(n) = S(n) – mean(S). A first order (low-pass) pre-emphasis ŝ(n) = S(n) – ã * S(n-1) network formula is used to compensate for the speech spectral fall-off at higher frequencies and approximates the inverse of the mouth transmission frequency response. Then standardization is done to entire set of values to have standards amplitude. This process will increases or decreases the amplitude of speech signal using the $S(n) = \hat{s}(n) - \max(|s|)$. Here we have used the constant value ã = *0.9955*.

The second stage is Detection of Voiced/ Unvoiced part in speech signal, also called speech signal segmentation. To solve this problem, using dynamic threshold approach, we have designed an algorithm for automatic segmentation of speech signal into sub-word or syllable [11]. Here we have combined the short time energy and magnitude of frame. Dynamic threshold for each frame is detected. Lastly, it is checked for voiced part in that frame using that frame threshold. This is achieved by following these steps

$$Thr_{STE} = \left(\left[\frac{\sum_{i=1}^{n} STE}{n}\right] - [\min(STE) * 0.5]\right) + \min(STE) \quad - (3.1)$$

$$Thr_{msf} = \left(\left[\frac{\sum_{i=1}^{n} msf}{n}\right] - [\min(msf) * 0.6]\right) + \min(msf) \quad - (3.2)$$

$$if\ (STE \geq Thr_{STE}) then\ marked\ has\ Voiced_{STE} = 1 \quad - (3.3)$$

$$if\ (msf > Thr_{msf})\ then\ marked\ has\ Voiced_{msf} = 1 \quad - (3.4)$$

$$if\ (Voiced_{STE} * Voiced_{msf} = 1)\ then\ that\ frame\ contains\ voice, \quad otherwise\ its\ unvoiced\ frame$$

where STE is Short Time Energy, msf is the Magnitude of Frame, n is number of samples in the frame. The fig 1 shows the voice part detected and segmented into syllable, sub-word or word level.

Feature Extraction is the Third stage: Here we have selected the voiced part of signal and then frame blocking was done for *N* samples with adjacent frames spaced *M* samples apart. Typical values for *N* and *M* correspond to frames of 20 ms duration with adjacent frames overlap by 6.5 ms. A hamming window is applied to each frame using frame same size. Next, the autocorrelation is applied to that part of signal. LPC method is applied to detect LPC coefficients. The LPC coefficients are converted into Real Cepstrum Coefficients. Here the outputted data will be of the size p*L, where p is the LPC order and it will be constant and L is the number of frames in that voice segmented parts. So it varies. In our experiment we have used LPC order p=24.

The Fourth stage is Speech model building: In this stage the real cepstrum coefficients are in dimension of p*L matrices. This matrix will be passed into k-means algorithm by keeping k=3 and outputted values are passed into 3 state Baum–Welch algorithm and each syllable or sub-word is trained. The Baum-Welch re-estimation procedure is the stochastic constraints of the HMM parameters

$$\sum_{i=1\ldots N} \bar{\pi}_i = 1 \quad - (3.5)$$

$$\sum_{j=1\ldots N} \bar{A}_{ij} = 1, 1 \leq i \leq N \quad - (3.6)$$

$$\sum_{k=1\ldots M} \bar{B}_j(k) = 1, 1 \leq j \leq N \quad - (3.7)$$

Are automatically incorporated at each iteration. The parameter estimation problem as a constrained optimization of P(O | λ). Based on a standard Lagrange optimization setup using Lagrange multipliers, P is maximized by

$$\pi_i = \frac{\pi_i(\partial P/\partial \pi_i)}{\sum_{k=1\ldots N} \pi_k(\partial P/\partial \pi_k)} \quad - (3.8)$$

$$A_{ij} = \frac{A_{ij}(\partial P/\partial A_{ij})}{\sum_{k=1\ldots N} A_{ik}(\partial P/\partial A_{ik})} \quad - (3.9)$$

$$B_j(k) = \frac{B_j(k)(\partial P/\partial B_j(k))}{\sum_{l=1\ldots M} B_j(l)(\partial P/\partial B_j(l))} \quad - (3.10)$$

Normal fit is applied for 2 consecutive HMM parameter

λ(A, B, pi) and Normal fit parameters are computed. Her the trained two consecutive λ(A, B, pi) are considered has sample data. So, we will be having a sample ($x_1 \ldots x_n$), for this a normal parameter $N(\hat{\mu}, \hat{\sigma}^2)$ is computed by using the

$$\hat{\mu} = \bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\sigma}^2 \sim \frac{\sigma^2}{n}.X_{n-1}^2 - (3.11)$$

The labelled $\hat{\mu}$ and $\hat{\sigma}^2$ value will be classified according to acoustic classes and then stored. Those data have representatives of syllables or sub-words in that particular class. In Language model we have designed bi-syllable language model for each word.

The Fifth stage is Recognition part / Testing Unknown Signal: Initially, for the unknown speech signals HMM parameters are computed and passed into normal fit method. Subsequently, the outputted $\hat{\mu}$ and $\hat{\sigma}^2$ value is identified and then matched with trained set of data by retaining threshold values. The outputted syllables or sub-words are matched with the bi-syllable language model. The concatenation of outputted syllables and sub-words are done for word building. On this basis decision is taken has recognized word by checking for top ranked.

4 Experimentation

In this paper experimentation are done on recognition of isolated Kannada words using HMM (3 state Baum-Welch Algorithm alone), GMM and compared with proposed model for same speech database. To experiment programs are written in mat lab and ruined on Intel Core i5 processor speed of 2.67 GHz and RAM of 3 GB. The table 1 shows the details of memory required to store speech models for different vocabulary size, figures are in Kilo bytes. This shows that our model requires the less memory to store speech models. The table 2 shows the average accuracy rate for different size of vocabulary.

TABLE 1
SHOWS THE MEMORY REQUIRED TO STORE SPEECH MODELS HAS A WORD REPRESENTATIVE IN KILO BYTES FOR DIFFERENT VOCABULARY SIZE.

Methods / Words	HMM + Normal fit	HMM	GMM
1000 Words	340.92	564	368
2000 Words	681.84	1128	736
3000 Words	1022.76	1692	1104
4000 Words	1363.20	2259.20	1473.60
5000 Words	1704.40	2825	1843
6000 Words	2045.28	3390	2211.60
7000 Words	2385.88	3954.44	2579.64
8000 Words	2726.72	4519.36	2948.16
9000 Words	3066.84	5084.28	3316.68
10000 Words	3407.60	5648.40	3685.60

TABLE 2
Shows the Average Accuracy Rate measured for different vocabulary size.

Methods/ Words	HMM + Normal fit	HMM	GMM
1000 Words	92.98%	83.45%	91.90%
2000 Words	92.13%	82.99%	91.54%
3000 Words	92.02%	82.21%	91%
4000 Words	92.73%	82.01%	90.78%
5000 Words	92.69%	81.90%	90.12%
6000 Words	91.11%	81.77%	90.01%
7000 Words	90.30%	81.01%	89.05%
8000 Words	89.44%	80.32%	88.75%
9000 Words	89.42%	80.15%	88.66%
10000 Words	89.39%	80.05%	88.45%
Average	91.22%	81.59%	90.03%

5 Discussion and Conclusion

In this paper, ASR model is designed by combination of HMM and Normal fit method and experimented for recognizing the isolated Kannada words. Our ASR model is compared with HMM (3-state Baum-Welch Algorithm alone) and GMM for same speech database. The space required to store the model datum has syllable or sub-word representatives in the HMM and GMM required more memory than storing the normal fit parameters. A normal fit method shows the better accuracy rate then the other two methods. This experiment shows that using normal fit (Normal Parameter estimation), ASR model can be designed and it takes less space with good accuracy rate compared to GMM and HMM models. Using our model ASR can be designed for small, medium and large vocabulary.

Acknowledgment

The authors would like to thank for Bharathiar University for giving an opportunity to pursuing part-time PhD degree. Authors would like to thanks for Prof M.R. Nandan, Former Principal, GCWM and all our friends, reviewers and Editorial staff for their help during preparation of this paper.

References

[1] Hemakumar G and Punitha P (2013), Speaker Independent Isolated Kannada Word Recognizer, published by P. P. Swamy and D. S. Guru (eds.) (2013), Multimedia Processing, Communication and Computing Applications, Lecture Notes in Electrical Engineering 213, DOI: 10.1007/978-81-322-1143-3_27, Springer India, Page No 333-345.

[2] Bishnu Prasad Das and Ranjan Parekh (2012), Recognition of Isolated Words using Features based on LPC, MFCC, ZCR and STE, with

Neural Network Classifiers, International Journal of Modern Engineering Research (IJMER), Vol.2, Issue.3, May-June 2012 pp-854-858.

[3] Siva Prasad Nandyala and T. Kishore Kumar (2012), Real Time Isolated Word Recognition using Adaptive Algorithm, International Conference on Industrial and Intelligent Information (ICIII 2012), IPCSIT vol.31 © (2012) IACSIT Press, Singapore.

[4] http://www.mathworks.in/help/stats/statset.html.

[5] Hemakumar G. and Punitha P., (2013), Speech Recognition Technology: A Survey on Indian Languages, International Journal of Information Science and Intelligent System, Vol. 2, No.4, 2013, Page No 1-38

[6] Santosh K.Gaikwad et al., (November 2010), A Review on Speech Recognition Technique, International Journal of Computer Applications (0975 – 8887), Volume 10– No.3.

[7] Rabiner L, Jung B-H (1993), Fundamentals of speech recognition, Pearson Education (Singapore) Private Limited, Indian Branch, 482 F.I.E Patpargans, Delhi 110092, India.

[8] David Doria (2009), Expectation-Maximization: Application to Gaussian Mixture Model Parameter Estimation, Lecture notes published on April 23.

[9] Lawrence R. Rabiner et al., (1979), Speaker-Independent Recognition of Isolated Words Using Clustering Techniques, IEEE Transactions on Acoustics, Speech, And Signal Processing, Vol. Assp-27, No. 4, August 1979, page No. 336-349.

[10] Carlo Tomasi, Estimating Gaussian Mixture Densities with EM – A Tutorial, by Duke University.

[11] Dimo Dimov and Ivan Azmanov (2005), Experimental specifics of using HMM in isolated word speech recognition, International Conference on Computer Systems and Technologies – *CompSysTech*..

[12] Sukhminder Singh Grewal1 & Dinesh Kumar (2010), Isolated Word Recognition System For English Language, International Journal of Information Technology and Knowledge Management, July-December 2010, Volume 2, No. 2, pp. 447-450.

[13] A.Revathi (2009) et al., Text Independent Speaker Recognition and Speaker Independent Speech Recognition Using Iterative Clustering Approach, International Journal of Computer science & Information Technology (IJCSIT), Vol. 1, No 2, November 2009.

[14] Hemakumar G. and Punitha P. (2013), Automatic Segmentation of Kannada Speech Signal into Syllables and Sub-words: Noised and Noiseless signals, International Journal of Scientific & Engineering Research, Volume 5, Issue 1, January-2014, pag no. 1707-1711.