Large Vocabulary Speech Recognition: Speaker Dependent and Speaker Independent

G. Hemakumar and P. Punitha

Abstract This paper addresses the problem of large vocabulary isolated word and continuous Kannada speech recognition using the syllables and combination of Hidden Markov Model (HMM) and Normal fit method. The models designed for speaker dependent and speaker independent mode of working. This experiment has covered 6 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 labeled has syllable or sub-word. In terms of memory requirement and recognition rate the proposed model is compared with Gaussian Mixture Model and HMM (3-state Baum–Welch algorithm). This paper clearly shows that combination of HMM and normal fit technique will reduce the memory size while building and storing 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.

Keywords Speaker independent · Speaker dependent · Normal fit · 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

G. Hemakumar (✉)
Department of Computer Science, Government College for Women, Mandya, India
e-mail: hemakumar7@yahoo.com

P. Punitha
Department of MCA, PESIT, Bangalore, India
e-mail: punithaswamy@gmail.com

J.K. Mandal et al. (eds.), *Information Systems Design and Intelligent Applications*,
Advances in Intelligent Systems and Computing 339,
DOI 10.1007/978-81-322-2250-7_8

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 five stages namely converting analog speech signal into Digitalization (Normalization part) form, Speech signal segmentation or Voice part detection, 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 [1–3].

The objective of modeling 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 [2–4].

This paper discussing the large vocabulary speaker dependent and speaker independent 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. This paper also discuss on large vocabulary continuous Kannada speech recognition for speaker dependent and speaker independent using syllable, combination of HMM and normal fit technique and tri-syllable language model.

The remaining part of the paper is organized into four different sections; Sect. 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 Information science, University of Hyderabad, Hyderabad, India in the year 2011. The top 10,000 words most frequently occurred in this corpus are taken. These 10,000 words have occurred 6 million times in Hampi text corpus. Those 10,000 words are record at sampling rate of 8 kHz, 16 bps, mono channel by one adult male speaker by uttering 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 of 50–12,500 Hz.

Secondly Kannada speech corpus is designed for selected 294 words to design the speaker independent recognition model. We have taken age group of 16–60 years native Kannada speakers. Here 5 districts dialects are recorded namely Mysore, Bangalore, Mandya, Chamarajnagar and Ramanagar districts located at southern part of Karnataka state. The signals are recorded at the sampling rate of 8 kHz, 16 bps with mono channel using mini microphone of frequency response of 50–12,500 Hz.

Thirdly, continuous Kannada speech corpus is designed by randomly selecting 250 sentences from Hampi text corpus. The details are shown in Table 1.

Fourthly, continuous Kannada speech designed by IIIT Hyderabad is collected, which consist of 1,000 unique sentences recorded at room environment and pulse code modulation with a frequency of 16,000 Hz/s and 16-bit mono channel.

Table 1 Continuous Kannada speech corpus designed by us for randomly selected sentences from Hampi text corpus

Language	Kannada
Speech type	Sentence read from documents
Number of sentences used	250 Sentence for minimum length of 2 words, maximum length of 47 words
Number of unique words	2,419 Words
Number of speakers for training	20 Speakers (10 female + 10 male)
Number of speakers for testing	2 (known) + 2 (unknown) = 4 female and 4 male, total = 8 speakers
Speech sampling rate	16 kHz/s, 16 bit mono channel
Recording conditions	Room environment
Number of signals used to training	5,000 Signals
Number of signals used to testing	2,000 Signals
Total signals using in experiment	7,000 Signals
Age categories	19–60 years aged
Total hours of recording	30 h
Total memory size	1.43 GB

3 Proposed Method

In this experiment we have designed algorithm in five stages for speaker dependent, speaker independent isolated Kannada word recognition and continuous Kannada speech. 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 or 16,000 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 $\hat{s}(n) = S(n) - \tilde{a} * S(n-1)$ 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 $\tilde{a} = 0.9955$.

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 [5]. 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

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

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

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

$$if\ \left(msf > Thr_{msf}\right) then\ marked\ has\ Voiced_{msf} = 1 \tag{3.4}$$

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

where STE is Short Time Energy, msf is the Magnitude of Frame, n is number of samples in the frame.

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

$$\sum_{\mathrm{i}=1\ldots\mathrm{N}} \bar{\pi}_1 = 1 \tag{3.5}$$

$$\sum_{\mathrm{j}=1\ldots\mathrm{N}} \bar{\mathrm{A}}_{\mathrm{ij}} = 1, 1 \leq \mathrm{i} \leq \mathrm{N} \tag{3.6}$$

$$\sum_{\mathrm{k}=1\ldots\mathrm{M}} \overline{\mathrm{B}}_{\mathrm{j}}(\mathrm{k}) = 1, 1 \leq \mathrm{j} \leq \mathrm{N} \tag{3.7}$$

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

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

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

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

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

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

The labeled $\hat{\mu}$ 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 isolated word and tri-syllable language model for continuous speech.

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 and continuous Kannada speech using HMM (3 state Baum-Welch Algorithm alone), GMM and compared with proposed model for same speech database. All experiment programs are written in mat lab and ruined on Intel Core i5 processor speed of 2.67 GHz and RAM of 3 GB. Table 2 shows the details of memory required to storing speech models for different vocabulary size, figures are in Kilo bytes and also shows the average accuracy rate for different size of vocabulary. This shows that our model requires the less memory to store speech models.

The average WRR for IIIT Hyderabad speech corpus is 95.87 % and average WER is 4.13 %. The average WRR for randomly selected 250 sentences from Hampi text corpus is 86.5 % and average WER is 13.5 %. Our model is tested on noised (little) and noiseless signal and the average success rate of noised continuous Kannada speech signals for known speaker and unknown speaker is 81 %.

Table 2 Shows the average accuracy rate measured with different vocabulary size

Methods / words	HMM		GMM		HMM + Normal fit	
	Accuracy rate (%)	Memory size	Accuracy rate (%)	Memory size	Accuracy rate (%)	Memory size
1,000 Words	83.45	564	91.90	368	92.98	340.92
2,000 Words	82.99	1,128	91.54	736	92.13	681.84
3,000 Words	82.21	1,692	91	1,104	92.02	1,022.76
4,000 Words	82.01	2,259.2	90.78	1,473.6	92.73	1,363.2
5,000 Words	81.90	2,825	90.12	1,843	92.69	1,704.4
6,000 Words	81.77	3,390	90.01	2,211.6	91.11	2,045.28
7,000 Words	81.01	3,954.44	89.05	2,579.64	90.30	2,385.88
8,000 Words	80.32	4,519.36	88.75	2,948.16	89.44	2,726.72
9,000 Words	80.15	5,084.28	88.66	3,316.68	89.42	3,066.84
10,000 Words	80.05	5,648.4	88.45	3,685.6	89.39	3,407.6
Average	81.59		90.03		91.22	

Memory required storing the speech models shown in kilobytes

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 and continuous Kannada speech. 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. And also ASR can be design for speaker dependent, speaker independent mode of working and isolated word, connected words and continuous speech.

Acknowledgments The authors would like to thank for Bharathiar University for giving an opportunity to pursuing part-time Ph.D. degree. Authors would like to thanks for all our friends, reviewers and Editorial staff for their help during preparation of this paper.

References

1. Swamy, P.P., Guru, D.S. (eds.): Multimedia Processing, Communication and Computing Applications. Lecture Notes in Electrical Engineering, vol. 213, pp. 333–345. Springer, India. doi:10.1007/978-81-322-1143-3_27 (2013)
2. Hemakumar, G., Punitha, P.: Speech recognition technology: a survey on Indian languages. Int. J. Inf. Sci. Intell. Syst. **2**(4), 1–38 (2013)
3. Gaikwad, S.K., et al.: A review on speech recognition technique. Int. J. Comput. Appl. (0975–8887) **10**(3), 16–24 (2010)
4. Rabiner, L., Jung, B.-H.: Fundamentals of Speech Recognition. Pearson Education, Singapore (1993)
5. Hemakumar, G., Punitha, P.: Automatic segmentation of Kannada speech signal into syllables and sub-words: noised and noiseless signals. Int. J. Sci. Eng. Res. **5**(1), 1707–1711 (2013)
6. Mat lab R2009a help menu on statistics toolbox online: www.mathworks.com/help/
7. http://en.wikipedia.org/wiki/normal_distribution#cite_note-kri127-33
8. http://www.mathworks.in/help/stats/statset.html
9. David, D.: Expectation-Maximization: Application to Gaussian Mixture Model Parameter Estimation. Lecture Notes Published on April 23 (2009)
10. Rabiner, L.R., et al.: Speaker-independent recognition of isolated words using clustering techniques. IEEE Trans. Acoust. Speech Sig. Process. **27**(4), 336–349 (1979)
11. Carlo, T.: Estimating Gaussian Mixture Densities with EM—A Tutorial. Duke University
12. Dimov, D., Azmanov, I.: Experimental specifics of using HMM in isolated word speech recognition. In: International Conference on Computer Systems and Technologies—CompSysTech (2005)
13. Grewal, S.S., Kumar, D.: Isolated word recognition system for English language. Int. J. Inf. Technol. Knowl. Manag. **2**(2), 447–450 (2010)

14. Nandyala, S.P., Kishore Kumar, T.: Real time isolated word recognition using adaptive algorithm. In: International Conference on Industrial and Intelligent Information (ICIII 2012), IPCSIT, vol. 31 © (2012). IACSIT Press, Singapore (2012)
15. Revathi, A., et al.: Text independent speaker recognition and speaker independent speech recognition using iterative clustering approach. Int. J. Comput. Sci. Inf. Technol. (IJCSIT) **1** (2), 30–40 (2009)
16. Das, B.P., Parekh, R.: Recognition of isolated words using features based on LPC, MFCC, ZCR and STE, with neural network classifiers. Int. J. Mod. Eng. Res. (IJMER) **2**(3), 854–858 (2012)