Original Article

Speaker Accent and Isolated Kannada Word Recognition

Hemakumar G.* and Punitha P.

[1]*Department of Computer Science Government College for Women, Mandya Karnataka, India*
[2]*Department of MCA PESIT, Bangalore Karnataka, India*

***Corresponding Email:** hemakumar7@yahoo.com*

ABSTRACT

Algorithm is designed for isolated Kannada word recognition of five districts Kannada speakers' accent. Isolated Kannada words recognition is designed using the syllables, Baum-Welch algorithm and Normal fit method. The novelty of proposed method is in recognition of five district Kannada speaker accents as well as spoken words. Our model is compared with baseline Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM). Our model is tested for both noisy (little) and noiseless signal. For experiment totally 7056 signals are used for training and 3528 signals are used for testing. The experiment showed that average WRR for known accent speaker is 95.56% and of unknown accent speaker is 90.86% and average recognition of Kannada speaker accents is 82%.

Keywords: ASR, Speaker accent, HMM, Normal fit.

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. The reason is, for human beings speech is the primary form of communication. So human beings wanted to communicate with machine or computer through speech input for computation and get output in the form text or speech. 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. ASR technology has the basic quality to help humans easily communicate with machines / computers and reap maximum benefit from them. The performance of ASR has improved

dramatically due to recent advances computer technology with continually improving algorithms and faster computing. But still ASR has facing lot of challenges like speaker accents, speaking style, emotion, confusing words, and memory requirement and so on.

The ASR system may be viewed as working in a four stages namely converting analog speech signal into Dig- ital (Normalization part) form, voice part detection in given speech signal and Feature extraction stage, Speech Model building and accent class building part, and testing the unknown speech signal. 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 [4,13].

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, Artificial Intelligence approach, stochastic approach, Knowledge based approaches, Statistical based approaches, and Learning based approaches[4,10,13].

ASR systems need to be general use systems, they have to support multiple speakers and it shall be have ability for adapting to all the speaker variations. The variations are in speech styles, pitch and anatomy which make each speaker unique. Also things like background noise, utterances, and accents can negatively affect the interpretation of speech. Even words that sound alike can create problems for ASR systems. In paper[1] focus on some variations within the speech signal that make the ASR task difficult. The variations detailed in the paper[1] are intrinsic to the speech and affect the different levels of the ASR processing chain. According to them variation of speech occurs by Speaker characteristics, Foreign and regional accents, Speaking rate and style, Age of speaker and speakers Emotions. In paper[12], discuss about two novel algorithms to improve dialect classification for text-independent spontaneous speech in Arabic and Spanish languages, along with probe results for Chinese. Problem they considers here is, there should be no transcripts but dialect labels are available for training and test data, and speakers are speaking spontaneously, which is defined as text-independent dialect classification. The Gaussian mixture model (GMM) is used as the baseline system for text-independent dialect classification. In the training phase, a symmetric version of the Kullback–Leibler divergence is used to find the most discriminative GMM mixtures (KLD-GMM), where the confused acoustic GMM region is suppressed. For testing, the more discriminative frames are detected and used via the location of where the frames are in the GMM mixture feature space, which is

termed frame selection decoding (FSD-GMM). The first KLD-GMM and second FSD-GMM techniques are showed to improve dialect classification performance for three-way dialect tasks. The two algorithms and their combination are evaluated on dialects of Arabic and Spanish corpora.

This paper shows the speaker independent isolated Kannada word recognition and accent identification using LPCC, Syllable, HMM and Normal fit technique and compared with HMM and GMM. The proposed model clear shows that memory size reduction while storing the speech model and increasing the accuracy of recognition. The programs written in mat lab for isolated word recognition will recognize only trained set of words for five districts Kannada accents. This model recognizes the accents of Kannada speaking at South-Eastern region of Karnataka. The districts covered are Mysore, Bangalore, Mandya, Chamarajnagara and Ramanagara. While designing the Kannada isolated word speech corpus, we have selected the 294 unique words by our self and asked to speak the same words from each speaker. These words are selected to train the different phoneme present in Kannada. This paper will not discuss about difference between the dialects and phonemes, instead of that it discuss only the design of speaker independent recognition for above mentioned accents.

The remaining part of the paper is organized into four different sections; Section 2 deals with proposed model. Section 3 deals with Normal Fit method. Section 4 deals with Experimental results. This section further sub-divided into 2 sub-sections which provide details of Speech Signal Database, and experimentation results. Section 5 deals with discussion and conclusion.

Proposed Method

The proposed system works in offline mode, where the speech signal is prerecorded and stored for processing. First stage is Pre-processing or Normalization of speech signal: In this stage analog speech signal is converted into digital form values. Then DC component is removed from digitalized sample values using the formula S(n) = S(n) – mean(S). A first order (high-pass) pre-emphasis network formula ŝ(n) = S(n) – ã * S(n-1) is used to compensate for the speech spectral falloff at higher frequencies and approximates the inverse to the mouth transmission frequency response. Then standardization of amplitude is done using the formula S (n) = ŝ (n) / max (| (ŝ (n) |).

The second stage is Feature Extraction stage: For standardized speech signal the frame blocking is done for *N* samples, with adjacent frames spaced *M* samples apart. Typical values for *N* and *M* correspond to frames of duration 20 ms, with adjacent frames overlap by 6.5 ms in this case. A Hamming window is applied to each frame using same size of frame, and then autocorrelation is applied to that part of signal. For each frames short time energy (STE) and magnitude of frame (MSF) is computed. Then dynamic threshold is computed by fusion of STE and MSF for each frame. Then decision is taken whether that frame is voiced or not. The formula gives information regarding how voice part is detected by us in each frame by computing dynamic thresholds for each frame[513].

$$Thr_{STE} = \left(\left[\frac{\sum_{i=1}^{n} STE}{n}\right] - [\min(STE) * 0.5]\right) + \min(STE) \quad (2.1)$$

$$Thr_{msf} = \left(\left[\frac{\sum_{i=1}^{n} msf}{n}\right] - [\min(msf) * 0.6]\right) + \min(msf) \quad (2.2)$$

$$if\ (STE \geq Thr_{STE}) then\ marked\ has\ Voiced_{STE} = 1 \quad (2.3)$$

$if\ (msf > Thr_{msf})\ then\ marked\ has\ Voiced_{msf} = 1$ (2.4)

$if\ (Voiced_{STE} * Voiced_{msf} = 1)\ then$

that frame contains voice, otherwise its unvoiced frame

If that frame is voiced, then LPC method is applied to detect LPC coefficients. The LPC Coefficients are converted into Real Cepstrum Coefficients.

The third stage is Speech model building: In this stage the real cepstrum coefficients are passed into k-means clustering algorithm by keeping k=3. Then using 3-state Baum–Welch algorithm each syllables or sub-words values are trained. The trained output parameter are λ(A, B, Pi). Multiplication of matrices A and B is done, then we get 3*P ordered matrices. For this matrices row wise summation is done. Then we get vector, for this vector normal fit is applied and outputted mean hat and sigma hat values will be labeled has syllable or sub-word. Those mean hat and sigma hat is stored in our database has representative of the labeled syllable or sub-word. During storing in database the labeled mean hat and sigma hat value will be identified has in which accent it belong and inside that accent again classification is done according to acoustic feature and labeled values are stored in that particular class. Consecutive of two normal fit values is summed and bi-syllable language model is designed for the purpose of word recognition. Here first accent classification is done and then syllable or sub-word classifications are done according to its acoustic features inside the each accent class.

The final stage is Recognition part or Testing Unknown Signal: First for the unknown signals feature extraction is done and then using 3-state Baum–Welch algorithm, HMM parameters is computed. Normal fit is applied for HMM parameter. Successor of two normal fit is summed. Then mean hat and sigma hat values are matched with stored speech models by retaining some threshold values. The outputted Syllable or Sub-word is matched with the Bi-Syllable language model for the isolated word recognition. The decision is taken and outputted the top ranked matched word has recognition. After recognition of word, identification of accent is done by counting the maximum number of outputted syllable in which class it lies. Then declared the speaker utterance lies in that particular accent is class.

Normal Fit (Normal Parameter Estimates)

For the output of 3-state Baum–Welch algorithm parameter the normal fit is applied. First matrices A and B are multiplied and then row wise summation is done (column wise addition will always equal to 1). The outputted vector is considered has the sample ($x_1 \ldots x_n$) data, for this samples data normal parameter $N(\mu, \sigma^2)$ is computed. Here, we would like to learn the approximate the values of parameters μ and σ^2. The standard approach to this problem is the maximum likelihood method, which requires maximization of the log-likelihood function[6,7]:

$$\ln L(\mu, \sigma^2) = \sum_{i=1}^{n} \ln f(x_i; \mu, \sigma^2)$$

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \quad (3.1)$$

Taking derivatives with respect to μ and σ^2 and solving the resulting system of first order conditions yields the maximum likelihood estimates using[6]:

$$\hat{\mu} = \bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (3.2)$$

Estimator μ is called the sample mean, since it is the arithmetic mean of all observations. The statistic $\bar{x}$ is complete and sufficient for μ, and therefore by the Lehmann-Scheffe theorem, $\hat{\mu}$ is the uniformly minimum variance unbiased (UMVU) estimator[6,78]. The estimator $\hat{\sigma}^2$ is

called the sample variance, since it is the variance of the sample ($x_1 \ldots x_n$) used in computing mean squared error (MSE) criterion. It is a biased estimator.

In this paper the outputted $\hat{\mu}$ and $\hat{\sigma}^2$ is labeled and then stored has representative of syllables or sub-words. The concatenation of this word representative is designed.

EXPERIMENTATION

Signal Preprocessing

See table: 1.

In this experiment, signals were recorded at a room environment. The pulse code modulation with a frequency of 8000 Hz/Sec, 16-bit mono channel is used. For recording purpose we have designed recording program using mat lab. The recording is done with the help of mini microphone of frequency response 50 – 12500 Hz. Each word signal contains with a silence region before and after right signal. The detail of speech database is described in table-1.

Experimental Results

In this paper experimentation are done on recognition of isolated Kannada words uttered by 5 districts Kannada speakers accents using baseline model has HMM, GMM and compared with our proposed model for same speech database. To experiment programs are written in mat lab and run on Intel Core i5 processor speed of 2.30 GHz and RAM of 4 GB. This model programs recognize only the trained set of words for an accent of southern Karnataka region speaking at normal condition in room environment. The table-II shows the details of memory required to store speech models for different vocabulary size, figures are in Kilo bytes. This experiment shows that our model requires the less memory to store speech models. The table-III shows details of average of 12 speakers Average Accuracy Rate measured for different vocabulary size. It's clearly shows that our proposed model gives better Recognition rate comparing to those other two methods.

The average recognition of Kannada speaker accents showed 82% accuracy rate. In this experiment our goal was only to recognize the speaker accent and Kannada word. So this speech model will not gives good accuracy rate in gender classification and age classification.

See table: 2, 3.

DISCUSSION AND CONCLUSION

In this experiment we agree that our speech database is very small, because only 30 speakers' are selected from 5 districts. But still we got good WRR. Our model shows that one speech model can be designed to recognize the speakers' accent and the word recognition. If the speech corpus is huge then definitely we can design the one speech model which can recognize the accent, age, gender and the spoken words. In this paper, ASR model is designed by combination of 3-state Baum–Welch algorithm and Normal fit method and experimented for recognizing the isolated Kannada words. Our ASR model is compared with baseline HMM (3-state Baum-Welch Algorithm alone) and GMM for same speech database. The memory size required to store the speech model has syllable or sub-word representatives in the HMM the lambda values are used to stored. So it takes more memory. In combination of 3-state Baum–Welch algorithm and normal fit only mean hat and sigma hat will be stored for each syllables or sub-word has representative (mean of syllables are taken according to their classification of gender and five districts accents). The memory required will be very less compared to other models.

Normal fit is also derived from Gaussian method. So, GMM and Normal fit methods showed the very high accuracy rate in our experiment. But still GMM accuracy rate is less then proposed model. The memory required to store speech models and time taken to testing the speech signal in GMM takes little more than our model. This experiment shows that using Baum-Welch algorithm and normal fit (Normal Parameter estimation) method, ASR model can be designed and it takes less space compared to GMM and HMM models. The experiment shows that the identification and storing the speech model in proper class of accents and acoustic feature will increases the accuracy rate. Using our model speaker independent recognition system can be designed for small, medium and large vocabulary with good word recognition rate and with less memory.

ACKNOWLEDGMENTS

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

REFERENCES

1. Benzeguiba M et al (2006), "Automatic Speech Recognition and Intrinsic Speech Variation*", Proc of IEEE ICASSP 2006*, page no. 1021-1024. DOI: 142440469X /06/$20.00 ©2006 IEEE
2. Carlo Tomasi, "*Estimating Gaussian Mixture Densities with EM – A Tutorial",* Duke University. Online: *https://www.cs.* ***duke****.edu/courses/spring04/cps196.1/.../****toma si****EM.pdf*
3. David Doria (2009), "*Expectation-Maximization: Application to Gaussian Mixture Model Parameter Estimation"*, Lecture notes published on April 23. www.*engineering****notes****.net/****Notes****/EE/Presentations/GMM_MLE_EM.pdf*
4. 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, ISSN 2307-9142.
5. 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, page no. 1707-1711. ISSN 2229-5518.
6. http://en.wikipedia.org/wiki/normal_distribution#cite_note-kri127-33.
7. Math work Documentation center on statistics toolbox, exploratory data analysis, cluster analysis, Gaussian mixture models, gmdistribution.fit Online: http://www.mathworks.in/help/stats/statset.html.
8. Mat lab R2009a help menu on statistics toolbox online : www.**math**works.com/**help**/
9. 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.
10. 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.
11. Umesh S, "Studies on inter-speaker variability in speech and its application in automatic speech recognition", *Sa¯dhana¯* (Vol. 36, Part 5, October 2011), pp. 853–883, © Indian Academy of Sciences.
12. Yun Lei and John H. L. Hansen, "Dialect Classification via Text-Independent Training and Testing for Arabic, Spanish, and Chinese", *IEEE Transactions On Audio, Speech, And Language Processing*, (VOL. 19, NO. 1, JANUARY 2011), Page no 85-96.
13. Hemakumar G. and Punitha P., "Large vocabulary isolated word recognition using syllable, hmm and normal fit", published by *International Journal of Scientific &*

Engineering Research, Volume 5, Issue 9, Sept-2014, ISSN 2229-5518.

Table 1. Speaker Independent Speech Database Description

Language	Kannada
Speech type	Read Speech
Number of Words Used	294 Words
Number of Speakers	30 speaker of different age categories
Speakers Age	Considered from 16 - 63 years old
Recording Conditions	Room Environment
Number of signals used to Training	24 speakers (12M + 12 F) = Total 7056 Signals
Number of signals used to Testing	12 speakers (6 M + 6 F), among 12 speakers 6 speakers are known speakers and 6 are unknown speakers. Total 3528 Signals used.
Total Memory Size	763 Mega Byte
Total Signals used in experiment	10,584 Signals are used.

Table 2. Shows the memory size required to store speech models in kilo bytes (24 speakers for each word) for speaker independent recognition in different vocabulary size.

Method	50 Word	100 Word	150 Word	200 Word	250 Word	294 Word
HMM+ Normal fit	16.5	33	49.5	67.4	85.23	98.5
HMM	28.2	56.5	84.8	113	141	169.2
GMM	19.9	36.4	54.6	72.8	92	108

Table 3. Shows the average of 12 speaker accuracy rate measured for speaker independent recognition in different vocabulary size.

Methods	50 Word	100 Word	150 Word	200 Word	250 Word	294 Word
HMM + Normal fit	91.78%	94%	93.34%	93.56%	93.96%	93.25%
HMM	83.5%	82.89%	83.70%	84%	84.99%	84.74%
GMM	90.5%	91%	90.98%	91.19%	91.98%	91.82%