Isolated Word Recognition using Morph – Knowledge for Telugu Language

Dr. K. V. N. Sunitha
Prof & Head of CSE
G. Narayanamma Institute of Tech & Science
Hyderabad, India

N. Kalyani
Assoc Prof., CSE
G.Narayanamma Institute of Tech & Science
Hyderabad, India

ABSTRACT
Building a speech recognition system for Indian languages is an open question and requires focus. This paper highlights on a new model for speech recognition system and uses syllable as the basic unit. This model has five phases, the first three phases focused on training the data and building Trie structure to reduce the time and space and the last two phases are for testing. Training includes, first phase for syllable extraction from text and speech and annotating data sets. Second phase focuses on building the three state model for each syllable unit and third phase, for building Trie structure using morph knowledge of Telugu language. Testing includes the fourth and fifth phase. Fourth phase is to mark the rough boundary of the syllable using the intensity of the signal and these sequence of syllables are recognized during fifth phase. The experiment is conducted on CIIL Telugu corpus and achieved good results in recognizing the words that were not used for training. For training we have used 300 words and for testing we recorded 100 new words and 80% of the words were recognized.

General Terms
Computer Science - Speech Processing

Keywords
Five phase system, three state model, Trie structures, Syllable units, Word Model.

1. INTRODUCTION
Automatic Speech Recognition is process of converting speech utterance to text form. This process requires segmentation of speech signal into representable units and recognizing it using different approaches. The largest unit could be a sentence, and the smallest unit could be a single phone. There are reported works on speech recognition for few languages, each focused on choosing a different unit size. There are applications that were developed which have limited vocabulary like digits *Plauche et al.[1]* Automatic speech recognition system was developed at word level by *Lippmann et al.[2], Rabiner et al. [3]*. They represented the word utterances using the acoustic representation taking care of all the contextual effects with it. This system proved was best for limited vocabulary. The limitations with this system was, the training process is done with the individual word which can appear in any context as mentioned by *Huang et al.[4]* and the actual words occur in continuous form in natural speech. Second limitation is decrease in performance as training set increases. The third limitation is increase in memory requirements and increase in process time.

Syllable is a sub word unit which is found to be a promising unit for recognition. The importance of syllable was first reported by *Fujimura O[5]***.** The first successful robust LVCSR system was developed by *Ganapathiraju et al[6]* and he used syllable level acoustic unit in telephone bandwidth spontaneous speech. There are many papers published by *Nagarajan et. al.[7,8]***.** Their contributions being mainly on automatic segmentation of speech signal into syllabic unit using the short-term energy as magnitude spectrum and using group delay function to identify syllable boundaries. Our previous work was on analysis of coverage of syllables in words of the language. This analysis was done for Telugu text corpus developed by CIIL Mysore**.** *Dr.K.V.N.Sunitha,N.Kalyani [9]*.

Structurally, a syllable consists of three parts, an onset, a nucleus and a coda. An example structure of word ***BAgyaM*** with two syllable units is shown in the Fig 1. In general the syllables are represented as C*VC*. The appearance of consonants in preceding and succeeding positions is language dependent. In some languages there may be more than two consonants in either position. In English there is monosyllabic word "strength", this word in its canonical pronunciation has CCCVCCC form which is complex structure. Such complex structures are relatively rare in natural speech. In Telugu we rarely find such complex structures. The frequently found syllables are of the form V, CV, VC, CVC, and CCVC for most of Indian Languages.

The structure of the syllable is as shown where the onset corresponds to preceding consonant, a nucleus with the vowel and a coda to succeeding consonant respectively. In some cases the syllable can be formed only with a single vowel in such cases onset and coda is absent. In the above example the word ***BAgyaM*** (fortune) there are two syllables. The syllable ***BA*** has one consonant , one vowel and syllable ***gyaM*** has two consonants one at onset and the other at coda, and one vowel. In word ***ara*** - (rack) there are two syllables ***a*** and ***ra***. First syllable has only vowel and second syllable has one consonant and one vowel. It is observed that the energy levels increases in onset region and reaches to peak in nucleus and decreases in coda region. The words can be represented as sequence of syllables and most of the words have same common syllable sequence either as prefix or as suffix. Morphological study enables to group the words depending on the commonality.

Morphological analysis is an integral part of larger language processing projects such as text-to-speech synthesis, information extraction, syllable identification or machine translation. Words in Telugu language are morphologically rich. Hence the words can be represented using their morph structure. Morphological analysis includes the separation of stems and affixes (prefixes, suffixes, infixes, crucifixes) and the identification of inflectional and derivational processes.

These processes may be productive and may also be combined making an enumeration of morphological forms unfeasible as described in *Jurafsky et al.[10]*.

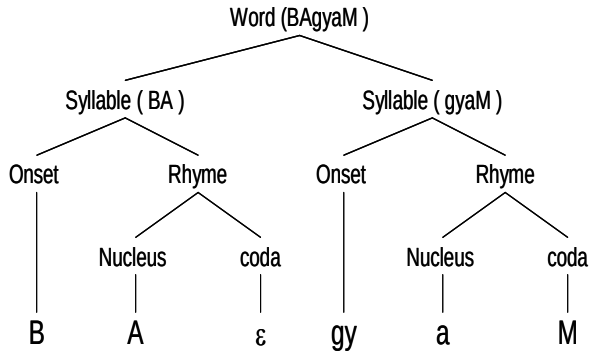

Fig 1: Detailed structure of a word BAgyaM

There is a body of related work that grows faster and faster as briefed in *Déjean H [11]* first induces a list of 100 most frequent morphemes and then uses those morphemes for word segmentation. His approach is thus not fully unsupervised. *Keshava et al.,[12]* combine the ideas of *Déjean H[11]* on the Morpho Challenge 2005 datasets, they achieved the best result for English. Other UMA learning algorithms exploit the Minimum Description Length (MDL) principle *(Mathias Creutz et al.,[13]*. *Brent et al. [14]* and were the first to introduce an information theoretic notion of compression to represent the MDL framework. *Goldsmith J A[15]* also used an MDL-based approach but applied a new compression system with different measuring of the length of the grammar. *Creutz et al.,[13]*. uses probabilistic distribution of morpheme length and frequency to rank induced morphemes. Our previous work, proposes an approach which presents a simple algorithm for unsupervised learning of morphological forms for inflectionally rich languages like Hindi and Telugu. Given a low coverage of morphologically related words and a corpus of raw text our approach can build all possible words belonging to similar group. This approach can be applied for any language that is morphologically rich ***Dr.K.V.N.Sunitha N.Kalyani, [16,17,18, 19] .***

This work proposes a new model for Isolated Word Recognition system that uses the morph knowledge. The next section gives the importance of the syllables and the statistical results obtained during the syllable analysis. Section 3 gives the block diagram of the proposed system that works in five phases and the details of the data structure build. Sample data used for training and new words used for testing are listed in section 4 followed by the conclusions and future scope.

2. SYLLABLE COVERAGE IN WORDS

The text segmentation is based on the linguistic rules derived from the language. Any syllable based language can be syllabified using these generic rules, to make the text segments exactly equivalent to the speech units. The syllable analysis is performed on Telugu corpus developed by CIIL Mysore.

Algorithms were developed to extract syllable units. The total distinct syllables observed are 12,378 and the frequency of occurrence of the syllables is plotted in the following chart. The number of Syllables with frequency less than 100 is 11057. It is observed that nearly 4903 syllables have frequency one. This is due to loanwords from English like (Apple, coffee, strength etc.) Fig 2 shows the count of Syllables with the frequencies in Hundreds.

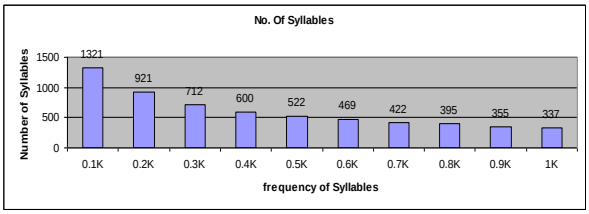

Fig. 2 Number of syllables with frequency in the range 100 to 1K.

It is observed that there are nearly 71 syllables that have frequency more than 10K A study is also made in terms of the words which have varying number of syllables with varying frequencies. Here in the following figure, plot is given for words which have syllables with cut-off frequency specified on X-axis and Y-axis indicates number of words having the syllable index above cut-off frequency and syllable index 0.5, 0.8 and 1.0.

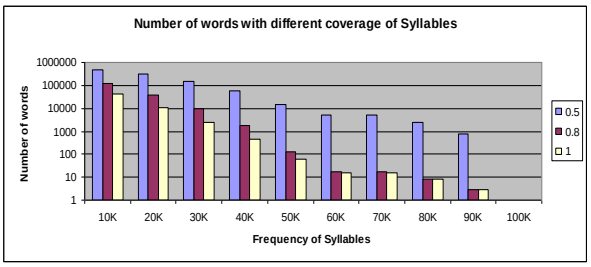

Fig. 3 Number of words having 50%, 80% and 100% syllables with syllable frequency in the range 10K to 100K.

Words count, with syllable index 50%, 80% and 100% and cut-off frequency varying in the range from 10K to 100K is shown in Fig.3. It is observed from the above figures that as the frequency increases the number of words included decreases. Importance of the word is dependent on syllable index and on the cut-of frequency.

This analysis is useful in selecting good set of words that would cover all possible syllables in large vocabulary. Optimal selection of words reduced the collection of speech samples for training the system.

3. PROPOSED SYSTEM

The proposed system is divided into five phases. The first phase is for extracting the syllables in both text and speech data. Text data is syllabified using the linguistic rules of Telugu and speech data is segmented into syllable unit semi automatically. Phase two is for building the three – state model for each syllable. For building the model the different samples of same syllable are identified and grouped together and the same units are used. Phase three builds the Trie structure which represents all the words as a single data structure.

This structure has the advantage that the search space of the word depends on the number of syllables in the word. Phase four is for preprocessing the collected speech sample for testing. Rough boundaries based on the intensity levels are marked using praat tool and syllable samples are extracted and placed in separate folder. The fifth phase reads these units

and compares with the syllables arranged in the Trie structure. Once the leaf node is reached then it concatenate all the syllables that are in the path from the root node to the leaf which forms the syllable sequence of the word. The block diagram is shown in the following Fig 4.

3.1 Syllable Extraction

Speech samples are collected and its corresponding text is selected and annotated at syllable level. For extracting the syllable units from text, linguistic rules of Telugu were applied and for extraction the syllable units in speech sample Praat tool is used to mark the boundaries.

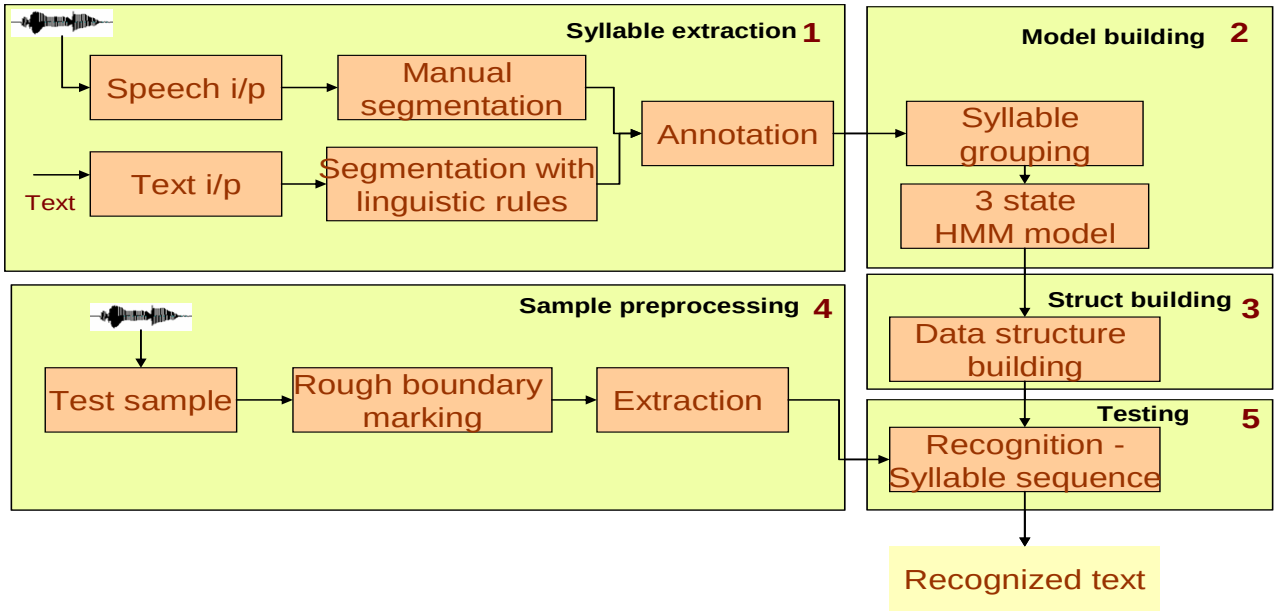

Fig. 4 Block diagram of the proposed system

3.1.1 Syllable extraction form speech

Speech corpus is generated by many institutes for various Indian languages. CEERI, Delhi developed for Hindi and Bengali at Syllable and phonemes level. *Samudravijaya K et.al [20]*. TIRF Mumbai focused on developing database for Hindi, Bengali, Marathi and Indian English (*Samudravijaya K et.al[21]*. IIT Chennai was the first institute which contributed database that was at syllable level which is prepared by an automatic segmenter. *T.Nagarajan, Hema et.al[22]*.

We propose a semi-automatic procedure for segmenting the speech signal into syllable units. This system is semi-automatic as it uses manual procedure for marking the boundaries of syllable and Praat scripts to label and store in folder for next processing. We open the speech file in Praat and mark the valley points on the first tire as shown in the Fig 5. The portion of the signal between two markings with intensity more than 50db(estimated noise level) is corresponding to one syllable. These portions are extracted and stored with temporary names as s_1, s_2, ... s_n in separate directory using Praat scripts.

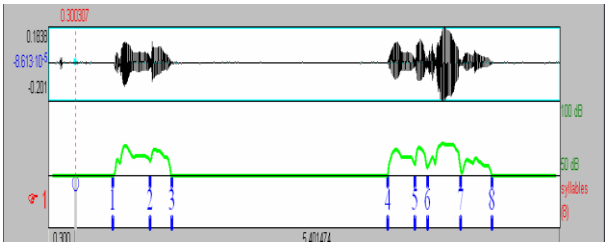

Fig.5. Marking syllable boundaries based on intensity information

3.1.2 Syllable extraction form text

The text segmentation is based on the linguistic rules derived from the language. Any syllable based language can be syllabified using these generic rules. To make the text segments exactly equivalent to the speech units the following algorithm is applied.

1. Read from the file which has text in WX notation.
2. Label the characters as consonants and vowels using the following rules
   - Any consonant except(*y, H, M*) followed by *y* is a single consonant, label it as C
   - Any consonant except (*y, r, l, lY, lYY*) followed by *r* is taken as single consonant
   - Consonants like(*k, c, t, w, p, g, j, d, x, b, m, R, S, s*) followed by *l* is taken as single consonant.
   - Consonant like (*k, c, t, w, p, g, j, d, x, b, R, S, s, r*) followed by *v* is taken as a single consonant.
   - Label the remaining as Vowel (V) or Consonant(C) depending on the set to which it belongs.
   - Store the attribute of the word in terms of (C*VC*)* in temp2 file.

3. For each word in the corpus get its label attribute from temp2 file.
   - If the first character is a C then the associate it to the nearest Vowel on the right.
   - If the last character is a C then associate it to the nearest Vowel on the left.
   - If sequences correspond to VV then break is as V-V.
   - Else If sequence correspond to VCV then break it as V-CV.
   - Else If sequence correspond to VCCV then break it as VC-CV.
   - Else If sequence correspond to VCCCV then break it as VC-CCV.
   - The strings separated by – are identified as syllable units.
4. Repeat 3 until end of file.
5. Store the result in output file.

The following Table 1: shows the output obtained for the input in Telugu text in UNICODE.

Table 1. OUTPUT FOR ALGORITHM

S . N o	Input word	Output of Algorithm
1	kaMpeVnIkaMteV	kaM-peV-nI-kaM-teV
2	KarcukaMteV	Kar-cu-kaM-teV
3	lABAlaku	lA-BA-la-ku

The syllable units extracted from the speech are labeled with the corresponding syllable name extracted from text. Once all the syllable units are labeled then a three state model is build in the second phase.

3.2 THREE STATE MODEL FOR SYLLABLE

Once the syllable units are extracted, group similar syllable units to build a three – state model. The model is defined with three states, where each state corresponds to segment of frames with small variance. All segments are not of uniform size since some segments are naturally longer than others.

E.g., Syllables with only vowel, may have few frames in the onset and coda segment. This difference in segment lengths is different from the variation within a segment. Segments with small variance could still persist very long for a particular sound or syllable.

To build the model for each syllable we divided the feature vectors into three segments where each segment represents the averaged model of features in that state. Initially divide the training sequence into equal segments and compute the average of each segment and adjust the segments iteratively as defined below.

- Divide the sequence vectors into three equal segments.
- Compute the average model.
- Align each template to the averaged model to get new segmentations
- Re-compute the average model from new segments with varied number of frames.
- The procedure can be continued until convergence.

Convergence is achieved when the total best-alignment error for all training sequences does not change significantly with further refinement of the model.

The following Fig 6 shows three samples for the syllable *rA* represented as *rA1, rA2, rA3*. Initially each sample is segmented into more or less equal segments. Average model is computed and segmented boundaries are adjusted. Finally all samples are represented with three states where first state corresponds to onset *r*, second part corresponds to nucleus *A* and third part corresponds to coda *ε*. Each state has a probabilistic function that describes the sound produced when in that state. Thus, the state labeled onset, nucleus and coda would have a very high probability associated with feature vectors for the sound corresponding to consonant *r*, vowel *A* and coda as *ε* as there is no succeeding consonant.

These segments are adjusted by aligning the average model with each of the sample as shown in the Fig 7. This process is repeated until the number of frames in each segment remains same.

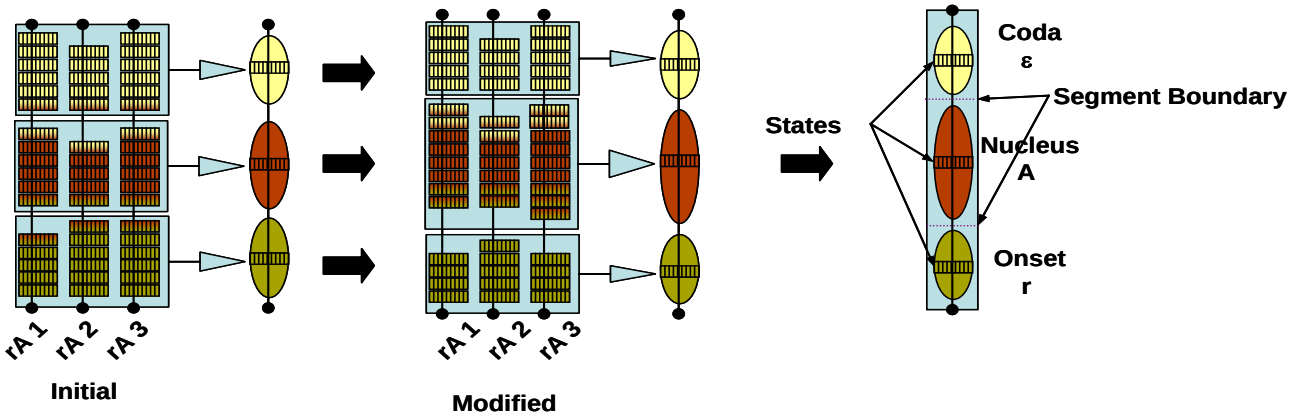

Fig 6: Building Tri state syllable model.

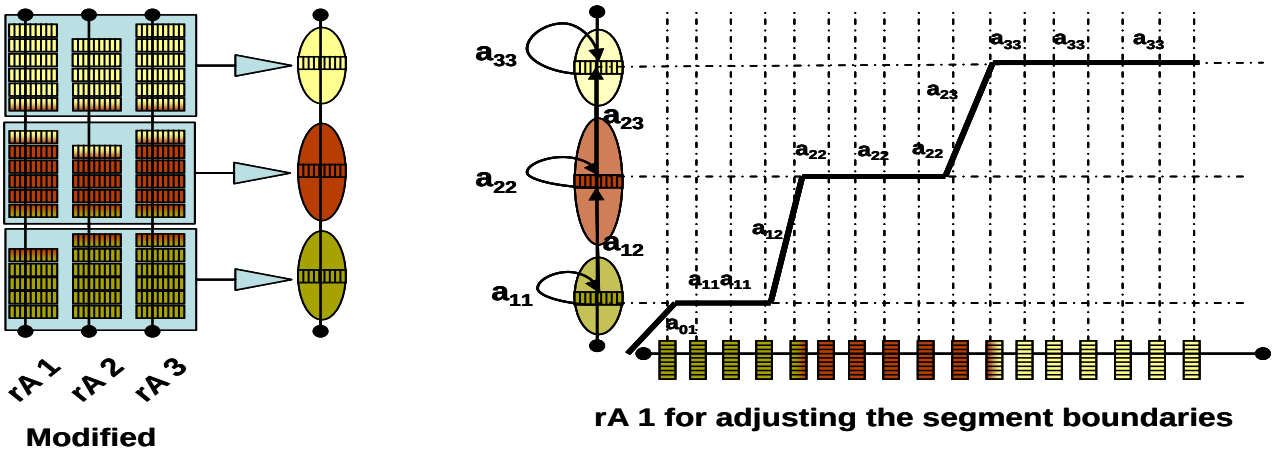

Fig 7: Alignment of Input vector with three state model using feature vector of Input vector and model parameters

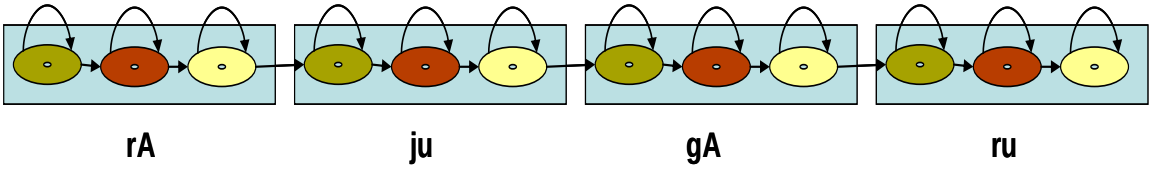

Fig 8: Word model for the word *rAjugAru* using syllable units

3.2.1 Word Models with Syllables

The word is a sequence of syllables, uttered in continuous speech. Since there is no need to surround the syllable with silence we start with the beginning of next syllable when it ends in last state of previous syllable. The Fig 8 illustrates the word model for "***rAjugAru***" which has four syllables. Each syllable may be succeeded by many possible syllables. It is there fore essential to represent all the words with a suitable data structure that would help during recognition process. We proposed a Trie data structure for representing the data.

3.3 TRIE STRUCTURE BUILDING USING MORPH KNOWLEDGE

Data representation is an important task for recognition. We propose a method in which the words are first represented as sequence of syllables. HMM models are built for each syllable and all the words are represented as a Trie structure.

A **trie**, or **prefix tree**, is an ordered tree data structure that is used to store an associative array where the keys are usually strings with related information. Unlike a binary search tree, no node in the Trie stores the key associated with that node and other information. All the descendants of a node have a common prefix of the string associated with that node, and the root is associated with the empty string. The node structure is defined to store the information relating to the name of the node, link to the model for the syllable, link to the wave file, link to the sibling node, link to the descendent node and bit field to indicate the end of the file.

Trie Structure contains 6 fields. First field is Name of syllable of string data type which contains name of syllable. Second field is model of string array data type which stores file name which has the three state model of the syllable. Third field is string type to store the file name which has the wave file it. Fourth field is Bit field of Integer which indicates the end of word by storing 1 else by 0. Fifth field is for the sibling node which has the common prefix as that of the current syllable. Sixth field is link to the child node that corresponds to possible suffix for the partial word formed by concatenating the syllables from root to the current node.

Root node has null set in the fields of syl_name, link_model, link_wave and word_end, to satisfy the properties of trie structure. The main advantage with trie structure is when the sample syllable is close to one of the syllable in the trie structure then the succeeding comparisons are made with the possible syllable in the sequence.

Structure node

```
{
   char    syl-name[10];
           char      link_model[20];
           char      link_wave[20];
           int       word_end;
           node      *link_sib_pointer;
           node      *link_down_ptr;
}
```

Let us consider the words *reVMdu, reVMdusArlu, reVMduvela, reVMdurakAlu, reVMduvaMxala, reVMduvEpula and reVMdurojulu.* It is clear that the all these words have common prefix *reVMdu* which has two syllables. We place first syllable *reVM* in first level and second syllable *du* in the second level as child node for *reVM*. The node that has syllable *du* in second level has four child nodes each containing the syllables *sAr, ra, ve, vaM*. The fourth level has

lu under *sAr*, *kA* under *ra*, *la* under *ve*, *xa* under *vaM*. In fifth level two nodes are inserted *lu* under *kA* and *la* under *xa*.

The structure formed is a standard trie and shown in the Fig 9. The advantage of this structure is, for example to search for the word *Gadiya*(*Ga-di-ya*) first the syllable sample of *Ga* is compared with the syllables models in the first level, i.e *reVM, vac, A, Ga, Pre*. The distance is measured with each of the model and uses the decedents of the syllable that gives less distance. The next comparison is done with the child nodes in the next level of the closest syllable. If *Ga* has minimum distance then the next syllable is compared with the model corresponding to *di*. Third syllable unit is compared with the models corresponding to *ci* and *ya*.

If the recognition task is done by comparing the words, it requires the comparison of test sample with all 300 words and if recognition is wrong we cannot correct the word. If this Trie data structure is used the number of comparisons are less and depends on the nodes that lie in each level and is also useful to correct the recognized word if it is partially recognized.

3.4 SAMPLE PROCESSING

The test sample is collected using Praat tool. This sample is preprocessed to eliminate the silence preceding and succeeding the utterance. This filtering is done by estimating the noise for the first 5 – 7 frames and a threshold is fixed for the noise. Using these threshold values word boundaries is marked. Now the intensity couture is used to mark the rough boundary of each syllable. These syllables are extracted and named sequentially and are recognized by the fifth phase.

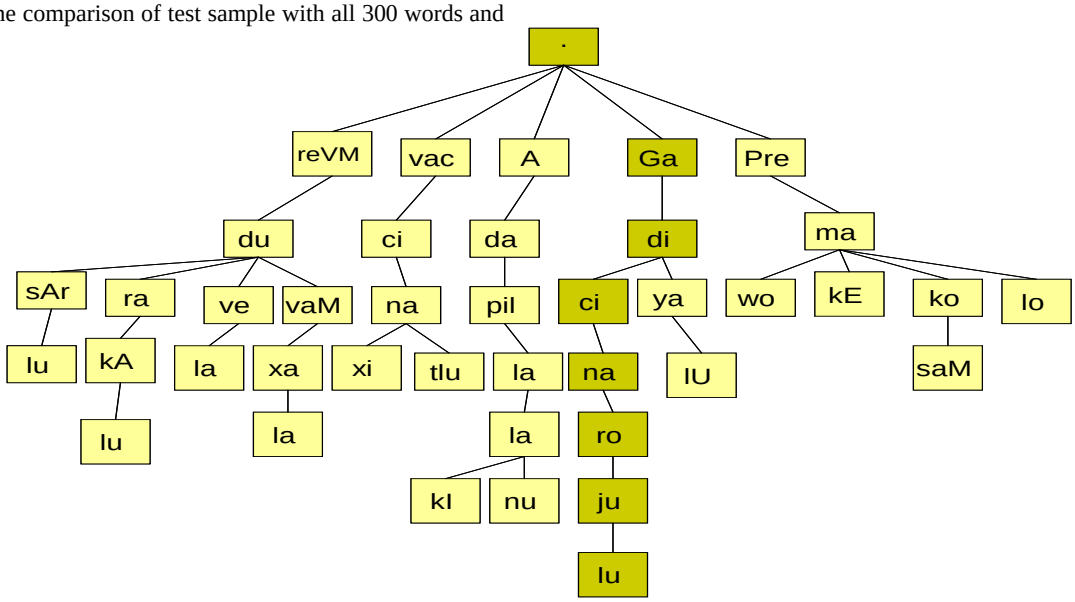

Fig. 9 Trie structure for the set of words in the language.

3.5 SYLLABLE RECOGNITION

Testing phase is to recognize the sequence of syllables and to identify the word. Each syllable is compared with the syllables in the first level. The distance is measured using Mahalanobis distance measure which considers the inherent variation between vectors of different segments. We defined the covariance for each segment using the standard formula for covariance. The close syllable is selected and the decendent syllables in next level are compared next. The following Fig 10 shows the sample of *lu* compared with the three models of *rA*, *ju* and *lu.*

3.6 TEST DATA

The following Table II is few set of words considered for experimental work represented in syllabified form. From these 30 sample words it is clear that most of the words have few syllables in common. These words could be covered with 48 distinct syllables. Training is done with 30 words by building three state models for 48 distinct syllables we could recognize the words that were not used for training. Table 3 has 16 words that were nor used for training but were recognized by our system. We can even test with the other words which is formed with these 48 syllables

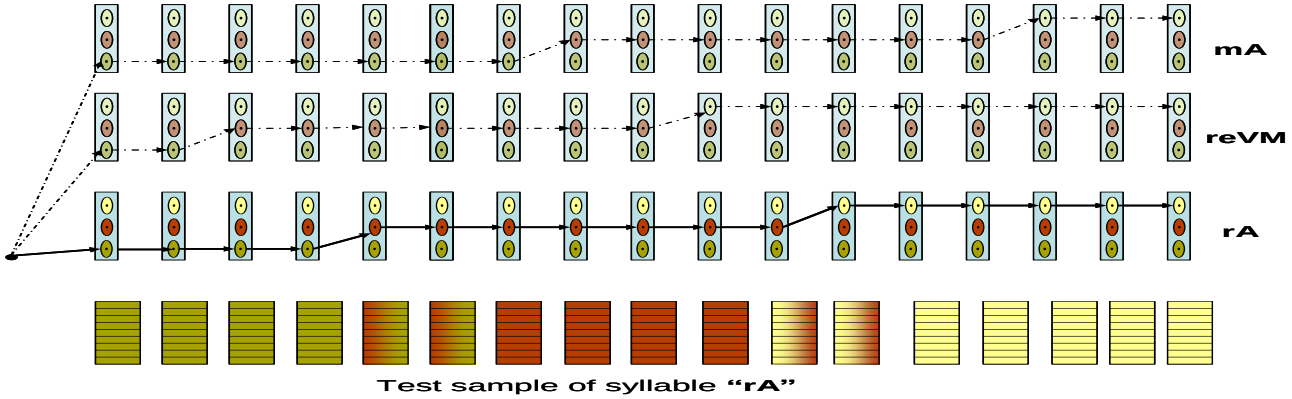

Fig 10: Recognition of given syllable "rA"

Table 2. SET OF SELECTED WORDS

Word	Syllable sequence	Distinct syllable
reVMdu	reVM-du	reVM, du
reVMdusArlu	reVM-du-sAr-lu	sAr, lu
reVMduvela	reVM-du-ve-la	ve, la
reVMdurakAlu	reVM-du-ra-kA-lu	ra, kA
reVMduvaMxala	reVM-du-vaM-xa-la	vaM, xa ,la
reVMduvEpula	reVM-du-vE-pu-la	vE, pu
reVMdurojulu	reVM-du-ro-ju-lu	ro, ju
vaccine	vac-ci-na	vac,ci, na
vaccinaxi	vac-ci-na-xi	xi
vaccinatlu	vac-ci-na-tlu	tlu
vacciMxo	vac-ciM-xo	ciM, xo
vaccinayi	vac-ci-na-yi	yi
vaccinavi	vac-ci-na-vi	vi
vaccinMxa	vac-ci-nM-xa	nM
vacciMxani	vac-ciM-xa-ni	ni
Adapilla	A-da-pil-la	A, da, pil
Adapillalanu	A-da-pil-la-la-nu	nu
AdapillalakI	A-da-pil-la-la-kI	kI
AdinatlugA	A-di-na-tlu-gA	gA
Adinapudu	A-di-na-pu-du	pu
Gadiya	Ga-di-ya	Ga, di, ya
GadiyalU	Ga-di-ya-lU	lU
Prema	Pre-ma	Pre, ma
Premawo	Pre-ma-wo	wo
PremakE	Pre-ma-kE	kE
PremakosaM	Pre-ma-ko-saM	saM
Premalo	Pre-ma-lo	lo
aBiruci	a-Bi-ru-ci	a, Bi,ru
Manamu	ma-na-mu	mu
Aneka	a-ne-ka	ne

Table 3. NEW WORDS GENERATED

New Word	Syllable sequence
vaccinapudu	vac-ci-na-pu-du
vaccinMxa	vac-ciM-xa-ni
vaccinavAru	vac-ci-na- vA-ru
vaccinatlugA	vac-ci-na-tlu-gA
AgAdu	A-gA-du
Gadiyalalo	Ga-di-ya-la-lo
Gadupunu	Ga-du-pu-nu
Gadicinaroju	Ga-di-ci-na-ro-ju
Gadicinarojulu	Ga-di-ci-na-ro-ju-lu
Adinavi	A-di-na-vi
mana	ma-na
manawo	ma-na-wo
manakE	ma-na-kE
anekasArlu	a-ne-ka-sAr-lu
PremagA	Pre-ma-gA
aBirucilawo	a-Bi-ru-ci-la-wo

Once the syllable sequence is identified concatenate the sequence to form the word. From the Table III it is clear that if the morph knowledge is used then we can recognize large vocabulary with limited training. If there are new words to be recognized then first enhance the Trie with the insertion of new syllable sequence including the existing syllable models. If the word has syllables that are not available then collect the sample and build the model and add it to the Trie structure.

4. CONCLUSIONS

First the analysis of the Telugu corpus was performed and limited words were selected for the experimental work. Based on this analysis words were identified which are formed with high frequent syllables. Morphological information is used to construct the Trie structure which places all the words with common prefix under same path.

Our proposed system can recognize large vocabulary of isolated words with small training. The experiments were conducted to recognize the words that are not used in training the system. To recognize a new word, we need not collect multiple samples of the word and build the model for the same. With the proposed system if the syllable models already exist then the sequence is just added to the structure. This takes care of building the path for syllable sequence of the word. For example the word Gadicinarojulu (Ga-di-ci-na-ro-ju-lu) is new word not used in training. If this word is to be recognized, we first check if all the syllables are there in model form. Since this is the word with existing syllables we add the sequence in the structure as shown in the Fig 9. When the word is to be recognized there would exist a path for this new word. Similar tests were conducted on many new words that are not used in training.

This system is also useful in continuous speech provided if there is a preprocessing unit which would identify the boundaries of syllables and words. The performance of the system is remarkable even for large vocabulary

5. REFERENCES

[1] Plauche. M, Udhyakumar.N, Wooters.C, Pal.J, & Ramachandran. D (2006). *Speech recognition for illiterate access to information and technology*. In proceedings of first international conference on ICT and development.

[2] Lippmann, R.P Martin, E.A &Paul, D.P.(1987). *Multi-style training for robust isolated-word speech recognition.* In Proc. IEEE international conference on acoustics, speech, signal processin(pp. 705-708).

[3] Rabiner, L.R, Wilpon, J.G., & Soong, F.K.(1988). *High performance connected digit recognition using hidden Markov models.* Presented at the IEEE international conference. Acoustics, speech, signal processing.

[4] Huang, X., Acero, A., & Hon, H. W. (2001). *Spoken language processing – a guide to theory, algorithm and system development.* Englewood Cliffs: Prentice – Hall PTR. ISBN:0-13-022616-5.

[5] Fujimura.O (1975). *Syllable as a unit of speech recognition.* IEEE transactions on Acoustics, Speech and Signal Processing, ASSP-23(1),82-87.

[6] Ganapathiraju.A, Hamaker.J, Picone.J, Ordowski.M & doddington.G.R(2001). *Syllable based large vocabulary continuous speech recognition.* IEEE Transactions on Speech and Audio Processing. 9(4), 358-366.

[7] Nagarajan.T, Kamakshi Prasad .V & Hema.A.M (2001). *The minimum phase signal derived from the magnitude spectrum and its application to speech segmentation.* In Sixth biennial conference of signal processing and communications.

[8] Nagarajan .T, Hema . A.M & Hegde .R.M(2003). *Segmentation speech into syllable – like units.* In EUROSPEECH – 2003 (pp.2893-2896).

[9] Sunitha .K.V.N & Kalyani.N (2009). *Syllable analysis to build a dictation system in Telugu language.* International Journal of Computer Science and Information Security. Vol 30. No 30.

[10] Daniel Jurafsky and Patrick Schone. 2000. *Knowledge free induction of morphology using latent semantic analysis.* In Proceedings of CoNLL-2000 and LLL-2000, pages 67–72, Lisbon.

[11] Dejean, H. 1998. *Morphemes as necessary concept for structures discovery from untagged corpora.* Workshop on Paradigms and Grounding in Natural Language Learning. Adelaide, Australia. 295–299.

[12] S. Keshava and E. Pitler. 2006. *A simpler, intuitive approach to morpheme induction.* In Proceedings of 2nd Pascal Challenges Workshop, pages 31–35, Venice, Italy.

[13] Creutz,M. 2003. *Unsupervised segmentation of words using prior distributions of morph length and frequency.* In Proceedings of the Association for Computations Languages (ACL'03). Sapporo, Japan. 280–287.

[14] Brent, Michael R., Sreerama K. Murthy, and Andrew Lundberg. "*Discovering Morphemic Suffixes: A Case Study in MDL Induction."* The Fifth International Workshop on Artificial Intelligence and Statistics. Fort Lauderdale, Florida, 1995.

[15] Goldsmith, J.A. (2001). *Unsupervised Learning of the Morphology of a Natural Language.* Computational Linguistics, 27:2 pp. 153-198.

[16] Sunitha .K.V.N & Kalyani.N (2009). *A Novel approach to improve rule based Telugu Morphological Analyzer*. CISIM 978-1-4244-5612-3/09/$26.00_c 2009 IEEE.

[17] Sunitha .K.V.N & Kalyani.N (2009). *Improving the word coverage by using Unsupervised Morphological Analyzer.* Sadhana-Academy Proceedings in Engineering Sciences ISSN0256-2499 (Print) 0973-7677 (Online), Springer India, in co-publication with Indian Academy of Sciences.

[18] Sunitha .K.V.N & Kalyani.N. *YAST – Yet Another Statistical Trimmer.* International Journal of Computer applications in Engineering Technology and Sciences. IJ-CA-ETS, ISSN:0974-3596, Oct 08.

[19] Sunitha .K.V.N & Kalyani.N. *Unsupervised stemmer to improve rule based morph analyzer.* International Journal of Computer Information Systems and Industrial Management Applications ISSN:2150-7988 Vol.2(2010), May 10.

[20] Samudravijaya K, P.V.S.Rao & S. Agrawal. *Hindi Speech Database.* Proceedings of International Conference on Spoken Language Processing(ICSLP 00), Beijing, China Oct 2000.

[21] .Samudravijaya K, K.D.Rawat, and P.V.S.Rao, *Design of Phonetically Rich Sentences for Hindi Speech Database*, J. Ac. Soc. Ind. vol. XXVI, December 1998, pp. 466-471.

[22] T.Nagarajan, Hema A. Murthy and Rajesh M. Hegde. *Segmentation of speech into syllable-like units.* EUROSPEECH 2003 – GENEVA, page 2893-2896.