Volume 3, Issue 10, October 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Minimum Data Set Based on Syllable Position for Telugu Speech Systems

Dr. N.Kalyani *
Professor, CSE Dept, GNITS, Hyderabad, AP, India

Dr K.V.N.Sunitha
Principal, BVRIT for Women, Bachupally, Hyderabad,AP,India

Mr. N. Sreekanth
Assoc Prof, Vignan college of Engg Nalgonda, Hyderabad, AP, India

Abstract— Building a speech recognition system for any language requires large training data to improve the efficiency of recognition. Collecting huge training data for every language is trivial task. For the languages that are syllabic centred, problem can be solved by proper analysis of text data and arriving at minimum words that would cover all possible syllables. These words are collected as training data and can be used to build robust HMM models. This paper highlights the importance of a syllable unit and the results relating to the statistical analysis carried on the CIIL text corpus for Telugu language. The results obtained are used to analyse the possible position of occurrence of each syllable in the words. This analysis is used in generating minimum training data set that covers maximum vocabulary. The minimum data set is generated by including high frequent syllables based on the position of occurrence in the words.

Keywords—Syllable Structure, Position of Occurrence, Minimum data set, Word Coverage and High frequent syllables.

I. INTRODUCTION

Automatic Speech Recognition is processing 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, which makes the process of building database a difficult task as the variations in the formation of sentence are enormous. The next possible unit is a word, it has well defined acoustic representation and can represent all contextual effects within it. It is best for limited vocabulary ASR applications **(Lippmann et al. 1987; Rabiner et al. 1988).** The disadvantage with this unit representation is, the training process is done with the individual word which can appear in any context **(Huang et al. 2001)** and the actual words occur in continuous form in natural speech. Second limitation is as the training set increases there is slight decrease in the recognition accuracy. The third limitation is increase in memory requirements and increase in process time. There are applications that were developed which have limited vocabulary like digits **(Plauche et al. 2006).** The next choice of sub-word units are phones. Since the number of phones is more or less 50 in most of the languages it becomes easy to build phone models. The advantage in such unit selection is the reasonable size of training corpus. The limitation with phone units is that the same phone in different words would have different pronunciations. This is because the articulators that produce the sound unit are affected by the previous and succeeding phones. The variations in pronunciations are strongly affected by adjacent phones.

There is reported work which showed that word based DTW performs significantly well when the vocabulary is small. **(Bahl et al. 1988; Paul and Martin 1988).** There are also related works that showed that tri-phone models are powerful as it accounts for the left and right context. The work carried out **by (Bahl et al. 1980; Schwartz et al. 1984)** reported that word error rate was reduced by more than 50% as compared with word models and phone models. The disadvantage was building the training set which has to incorporate large number of triphones. Syllable is next higher level to phone and low compared to word unit which is found to be a promising unit for recognition. The importance of syllable was first reported by **Fujimura (1975).** The first successful robust LVCSR system was developed by **Ganapathiraju et al(2001)** and he used syllable level acoustic unit in telephone bandwidth spontaneous speech. There are many papers published by **Nagarajan et. al.(2001, 2003).** 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. This work proposes the basic structure of a syllable in text data and the intensity details of variation in speech signal in section 2. The procedure to identify the syllable unit in text and speech utterance, with a suitable example in Telugu is elaborated in section 3. Section 4 gives the details of the corpus selected and the details relating to the syllable analysis. The selection of minimum data set using the syllable position is presented in section 5, followed by conclusions in section 6.

II. SYLLABLE STRUCTURE

Structurally, a syllable consists of three parts, an onset, a nucleus and a coda. The nucleus and coda together is called rhyme. In general the syllable consists of sequence of phone units which is of the form C*VC*. The Figure 1. a shows three components of the syllable.

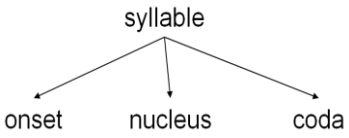

Fig. 1 Structure of Syllable

An example structure of word ***reVMdurakAlu*** **(రెండురకాలు)** comprising five syllable units is shown in the Figure 2.a and Figure 2.b shows the variations of intensity contour of syllable units in the word. 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.

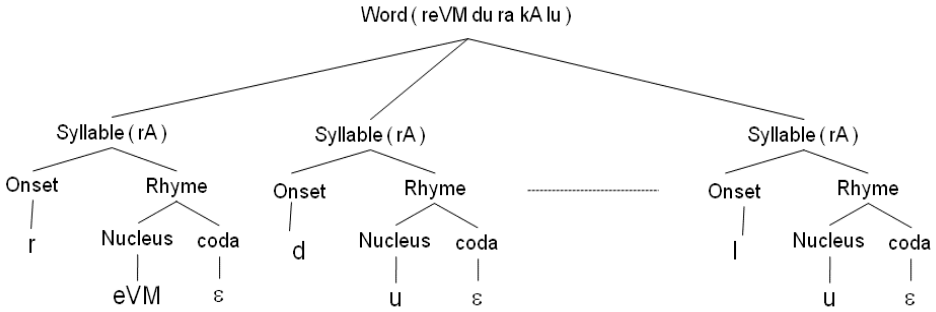

Fig. 2.a Detailed structure of a word ***reVMdurakAlu*** (reVM-du-ra-kA-lu)

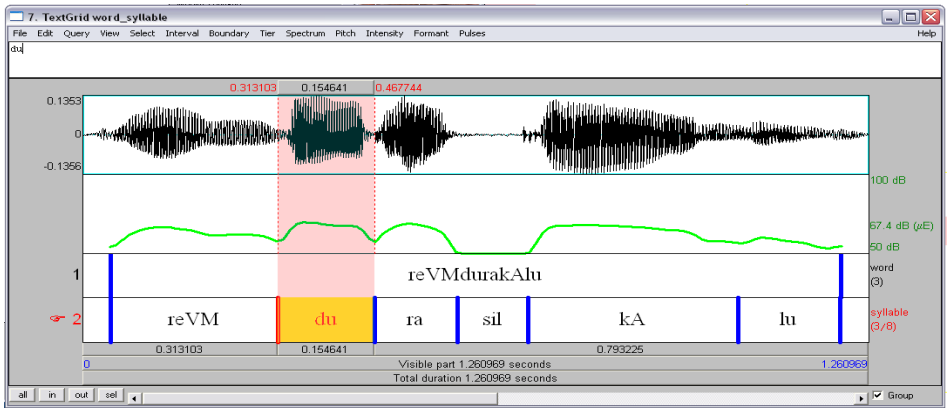

Fig. 2.b Intensity contour showing the syllable variation for the word ***reVMdurakAlu*** (reVM-du-ra-kA-lu)

The structure of the syllable is as shown in the Figure 2.a 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 word ***vixya*** (విద్య - education) there are two syllables. The syllable ***vix*** has two consonants , one vowel and syllable ***ya*** has one consonant in coda & one vowel. In word ***ara*** (అర - 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. This variation is shown in the Figure 2.b.

III. SYLLABLE IDENTIFICATION

The work presented in this paper is carried out on CIIL Telugu text corpus of size 3 GB. This corpus has the raw text data relating to different domains segregated in 776 files. The given raw text corpus is pre-processed to eliminate the extraneous characters and list of unique words are listed. The pre-processing is to remove header information, punctuation marks and special symbols. After pre-processing it has 3.7 million words of which 8.2 lakh words are unique. From the words, the syllables units are extracted using linguistic rules of Telugu language. These syllable units are analysed for extracting the information relating to frequency of occurrence of each syllable and the frequency relating to the position of occurrence in the words. Using this information the minimum list of words is generated that covers high frequent syllables in desired position. The algorithm for identifying the syllable boundaries in the text is given in Algorithm 3.1.

A. *Algorithm 3.1*

This algorithm breaks the words into sequence of syllables. The input is words in WX notation and the output is words broken into syllables.

- Read raw text from the file in WX notation and pre-process to eliminate extraneous characters.
- Mark each phoneme as V for vowel and C for consonant using the following rules.
  - All consonants with succeeding element y are marked as single unit and named as consonant C, except for the elements y, H and M.
  - Consonants with succeeding element r are marked as single unit and is named as consonant C, except for the elements y, r, l, lY and lYY
  - Few consonants like k, c, t, w, p, g, j, d, x, b, m, R, S and s with succeeding element as l is marked as a single consonant.
  - Few consonant like k, c, t, w, p, g, j, d, x, b, R, S, s and r with succeeding element as v is marked as a single consonant.
  - Mark the remaining symbols as either vowels or consonants depending on the class to which it belongs.
  - Store the labels assigned for each phonemes in the word to a separate file and name it as temp2 file.
- Read the file temp2 which has the label sequence for each word in the text file to identify the syllable boundary.
  - First check if the word is beginning with a consonant then associate it with the nearest vowel to its right.
  - Check if the word is ending with a consonant then associate it to the nearest vowel to its left.
  - In the sequence if there are two consecutive vowels then break it and place a hyphen between as V-V.
  - Check for the sequence VCV if found, then break it and place hyphen between V and C to make it as V-CV.
  - Check for the sequence VCCV if found, then break it and place hyphen between C and C to make it as VC-CV.
  - Check for the sequence VCCCV if found, then break it and place hyphen between C and C to make it as VC-CCV.
  - The strings that appear between two – are identified as syllable units.
- Repeat the steps and identify the syllable units for the entire file.
- Store the result is separate file.

B. *Example :*

When the word వాసంతి (vAsaMwi) is syllabified it takes the following forms

Word	vAsaMwi
Phone sequence	v-A-s-a-M-w-i
Applying rules of (yrlvM)	v-A-s-aM-w-i
Labels of each phone	C-V-C-V-C-V
After syllabifying	CV-CV-CV
Syllable sequence	vA-saM-wi (వా-సం-తి)

IV. STATISTICAL ANALYSIS OF SYLLABLES

A. *Analysis on frequency of occurrence*

Telugu being a syllabic centred language, the main focus was on analysing the frequency of occurrence for each syllable and the words having high frequency syllables. The prime motivation for this analysis is, the syllable boundary in speech signal is approximately identified by the signal properties. The second motivation was to develop a speech recognition system at syllable level that can be built with minimum training data. On analysis it was found that there were 12,378 distinct syllables. Nearly 337 syllables had the highest frequency more than 1000. These syllables are of primary interest as building models for them helps in recognizing more possible words. More than 10000 syllables have very less frequency even less than 100. It implies these most of the syllables are less frequently used in the vocabulary. It is also observed that there are syllables that had the frequency count as one, which implies that these syllables occurred only once in the Telugu corpus.

It is clear that these syllables occurred corresponding to the words like Apple, coffee, strength etc. These words are lone words and when written in Telugu it normally takes the same pronunciation. Syllables formed from loan words are different and do not occur in native language. The results of the analysis on syllable frequency are shown in Figure 3. The plot in Figure 3. shows the details with syllable frequency in Thousands. From the figure it is clear that there are around 337 syllables whose frequency is more than 1000 and there are 71 syllables with frequency more than 10,000.

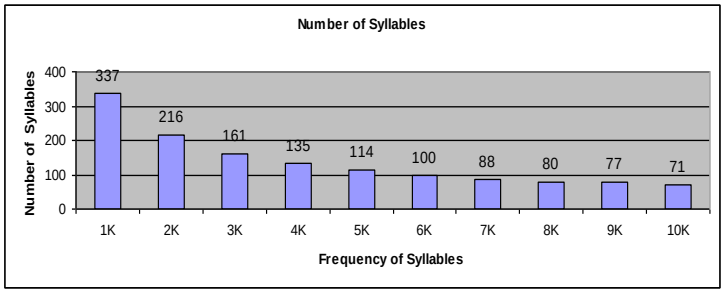

Fig. 3 Number of Syllables in frequency range 1000 to 10000.

From the above figure 50% syllables out of 337 are with frequency of 1K to 3K. Only 114 syllables have the frequency more than 5000. It is observed that there are nearly 71 syllables that have frequency more than 10K. Only one syllable has the highest frequency and that syllable is "na".

The analysis is also performed to identify the words that is formed with syllables(100%) with cut-off frequency. Similar analysis is carried to identify the words with 80% and 50% syllables having required cut-off frequency. The results plotted in the Figure 4 are relating to frequency range of ten thousands. The colours used in these charts indicate the percentage of syllables with the frequency more than the specified value.

- Blue indicates that in the word there are at least 50% of syllables that have the required frequency.
- Pink indicates that in the word there are at least 80% of syllables that have the required frequency.
- Yellow indicates that all the syllables in the word have the required frequency.

The motivation for this analysis is to identify the words that are formed with 100% of syllables with required cut-off frequency. This analysis also indicates if the words are chosen with 80% or 50% of syllables to have required frequency what would be the coverage of words that can be recognized.

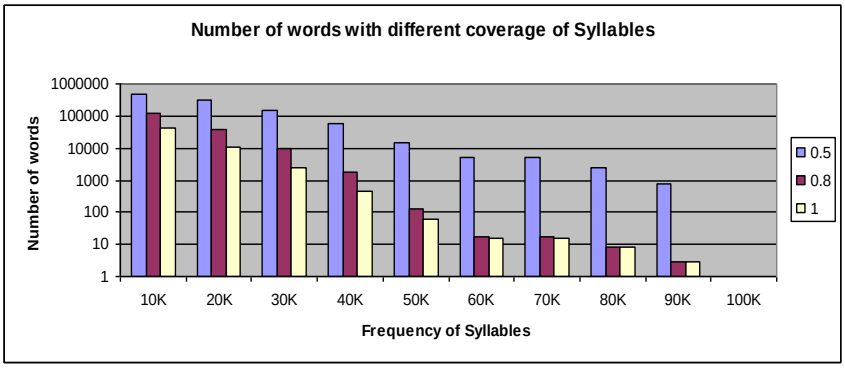

Fig. 4: Number of words with 50%, 80% and 100% syllables in cut-off frequency range 10000 to 100000.

Bar graph with words having syllables at least 50%,80% and 100% with frequency of syllables varying in the range from 10K to 100K is shown in Figure 4. It is observed from the above graph, the number of words formed with syllables of frequency more than 90K there are very few valid words are formed. If the word is considered with all high frequent syllables with the given frequency cut off value, it indicates it is most important word.

B. : Analysis on position of occurrence

When the cut-off range is set to 10,000 there are 71 syllables and the details are listed in the table shown in the appendix. When a thorough analysis is made relating to the position of occurrence and the frequency of occurrence the observations are listed in Table 1 to Table 3.

There are only **5** syllables that appear in the initial position with more than 50% of its frequency. These syllables are listed in Table 1.

TABLE I
MOST FREQUENT SYLLABLES OCCURRING IN INITIAL POSITION

S.No	Syllable	Total Frequency	Initial Position %	Middle Position %	Final Position %
1	a	24679	97.02%	2.92%	0.05%
2	A	13728	96.79%	2.92%	0.26%
3	pra	15301	79.46%	20.46%	0.05%
4	saM	11351	61.98%	26.85%	11.14%
5	vi	27273	50.69%	31.38%	17.71%

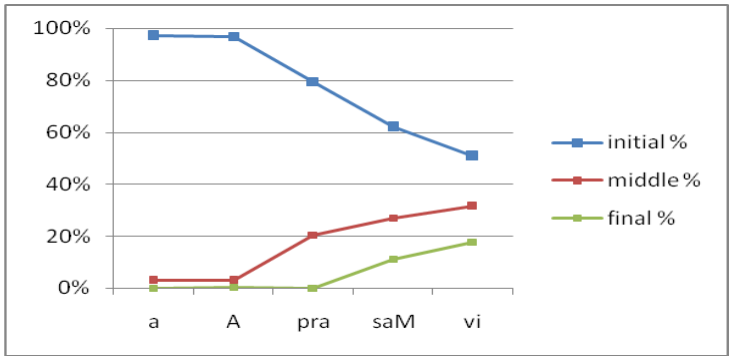

Fig. 5: Analysis of syllables occurring in the initial position

From the graph in Figure 5. it is clear that the syllables ***a, A*** and ***pra*** are more prominent in the initial position. In the data set while selecting them it is preferred in initial position where as the other two ***saM*** and ***vi*** are more desirable in initial position and then in the middle position and less in final position.

There are **54** syllables that appear in the middle position with more than 50% of its frequency. These are listed in the Table 2.

TABLE II
MOST FREQUENT SYLLABLES OCCURRING IN MIDDLE POSITION

S.No	Syllable	Total Frequency	Initial Position %	Middle Position %	Final Position %
1	Ra	10301	1.91%	91.11%	6.38%
2	wu	13419	4.64%	88.40%	6.71%
3	Na	12967	0.23%	85.04%	13.98%
4	da	26965	0.99%	84.54%	12.46%
5	ra	47087	7.25%	84.49%	6.56%
6	va	40112	10.05%	84.43%	3.69%
7	ya	28711	5.76%	83.95%	8.89%
8	xa	30269	5.06%	82.30%	11.57%
9	cu	19228	0.82%	81.26%	17.38%
10	ca	24863	15.26%	80.25%	3.48%
11	ri	25677	2.98%	78.69%	17.59%
12	yA	11442	8.44%	78.32%	12.36%
13	wA	17642	11.16%	77.55%	11.09%
14	po	14203	21.70%	77.33%	0.79%
15	la	93234	1.51%	76.98%	18.56%
16	ga	30049	10.80%	76.37%	11.51%
17	ta	25951	0.67%	74.73%	23.80%
18	ko	12565	18.04%	74.71%	7.04%
19	su	11345	15.21%	74.08%	10.58%
20	li	17384	2.70%	74.06%	22.87%
21	naM	10115	5.34%	73.02%	21.49%
22	tu	12264	0.29%	72.47%	26.48%
23	pu	20003	12.15%	72.21%	14.95%
24	vA	24554	23.14%	72.15%	4.34%
25	gu	15256	12.63%	71.93%	15.04%
26	le	14898	11.87%	71.61%	16.30%
27	vu	14610	4.44%	70.77%	24.51%
28	wa	43376	13.52%	70.44%	14.64%
29	wi	20201	8.76%	70.29%	20.25%
30	ma	39103	25.40%	68.13%	5.05%
31	ti	18063	0.72%	68.10%	30.45%
32	ba	13591	30.65%	67.32%	1.85%
33	mu	48246	5.97%	67.26%	26.20%
34	kA	25241	27.02%	67.23%	4.95%
35	di	22573	2.41%	64.99%	31.82%
36	pa	48505	30.69%	64.85%	3.09%
37	nA	36639	14.03%	64.59%	20.81%
38	xA	12230	17.74%	63.91%	18.11%
39	si	12202	6.67%	63.85%	29.01%
40	lA	12321	9.38%	63.50%	26.79%
41	ka	58311	20.14%	62.55%	15.91%
42	mA	18287	29.26%	62.42%	7.93%
43	pA	14610	37.32%	61.54%	0.92%
44	na	89112	5.25%	61.22%	31.34%
45	ja	12159	34.47%	60.90%	4.18%
46	rA	24437	25.84%	60.07%	13.32%
47	ci	21623	9.02%	59.75%	30.93%
48	dA	11106	1.71%	56.03%	42.14%
49	xu	21978	2.77%	54.45%	42.42%
50	ce	17572	30.66%	51.88%	17.39%
51	sA	11009	46.24%	51.62%	1.93%
52	ru	34476	1.56%	51.52%	46.32%
53	sa	24001	44.86%	50.94%	3.49%
54	yi	17037	3.40%	50.09%	45.69%

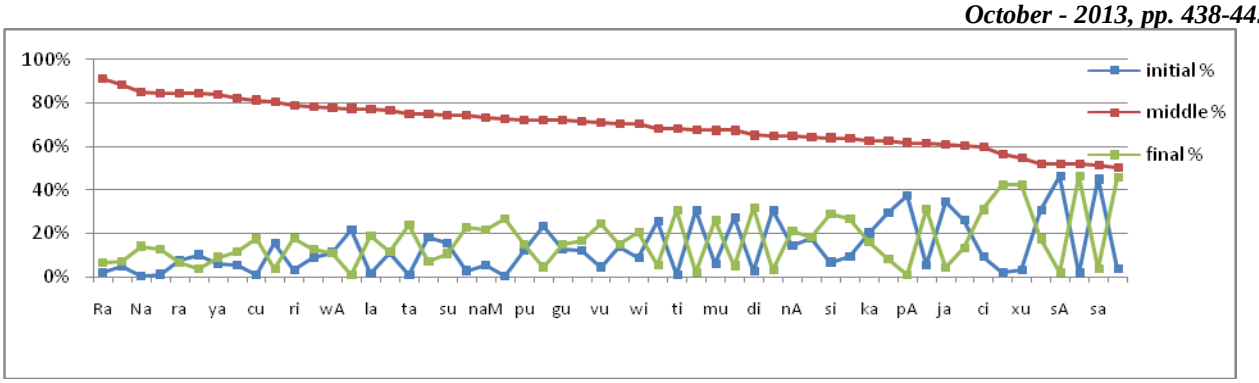

Fig. 5: Analysis of syllables occurring in the initial position

From the graph in Figure 6 it is clear that the there are only seven syllables ***dA, xu, ce, sA, ru, sa*** and ***yi*** are more prominent in middle position and also show prominence in either initial or final position. In the data set while selecting them it is first preferred in initial position and can also be included in other positions.

There are **10** syllables that appear in the final position with more than 50% of its frequency. These are listed in the Table 3.

TABLE III
MOST FREQUENT SYLLABLES OCCURRING IN FINAL POSITION

S.No	Syllable	Total Frequency	Initial Position %	Middle Position %	Final Position %
1	ki	20816	2.53%	20.13%	77.21%
2	xi	30894	3.72%	27.98%	68.03%
3	nu	50382	0.82%	31.30%	67.43%
4	wo	14213	6.30%	26.24%	67.42%
5	ne	15491	9.33%	26.95%	63.57%
6	nI	10479	16.84%	22.09%	61.04%
7	lu	42606	0.24%	39.88%	59.45%
8	du	32751	0.09%	43.18%	56.27%
9	lo	38247	4.28%	40.66%	54.93%
10	ni	72310	9.75%	36.99%	52.72%

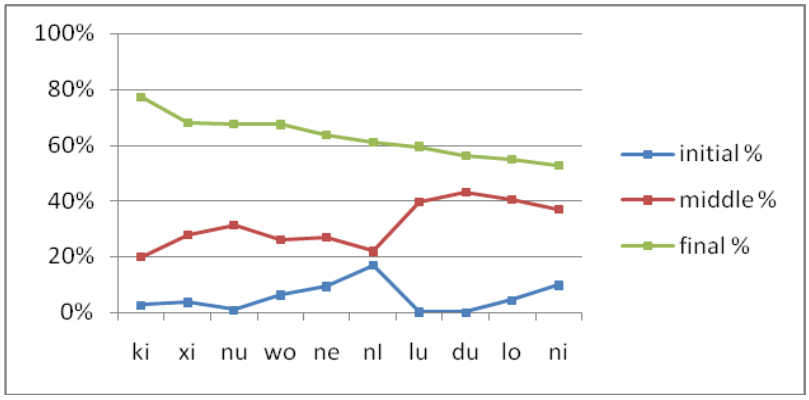

Fig. 7: Analysis of syllables occurring in the final position

From the graph in Figure 7 it is clear that the there are only four syllables ***lu, du, lo*** and ***ni*** are more prominent in final position and also show prominence in middle position. In the data set while selecting them it is first preferred in final position and can also be included in middle positions.

V. SELECTION OF MINIMUM DATA SET

The positional analysis shows that out of 71 syllables there are 69 syllables that have prominence in the position of its occurrence that is atleast 50% of the total frequency. There are only two syllables that can either take middle or final position. These syllables are ***ku*** and ***gA.*** Using this analysis the minimum number of words can be selected that would cover all possible syllables in the required positions are identified and listed in the Table 4.

TABLE IIV
MINIMUM DATA SET SELECTED BASED ON SYLLABLE POSITION

S No	Words	Syllable Sequence	Distinct Syllables list	Number of Distinct
1	vijayanagaramuwo	vi-ja-ya-na-ga-ra-mu-wo	vi, ja, ya, na, ra,	6
2	anaMwamayinaxi	a-naM-wa-ma-yi-na-xi	a, naM, wa, ma, yi,	6
3	adavikAparulu	a-da-vi-kA-pa-ru-lu	da, kA, pa, ru, lu	5
4	saMpraxAyAniki	saM-pra-xA-yA-ni-ki	saM, xA, yA, ki	4
5	prasaMgamunu	pra-saM-ga-mu-nu	pra, ga, mu, nu	4
6	AcaraNaloniki	A-ca-ra-Na-lo-ni-ki	A, ca, Na	3
7	Adiposukune	A-di-po-su-ku-ne	po , su , ku	3
8	kolAtamAduwAru	ko-lA-ta-mA-du-wA-ru	ko , lA, wA	3
9	jagadamAdarAxu	ja-ga-da-mA-da-rA-xu	mA , rA , xu	3
10	adigAdanI	a-di-gA-da-nI	di, gA, nI	3
11	viRayamamulo	vi-Ra-ya-ma-mu-lo	vi, Ra, lo	3
12	prakatanagAni	pra-ka-ta-na-gA-ni	ka, ta, ni	3
13	acirakAlamulone	a-ci-ra-kA-la-mu-lo-ne	ci, ne	2
14	adagavalasinavi	a-da-ga-va-la-si-na-vi	va , si	2
15	Jayapraxamagunu	ja-ya-pra-xa-ma-gu-nu	xa, gu	2
16	AdabaducugA	A-da-ba-du-cu-gA	ba , cu	2
17	bayativAriwo	ba-ya-ti-vA-ri-wo	ti , ri	2
18	adavulapAlu	a-da-vu-la-pA-lu	vu , pA	2
19	vAsaMwi	vA-saM-wi	vA , wi	2
20	kAlesinAru	kA-le-si-nA-ru	le , nA	2
21	sApucesi	sA-pu-ce-si	sA , pu	2
22	Apagaladu	A-pa-ga-la-du	la, du	2
23	acewanaM	a-ce-wa-naM	ce	1
24	adiyAsa	a-di-yA-sa	sa	1
25	rAjanIwulanu	rA-ja-nI-wu-la-nu	wu	1
26	rAlina	rA-li-na	li	1
27	adagadAniki	a-da-ga-dA-ni-ki	dA	1
			Total	**71**

VI. CONCLUSIONS AND FUTURE SCOPE

The analysis carried on the large text corpus is used to identify the position of occurrence of the syllables. This information is used in identifying the minimum data set that can represent the large data sets in speech corpus. It is observed that 71 syllables have frequency more than 10K. Out of these 5 syllables are more prominent in initial position, 54 syllables occur in middle position and 10 syllables are prominent in final position. There are only two syllables that occur in middle position and with equal frequency in either initial or final position. Using this analysis 27 words are derived that would cover 71 syllables in the possible place of occurrence. If the training data is collected for these words then effective models can be generated that will improve the performance of speech systems. Similar analysis can be carried with different cut-off frequencies that would include more number of syllables to increase the word coverage.

REFERENCES

[1] Lippmann, R P Martin, E A, Paul, D P "Multi-style training for robust isolated-word speech recognition" In Proceedings IEEE international conference on acoustics, speech, signal processing, 1987 PP 705-708.

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

[3] Huang X, Acero A, Hon H W, "Spoken language processing – a guide to theory, algorithm and system development." Prentice – Hall PTR ISBN:0-13-022616-5, 2001.

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

[5] Paul D B, Martin E A, "Speaker stress – resistant continuous speech recognition." Presented at IEEE international conference on Acoustics, speech, signal processing , 1988.

[6] Bahl L R, Brown P F, De Souza P V, Mercer R L, "Acoustic Markov models used in the Tangora speech recognition system." Presented at the IEEE international conference on Acoustics, speech, signal processing, 1988.

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

[8] Ganapathiraju A, Hamaker J, Picone J, Ordowski M, Doddington G R, "Syllable based large vocabulary continuous speech recognition." IEEE Transactions on Speech and Audio Processing Vol 9(4), 2001,PP 358-366.

[9] Nagarajan T, Hema A M, Hegde R M, "Segmentation speech into syllable – like units." In EUROSPEECH – 2003, PP 2893-2896.

[10] Nagarajan T, Murthy H A, "Non-bootstrap approach to segmentation and labelling of continuous speech", In National Conference on Communication, IISc Bangalore, India, 2004, PP 508–512.

[11] NagarajanT, Kamakshi Prasad V, Hema A M, "The minimum phase signal derived from the magnitude spectrum and its application to speech segmentation." In Sixth biennial conference of signal processing and communications, 2001.

[12] Dr.K.V.N.Sunitha, N.Kalyani "Improving the word coverage by using Unsupervised Morphological Analyser" , Sadhana-"Academy Proceedings in Engineering Sciences by IISC Bangalore, ISSN0256-2499 (Print) 0973-7677 (Online), vol 34, October 2009, Springer India, in co-publication with Indian Academy of Sciences

[13] Dr.K.V.N.Sunitha, N.Kalyani "Syllable Analysis to Build a Dictation System in Telugu language", in proceedings of International Journal of Computer Science and Information - (IJCSIS) Vol 6 N0 3.(pp. 171-176) in Dec' 09.

[14] Dr.K.V.N.Sunitha, N.Kalyani "Unsupervised stemmer to improve rule based morph analyser" is published in International Journal of Computer Information Systems and Industrial Management Applications - IJCISIM-2010 ISSN: 2150 – 7988, Vol.2 (2010), pp.179-186 in May'10.

[15] Dr.K.V.N.Sunitha, N.Kalyani "DIVINE-Desired Information for the Visually Impaired to Navigate Everywhere" is published in journal of Advances in Computational Sciences and Technology, ACST . Print ISSN 0973-6107, Online ISSN 0974 – 4738, 2011.

[16] Dr.K.V.N.Sunitha, N.Kalyani, "Isolated Word Recognition using Morph–Knowledge for Telugu Language" is published in International Journal of Computer Applications, ISSN 0975-8887, Vol 38 No 12, January 2012.

[17] Dr.K.V.N.Sunitha , N.Kalyani "Unsupervised Morphological Analyser for Hindi Language" at winter school IASNLP organized by IIIT, Hyderabad on 7th Jan 08.

