

Syllables Selection for the Development of Speech Database for Punjabi TTS System

Parminder Singh¹ and Gurpreet Singh Lehal²

¹ Dept. of Computer Science & Engineering, Guru Nanak Dev Engineering College,
 Ludhiana, Punjab 141006, India

² Dept. of Computer Science, Punjabi University,
 Patiala, Punjab 147002, India

Abstract

The Selection of the speech unit and then the number of speech units for the speech database is one of the important and tedious job. Syllables have been reported as good choice of speech unit for speech database of many languages. For this work also, syllables have been selected as the speech unit for the development of the Punjabi speech database. For minimizing the database size, efforts have been made for the selection of the minimal set of syllables covering almost whole Punjabi word set. To accomplish this all Punjabi syllables have been statistically analyzed on the Punjabi corpus having more than 104 million words. This analysis helped to select a relatively smaller syllable set (about first ten thousand syllables (0.86% of total syllables)) of most frequently occurring syllables having cumulative frequency of occurrence (FOO) less than 99.81%, out of 1156740 total available syllables. Also to improve the efficiency of the text-to-speech (TTS) system; interesting facts about Punjabi syllables have been obtained based on their FOO at the three (starting, middle and end) positions in the words.

Keywords: *Speech database, Punjabi syllables, Punjabi TTS system.*

1. Introduction

Speech database is a core component of a concatenative TTS system. The quality of the output synthesized speech signal depends upon how accurately speech database has been prepared [1]. Two main factors affect the quality of speech database. First, the selection of speech unit, and second the number of speech units to be stored in the database. The first factor controls the naturalness of the output speech signal and the second one affects the database size and hence the response time and portability of the TTS system. Since Punjabi is a syllabic language [2], so syllable has been selected as the basic speech unit for this TTS system, which preserves within unit co-articulation effects [3, 4]. Further, for limiting the number of syllables in the database, the frequency of occurrence of

all Punjabi syllables in Punjabi corpus has been found. The syllables occurring very less in the corpus have been ignored and are not selected for storage in the database.

1.1 Punjabi language

Punjabi is an Indo-Aryan language spoken by more than 100 million people those are inhabitants of the historical Punjab region (in north western India and Pakistan) and in the Diaspora, particularly Britain, Canada, North America, East Africa and Australia. It is written from left to right using the Gurmukhi (an abugida derived from the *Lanṭā* script and ultimately descended from Brahmi script) as well as Shahmukhi (a version of the Arabic script) scripts. In Gurmukhi script, which follows the “one sound-one symbol” principle, the Punjabi language has thirty eight consonants, ten non-nasal vowels and same numbers of nasal vowels (see Fig. 1).

ਸ	ਹ	ਕ	ਖ	ਗ	ਘ	ਙ	} Consonants			
ਚ	ਛ	ਜ	ਝ	ਟ	ਠ	ਡ		ਢ	ਣ	
ਤ	ਥ	ਦ	ਧ	ਨ	ਪ	ਫ		ਬ	ਭ	ਮ
ਯ	ਰ	ਲ	ਵ	ੜ	ਸ਼	ਖ਼		ਗ਼	ਜ਼	ਫ਼
ਇ	ਈ	ਏ	ਐ	ਅ	ਆ	ਐ	ਉ	ਊ	ਓ	Non-nasal Vowels
ਇੰ	ਈੰ	ਏੰ	ਐੰ	ਅੰ	ਆੰ	ਐੰ	ਉੰ	ਊੰ	ਓੰ	Nasal Vowels

Fig. 1 Punjabi Consonants and Vowels

1.2 Punjabi Syllables

Defining a syllable in a language is a complex task. There are many theories available in phonetics and phonology to define a syllable. In phonetics, the syllables are defined based upon the articulation [5]. However in phonological approach, the syllables are defined by the different sequences of the phonemes. So, combination of phonemes

gives rise to next higher unit called syllable. Further, combination of syllables produces larger units like morphemes and words. So, syllable is a unit of sound which is larger than phoneme and smaller than word. In every language, certain sequences of phonemes and hence syllables are recognized. Using these phonetic sequences and hence structures, all practically possible syllables can be formed those have been discovered so far in ancient and recent literary works. In addition, all theoretically possible syllables can be composed that may or may not yet be used in a language, but valid in the sense that these follow rendering rules for the language at present [6]. A syllable must have a vowel, without vowel, syllable can not exist. In Punjabi seven types of syllables are recognized [2] – V, VC, CV, VCC, CVC, CVCC and CCVC (where V and C represents vowel and consonant respectively), which combine in turn to produce words. The occurrence of syllables of last type CCVC is very rare, and has not been considered in the present work.

As said above, Punjabi language has thirty eight consonants, ten non-nasal vowels and same numbers of nasal vowels; so, the above said seven syllable types results 1127090 syllables in Punjabi with non-nasal vowels and the same number of syllables with nasal vowels and thus giving total of 2254180 syllables in Punjabi.

2. Statistical Analysis

For the development of speech database for this Punjabi TTS system, the syllables of first six types (V, VC, CV, VCC, CVC and CVCC) have been selected and the syllables of type CCVC have not been considered due to their less existence [2]. For selecting syllables for the speech database, the total available 1156740 syllables (nasal and non-nasal) have been statistically analyzed on a Punjabi corpus. For this purpose a carefully selected balanced corpus having 104425741 total and 232565 unique words have been used. Frequency of occurrence of the available 1156740 syllables in the corpus has been found. It has been observed that certain syllables are having frequency of occurrence zero; and these syllables have been declared as invalid syllables in Punjabi. The table 1 and table 2 shows the result of the statistical analysis and shows the valid non-nasal and nasal syllables (having $FOO > 0$) of each type respectively.

Table 1: Statistical analysis result of non-nasal syllables

Syllable Type	Total No. of Syllables	Syllables with Freq. of Occurrence >0
V	10	10
VC	380	297
CV	380	373
VCC	14440	1247
CVC	14440	6076
CVCC	548720	11831
Total	= 578370 (Syllables)	= 19834 (Non zero freq. syllables or valid syllables)

Table 2: Statistical analysis result of nasal syllables

Syllable Type	Total No. of Syllables	Syllables with Freq. of Occurrence >0
V	10	10
VC	380	145
CV	380	329
VCC	14440	194
CVC	14440	1881
CVCC	548720	1659
Total	= 578370 (Syllables)	= 4218 (Non zero freq. syllables or valid syllables)

During utterance stress, duration and articulation of the same syllable is different at the starting, middle and end positions of the words [7]. Hence for the better accuracy of the TTS system, the syllables sounds need to be stored separately for these three positions in the speech database. So, *FOO* at these three positions has been found separately as shown in Table 3 and Table 4 for non-nasal and nasal syllables respectively.

Table 3: Frequency of occurrence at starting, middle and end positions of Non-nasal Syllables

Syllable Type	Freq. of Occurrence at			Total Freq.
	Starting Position	Middle Position	End Position	
V	553025	171660	391167	1115852
VC	635054	27378	137821	800253
CV	5130503	1159233	2533916	8823652
VCC	59764	1470	17193	78427
CVC	2176062	165073	881906	3223041
CVCC	547610	16626	147106	711342

Table 4: Frequency of occurrence at starting, middle and end positions of Nasal Syllables

Syllable Type	Freq. of Occurrence at			Total Freq.
	Starting Position	Middle Position	End Position	
V	18232	29624	265492	313348
VC	27940	1538	3785	33263
CV	738213	76557	608541	1423311
VCC	14762	56	1105	15923
CVC	250759	15188	72046	337993
CVCC	40465	2152	37891	80508

It has been observed that the occurrence of syllables (non-nasal and nasal) at the starting position in the words is more than at middle and ending positions. Results show that there are 18013, 6251 and 10869 syllables having *FOO* 10192389, 1666555 and 5097969 at the starting, middle and end positions respectively. The analysis reveals that the syllables occurring at middle position of words are very less than at starting and end positions. Fig. 2 shows the plot for number of times the syllables are occurring at starting, middle and end positions in the words of the said Punjabi corpus.

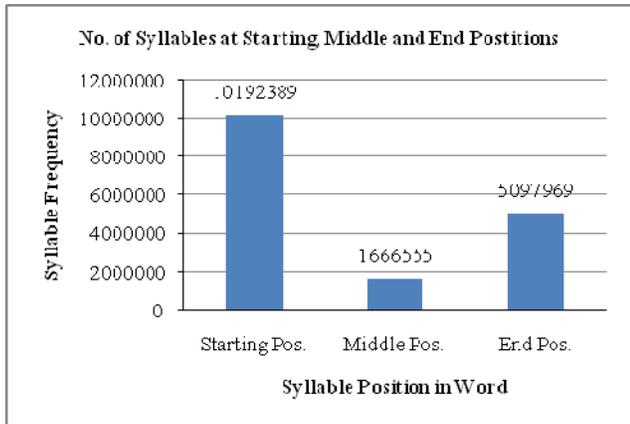


Fig. 2 Syllables frequency at starting, middle and end positions

It has been observed that there are 6674 syllables with unit *FOO*. This is mostly because of the words those are borrowed from other languages and are occurring very rare in Punjabi and their one time existence in the corpus has given the unit *FOO* to its syllables, those are not otherwise occurring in general in the language. There are 20099 syllables having their frequency in the range 1 to 100. Fig. 3 shows the plot for number of syllables against the *FOO* in the range 100 to 1000.

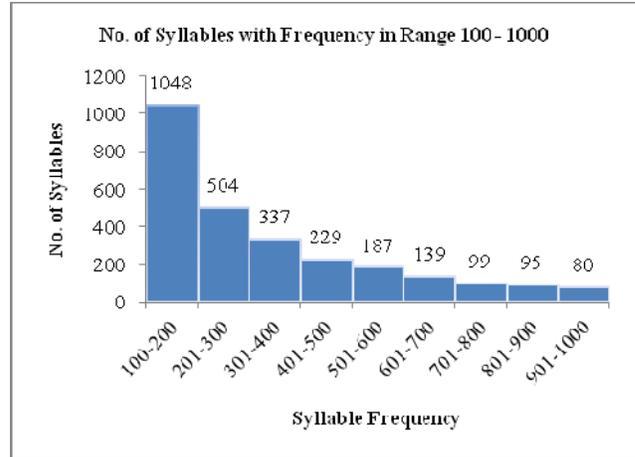


Fig. 3 Number of syllables with frequency in the range 100 – 1000

Fig. 4 shows the plot for number of syllables against the *FOO* in the range 1000 to 10000.

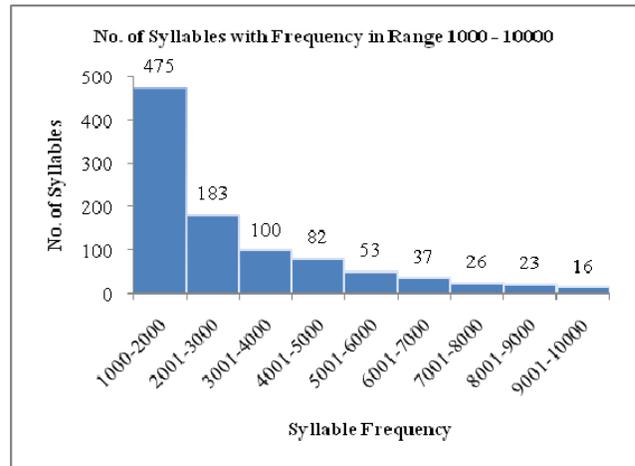


Fig. 4 Number of syllables with frequency in the range 1000 – 10000

Fig. 5 shows the plot for number of syllables against the *FOO* in the range 10000 to 100000. The number of syllables with *FOO* more than 100000 is only 39.

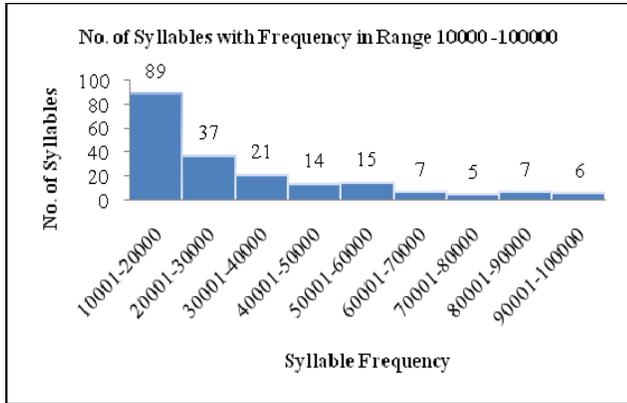


Fig. 5 Number of syllables with frequency in the range 10000 – 100000

The combined sorted list of nasal and non-nasal syllables results only 24052 syllables, out of total 1156740, having *FOO* (total of three positions) more than zero. Out of these, first 10000 (approx.) most frequently occurring syllables having cumulative *FOO* less than 99.81% have been selected for the development of speech database. Only the syllables having *FOO* less than 8 have been omitted and these will not affect the working of the TTS system because of their very less occurrence. So with these many syllables, the TTS system will be able to cover almost all Punjabi words as well as would be able to generate words borrowed from other languages and names of persons; and hence producing a general TTS system for Punjabi.

The word coverage (number of unique words in which a particular syllable is occurring) by the syllables provides important information about the syllables and it has also been found for the unique 232565 words of the above said corpus. Results show that non-nasal syllable type CV is having maximum number of word coverage (264349 unique words) than other syllables. Fig. 6 and Fig. 7 show these results for the non-nasal and nasal syllables respectively.

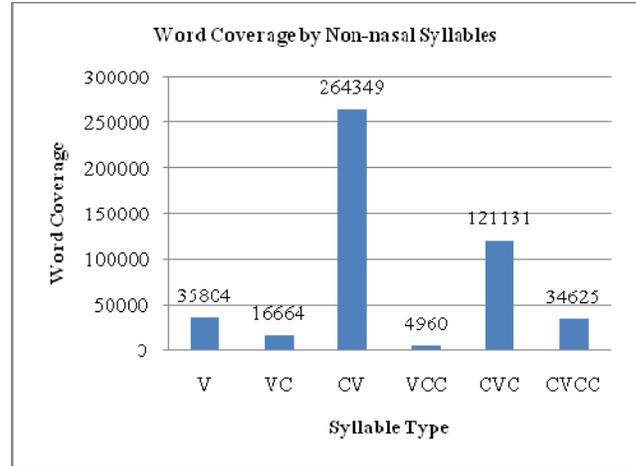


Fig. 6 Word coverage by non-nasal syllables

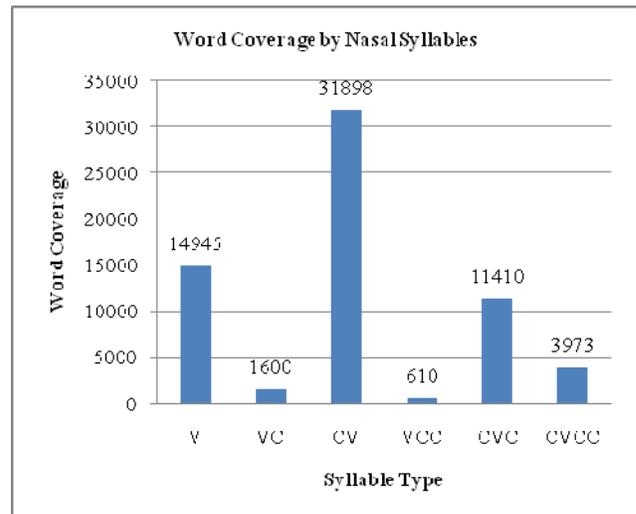


Fig. 7 Word coverage by nasal syllables

3. Conclusions

It has been observed from the above results that the statistical analysis of the Punjabi syllables over the Punjabi corpus plays a vital role in selection of syllables for the speech database. Results show that large number of syllables are not occurring even once in the corpus of about 104 million words and are declared as invalid syllables. Also, other large number of syllables are having comparatively very little frequency of occurrence and are ignored for the final selection. Also for improving the quality of speech database the syllables have been analyzed statistically for the three (starting, middle and

end) positions in the words of the corpus. This statistical analysis helped to select a relatively small syllable set (of about first ten thousand syllables, that are about 0.86% of total syllables) of most frequently occurring syllables having cumulative frequency of occurrence less than 99.81%, out of 1156740 available syllables. The results of this statistical analysis will also be very much helpful for the implementation of the other syllable based NLP systems.

References

- [1] N. Kalyani, and K.V.N. Sunitha, "Syllable analysis to build a dictation system in Telugu language", International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009, pp. 171-176.
- [2] P. Singh, Sidhantik Bhasha Vigeyan, Patiala (India): Madan Publications, 2002.
- [3] E.V. Raghavendra, S. Desai, B. Yegnanarayana, A.W. Black, and K. Prahallad, "Global syllable set for building speech synthesis in Indian languages", in IEEE Workshop on Spoken Language Technology, Goa (India), 2008, pp. 49-52.
- [4] M.L. Narayana, and A.G. Ramakrishnan, "Defining syllables and their stress in Tamil TTS corpus", in Workshop on Image and Signal Processing, IIT Guwahati (India), 2007, pp. 92-95.
- [5] R.A. Krakow, "Physiological organization of syllables: a review", Journal of Phonetics, Vol. 27, 1999, pp. 23-54.
- [6] R.K. Joshi, K. Shoff, and S.P. Mudur, "A phonemic code based scheme for effective processing of Indian Languages", in 23rd Internationalization and Unicode Conference, Prague, 2003.
- [7] H. Lee, and C. Seong, "Experimental phonetic study of the syllable duration of Korean with respect to the positional effect", in The Fourth International Conference on Spoken Language Processing (ICSLP'96), Philadelphia (USA), 1996, pp. 1193-1196.

Parminder Singh is presently working as Associate Professor in Department of Computer Science and Engineering at Guru Nanak Dev Engineering College, Ludhiana (India). He is M.Tech. and B.Tech. in Computer Science and Engineering and is currently pursuing Ph.D. His field of interest is Natural Language Processing (Speech Synthesis). He has guided sixteen M.Tech. thesis and has published about fifteen research papers in international and national journals and conferences.

Gurpreet Singh Lehal is Professor in Department of Computer Science and Director of Advance Centre for Technical Development of Punjabi Language Literature and Culture at Punjabi University, Patiala (India). He is a member of IEEE and ACM. He has about sixty publications to his credit in international journals and conferences. He has developed software like Punjabi Word Processor named Akhar, OCR for Gurmukhi, Gurmukhi-Shahmukhi transliteration system, Search

Engine for Punjabi, Devnagri to Urdu and Urdu/Kashmiri to Roman script transliteration systems, Punjabi morphological analyzer, and intelligent predictive Romanized typing utility for Gurmukhi text. His field of expertise is Natural Language Processing. He is team leader for many projects funded from abroad and Govt. of India.