Corpus Based Statistical Analysis of Punjabi Syllables for Preparation of Punjabi Speech Database

Parminder Singh
Associate Professor, Dept. of Computer Sc. & Engg., Guru Nanak Dev Engg. College, Gill Road, Ludhiana – INDIA

Gurpreet Singh Lehal
Professor, Dept. of Computer Science, Punjabi University, Patiala (Pb.) – INDIA

Abstract

Syllables preserve within unit co-articulation effects, which results syllables to be a good choice as speech unit for speech database of many languages. Punjabi is a syllabic language, so syllables have been selected as the basic speech unit for the preparation of speech database for Punjabi. In order to select a minimal syllable set covering almost whole Punjabi word set to be stored in the speech database, all the available Punjabi syllables have been statistically analyzed over a large Punjabi corpus having more than 104 million words. This statistical analysis reveals very interesting and important facts about the actual use of Punjabi syllables by its speaker and writers. From the results of this analysis a relatively smaller syllable set of about first ten thousand (0.86% of total available 1156740 syllables) most frequently occurring syllables having cumulative frequency of occurrence less than 99.81% have been selected. Also, frequency of occurrence of these syllables at the three positions (starting, middle and end) in the words of corpus has been obtained, for detailed storage and better utilization of these syllables in certain applications like Text-to-Speech synthesis.

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.

2. Research background

2.1 Punjabi language

Punjabi is an Indo-Aryan language spoken by more than hundred 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 Australasia. Punjabi is unusual among modern Indo-European languages in being a tonal language. It is written from left to right using the Gurmukhi (an abugida derived from the Laṇḍā script and ultimately descended from Brahmi script) as well as Shahmukhi (a version of the Arabic script) scripts. Punjabi language like other Indian languages includes segmental phonemes (vowels and consonants), but not supra-segmental phonemes (stress, intonation, juncture, nasality and tone) in its alphabet. Segmental phonemes having their independent existence, are the discrete units that can be identified either physically or auditory in the stream of speech. Supra-segmental phonemes on the other hand do not have independent existence; these exist only with segmental phonemes. In Gurmukhi script, which follows the *one sound-one symbol* principle, Punjabi language has thirty eight consonants, ten non-nasal vowels (ਇ, ਈ, ਏ, ਐ, ਅ, ਆ, ਔ, ਉ, ਊ, ਓ) and same numbers of nasal vowels (ਇੰ, ਈਂ, ਏਂ, ਐਂ, ਅਂ, ਆਂ, ਔਂ, ਉਂ, ਊਂ, ਓਂ) (see Figure 1). The consonants [ਙ] and [ਞ] are rarely utilized.

Vowels can appear alone in orthography (known as full vowels) however consonants can appear along with vowels only. Each consonant in written Punjabi is associated with inherent schwa ('ਅ' the third character of Punjabi alphabet and written as [ə] in International Phonetic Alphabet (IPA) transcription); if it is not associated with any of the other vowels.

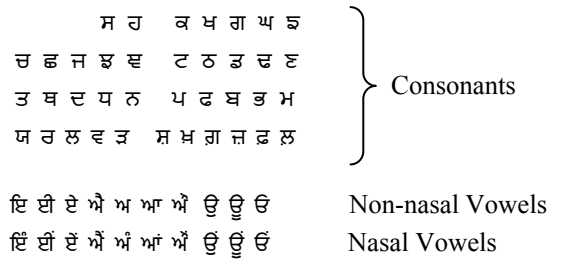

Figure 1. Punjabi consonants and vowels

While pronouncing the written word, the speaker retains the intervening schwa vowel associated with a consonant where required and eliminate it from pronunciation where it is not required. Vowels, except schwa, are represented diacritically when these come along with consonants (known as half vowels), otherwise as such. Diacritics in Punjabi can appear above, below, before or after the consonant they belong to. The consonant sound varies according to the vowel attached to consonant. For example, consonant [ਸ] conjoined with vowel [ਈ] (having diacritic ੀ) results a single orthographic unit "ਸੀ", having pronunciation of a consonant-vowel sequence /ਸ+ਈ/ (/si/) however when this consonant comes with vowel [ਆ] the resulting single unit [ਸਾ] will be pronounced as /ਸ+ਆ/ (/sā/). Consonants represented in orthography without any attached diacritic, basically have the attached schwa vowel that is not represented diacritically. There are three consonants of Punjabi [ਹ], [ਰ] and [ਵ] those can have their respective sounds intermixed immediately after the sound of other consonants. These consonants are used in Punjabi as conjuncts and appear at the bottom of a barer consonant. These are half *haha* as in ਨ੍ (=[ਨ੍]+[ਹ]), half *rara* as in ਪ੍ (=[ਪ੍]+[ਰ]) and half *vava* as in ਸ੍ (=[ਸ੍]+[ਵ]). In the conjugal pair of consonants the first consonant is vowelless however the second consonant (represented as half) has the associated vowel. Contrary to the orthographic representation of consonant conjuncts, first consonant of the conjunct has half phonetic effect whereas the second one has its full effect phonetically [6]. For example the consonant conjunct [ਪ੍ਰ] represented as a single unit in word ਪ੍ਰਭੂ (/prabu/ means God) gives the little sound of [ਪ] /p/ followed by the full sound of [ਰ] /r/ opposite to its orthographic representation. Gemination is a distinctive feature of Punjabi and is represented by the symbol [ੱ] called *addak* [7]. The first consonant of geminate is vowelless and the second consonant can occur with any of the vowels. The use of *addak* indicates that the following consonant is geminate. Gemination occurs at the middle or end position of the word, however not at the starting position. For example in word "ਅੱਗ" ([agg] means fire) the consonant [ਗ] is geminate.

2.2 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 different types of Punjabi syllables with examples are given in Table 1.

Table 1. Types of Punjabi syllables

Type	Pattern	Example
V	Vowel	ਆ
VC	Vowel-Consonant	ਉਹ (ਉ+ਹ)
CV	Vowel-Consonant	ਖਾ (ਖ+ਆ)
VCC	Vowel-Consonant-Consonant	ਅੱਗ (ਅ+ਗ+ਗ)
CVC	Consonant-Vowel-Consonant	ਗੀਤ (ਗ+ਈ+ਤ)
CCVC	Consonant-Consonant-Vowel-Consonant	ਪ੍ਰੀਤ (ਪ+ਰ+ਈ+ਤ)
CVCC	Consonant-Vowel-Consonant-Consonant	ਧਰਮ (ਧ+ਰ+ਮ)

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.

3. Statistical analysis of syllables

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 2 and Table 3 show the result of the statistical analysis and show the valid non-nasal and nasal syllables (having Frequency of occurrence > 0) of each type respectively.

Table 2. 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 (Total syllables)***	***= 19834 (Non zero freq. syllables or valid syllables)***

Table 3. 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 (Total 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, frequency of occurrence at these three positions has been found separately as shown in Table 4 and Table 5 for non-nasal and nasal syllables respectively.

Table 4. 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 5. 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 frequency of occurrence 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. The Figure 2 shows the plot for number of times the syllables are occurring at staring, middle and end positions in the words of the said Punjabi corpus.

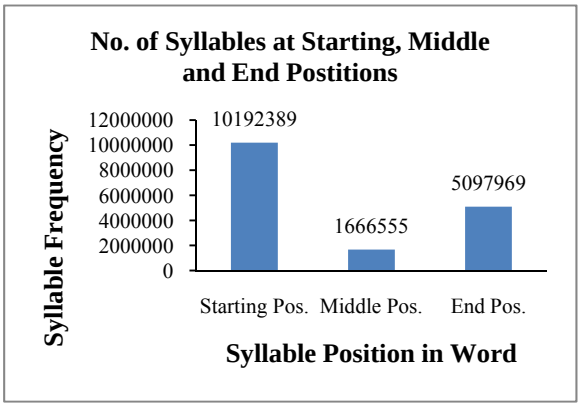

Figure 2. Syllables frequency at starting, middle and end positions

It has been observed that there are 6674 syllables with unit frequency of occurrence. 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 frequency of occurrence 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. Figure 3 shows the plot for number of syllables against the frequency of occurrence in the range 100 to 1000.

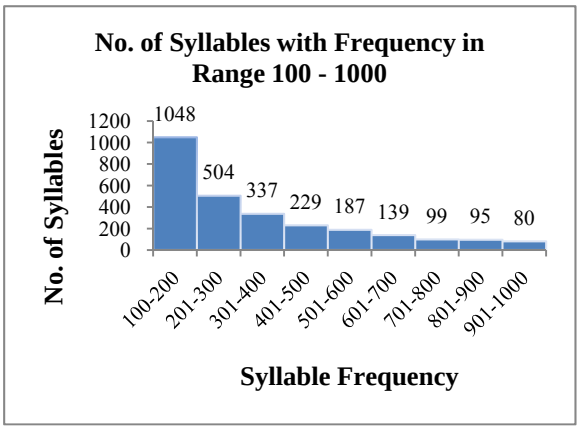

Figure 3. Number of syllables with frequency in the range 100 – 1000

Figure 4 shows the plot for number of syllables against the frequency of occurrence in the range 1000 to 10000. Figure 5 shows the plot for number of syllables against the frequency of occurrence in the range 10000 to 100000. The number of syllables with frequency of occurrence more than 100000 is only 39.

The combined sorted list of nasal and non-nasal syllables results only 24052 syllables, out of total 1156740, having frequency of occurrence (total of three positions) more than zero. Out of these, first 10000 (appox.) most frequently occurring syllables having cumulative frequency of occurrence less than

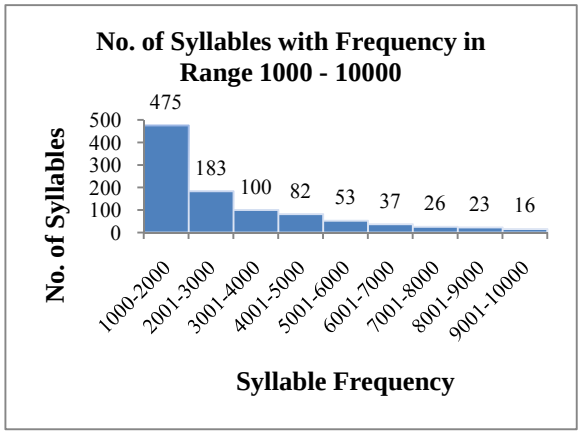

Figure 4. Number of syllables with frequency in the range 1000 – 10000

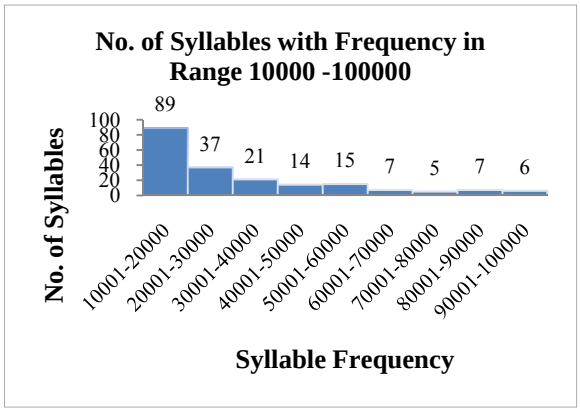

Figure 5. Number of syllables with frequency in the range 10000 – 100000

99.81% have been selected for the development of speech database. Only the syllables having frequency of occurrence 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 syllables of type CV are occurring in maximum number of words and thus having maximum word coverage (264349 unique words) than other syllables. The nasal syllables of type VCC are having minimum word coverage of 610 unique words than other syllables. Figure 6 and Figure 7 show these results for the non-nasal and nasal syllables respectively.

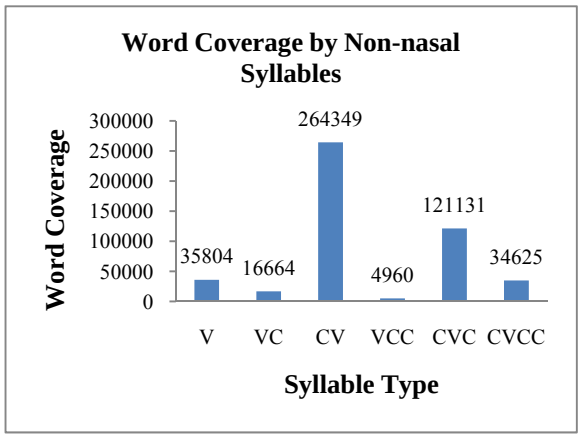

Figure 6. Word coverage by non-nasal syllables

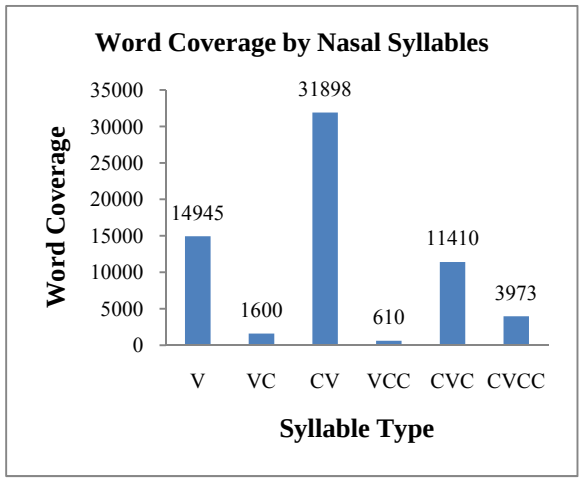

Figure 7. Word coverage by nasal syllables

4. 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 (staring, 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.

5. References

[1] N. Kalyani, 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*, 4th Edition, Madan Publications, Patiala India (2002), pp. 371-372.

[3] E.V. Raghavendra, S. Desai, B. Yegnanarayana, A.W. Black, K. Prahallad, "*Global syllable set for building speech synthesis in Indian languages*", in proc. of IEEE Workshop on Spoken Language Technologies (2008), Goa, India.

[4] Narayana, M.L., Ramakrishnan, A.G., "*Defining syllables and their stress in Tamil TTS corpus*", in proc. of Workshop in Image and Signal Processing (WISP-2007), 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, S.P. Mudur, "*A phonemic code based scheme for effective processing of Indian Languages*", in proc. of 23rd Internationalization and Unicode Conference (2003), Prague.

[7] H. Lee, C. Seong, "*Experimental phonetic study of the syllable duration of Korean with respect to the positional effect*", in proc. of 4th International Conference on Spoken Language Processing, Philadelphia, PA, USA, (1996), pp. 1193-1196.