Shipra J. Arora* and Rishipal Singh

Database Creation and Dialect-Wise Comparative Analysis of Prosodic Features for Punjabi Language

https://doi.org/10.1515/jisys-2019-2511
Received June 21, 2017; previously published online March 19, 2019.

Abstract: The paper represents a Punjabi corpus in the agriculture domain. There are various dialects in the Punjabi language and the main concentration is on major dialects, i.e. Majhi, Malwai and Doabi for the present study. A speech corpus of 125 isolated words is taken into consideration. These words are uttered by 100 speakers, i.e. 60 Malwi dialect speakers (30 male and 30 female), 20 Majhi dialect speakers (10 male and 10 female) and 20 Doabi dialect speakers (10 male and 10 female). Tonemes, adhak (geminated) and nasal words are selected from the corpus. Recordings have been processed through two mediums. The paper also elaborates some distinctive features of the corpus. This corpus is of quite significance for the speech recognition system. Prosodic characteristics such as intonation, rhythm and stress create a crucial impact on the speech recognition system. These characteristics vary from language to language as well as various dialects of a language. This paper portrays a comparative analysis of isolated words prosodic features of Malwi, Majhi and Doabi dialects of Punjabi language. Analysis is done using the PRAAT tool. Pitch, intensity, formant I and formant II values are extracted for toneme, adhak, nasal (bindi) and nasal (tippi) words. For all kinds of words, there is a significant variation in pitch (fundamental frequency), intensity, formant I and formant II values of male and female speakers of Malwi, Majhi and Doabi dialects. A detailed analysis has been discussed throughout this paper.

Keywords: Speech corpus; dialects; tones; prosodic features; pitch; intensity; formants.

1 Introduction

Agriculture is one of the main occupations of people of Punjab which forms the moral fiber of economy of the Punjab state. The state has the maximum growth rate in production of food. Approximately 1.54% of the country's geographical area comprises this state. It is surrounded by five rivers and can be considered as one of the most prolific areas on the earth.

There are various ways of communication, and speech is one of the most important and efficient methods of communication between human beings and their other environmental and non-environmental elements. Therefore, manufacturing of an automatic speech recognizer is very desirable and favorable. Speech recognition made it possible for the computers to understand human languages. As information technology has the impact on various aspects of human lives, and in recent years, it has affected a lot on human life, hence it becomes increasingly important to resolve the problem of communication between human beings and information processing devices. In the past, a lot of communication has been done through the usage of keyboards and screens, but speech is the most widely used, natural, convenient and the fastest means of communication for human beings.

Agriculture is the major livelihood of people in Punjab. Normally, farmers do not have scientific information about crops. They do not have interest in visiting an information center or a website to get information

***Corresponding author: Shipra J. Arora,** CSE Department, Guru Jambheshwar University of Science and Technology, Hisar 125001, Haryana , India, e-mail: jkshipra22@gmail.com
Rishipal Singh: CSE Department, Guru Jambheshwar University of Science and Technology, Hisar 125001, Haryana, India

about crops. Even they are not capable of typing on computer. The reasons for requirement of Punjabi corpora are as follows:

(a) to alleviate the communication of Punjabi speakers in India and abroad in their native language with smart devices;
(b) to upgrade the Punjabi speaker information retrieval system without the use of a mouse and keyboard;
(c) there is lack of resources;
(d) there is low literacy level; and
(e) visually impaired people find it inconvenient to use a Braille keyboard.

Punjabi speech corpora thus created are helpful to rural people who can be benefitted by automatic speech processing and can communicate with computer in their own native language.

Punjabi language is mainly used in two countries, India (East Punjab) and Pakistan (West Punjab), and on a small scale spoken in few other countries. It is a state language of Punjab in India. It has 10 vowels and 41 letters consisting of 38 consonants and 3 basic vowel sign bearers. Out of 38 consonants, there are 6 consonants with a dot below which are used to represent borrowed words from other languages. It has three conjunct vargs and also three signs, i.e. bindi, tippi and adhak. The script of Punjabi language is Gurumukhi which is based on the one-sound one-symbol rule. Punjabi is a tonal language. Segmental and super-segmental sounds are the taxonomy of speech sounds. Segmental sounds are further classified into vowels and consonants, but super-segmental sounds concentrate on tone, intonation and stress.

2 Related Work

TIFR Mumbai and IIT Bombay created a Marathi language speech database in a noisy environment. Data were recorded by 1500 speakers using cell phones and voice recorders. The database was created for the purpose of the automatic speech recognition system in the agriculture domain [2, 3, 16].

Punjabi University Patiala created a Punjabi language speech database consisting of 3312 syllables in a studio environment using a standard microphone. The database was created for the purpose of the text to speech synthesis system [9, 11].

KIIT, Gurgaon, created a Hindi and Indian English mobile-based speech database in a noise-free environment. Nokia Research Centre, China, sponsored this project. It consisted of 13 prompt sheets, each having 630 phonetically rich sentences. Data were recorded by 100 speakers (60 female and 40 male), using cell phones, and omni-directional and cardioid microphones [3].

TIFR Mumbai and CDAC Noida created a Hindi speech database in a noise-free environment. Data were recorded by 100 speakers using standard microphones. The database consisted of 10 phonetically rich Hindi sentences. It was created for the purpose of the speech recognition system [1, 2, 15].

IIT Hyderabad created a Telugu, Hindi and English speech database in a noise-free environment. Data were recorded by 15 speakers using standard microphones and laptops. Data were divided into the following domains: (i) local travel, (ii) hotel and restaurant transactions, (iii) tourism and (iv) emergency services, to overcome the problems faced by tourists because they were incapable of understanding native language [9, 10].

3 Text and Speech Corpora Collection

The fundamental step of speech recognition is anthology of vocabulary words which is to be recorded by native speakers. The corpus consists of 125 isolated words in the agriculture domain. These words have been recorded by 100 speakers (50 male and 50 female) in different dialects. Recordings have been processed in two recording mediums: laptop-mounted microphone and mobile phone using PRAAT software having a sampling rate of 16 kHz/16 bits. Speech corpora of 12,500 isolated words of 100 speakers have been gathered,

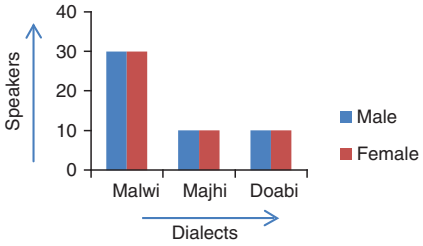

Figure 1: Dialect-Wise Data Distribution – Male and Female Speakers.

processed and organized in totality. Corpora characterize phonemic and phonetic characteristics of speech sounds. The dialect-wise distribution of data is shown in Figure 1.

4 Corpus Features

The text and speech corpora designed so far have many unique features.

4.1 Tonemes

Tone is the distinctive feature of Punjabi language. T. Grahame Bailey stated in 1914, about a century ago: "Variations in the tone of the voice form a very remarkable feature of Punjabi pronunciation. There are two special tones, apart from the ordinary tone of speaking. They occur in stressed syllables only" [15]. Velar (ਘ), Palatal (ਝ), Retroflex (ਢ), Dental (ਧ) and Bilabial (ਭ) are the toneme consonants. When these tonemes are used at the beginning position of words, they are pronounced as Velar (ਕ), Palatal (ਚ), Retroflex (ਟ), Dental (ਤ) and Bilabial (ਪ) with low vowel tone. But when these tonemes are used at the middle or final position of words, they are pronounced as Velar (ਗ), Palatal (ਜ), Retroflex (ਡ), Dental (ਦ) and Bilabial (ਬ) with high tone on the vowel before the consonants and low tone on the vowel after the consonants. Some of the toneme words in our corpora are ਝੋਨਾ (tʃòna), ਭੰਗ (pə̀ng), ਪੱਤਾਗੋਭੀ (pə̀ʈʈagobi), ਭਿੰਡੀ (pìnɖi), ਫੁਲਗੋਭੀ (p^{h}ʊlgobí), etc.

If ਹ (h) is used at the beginning of the word, then it is pronounced regularly, but if it is used at the middle and final position of the word, it is pronounced with high tone on the preceding vowel. Some of the words in our corpora are ਚਾਹ (tʃá), ਕਪਾਹ (kəpá), etc.

4.2 Nasal

Nasal phonemes are produced by using tippi and bindi. Tippi is used with vowels ਅ (Mukta), ਇ (Sihari), ਉ (Onkarh) and ਊ (Dulankarh), while bindi is used with vowels ਆ (kanna), ਈ (Bihari), ਏ (Lanva), ਐ (Dulavan), ਓ (Horha) and ਔ (Kanaorha). Both the tippi and bindi serve the same usage, i.e. used to emphasize nasal sound. Some of the nasal words in our corpora are: ਅੰਬ (əmb), ਮੁੰਗੀ (mʊngi), ਭੰਗ (pə́ng), ਮੁੰਗਫ਼ਲੀ (mungfəli), ਗੰਨਾ (gənna), ਹੀਂਗ (hing), ਲੋਂਗ (long), etc.

4.3 Adhak

Adhak is used to geminate the consonant sound. It is located between two letters. Letter which follows adhak is to be repeated. It is very important. Without it, words have different meaning. For example, *ਕੱਦ means height* and *ਕਦ means when*. Some of the words in our corpora are ਮੁਲੱਠੀ (mʊlə$ʈ^{h}ʈ^{h}$i), ਸ਼ੱਕਰਕੰਦੀ (ʃəkkərkə̀nɖi), ਪੱਤਾਗੋਭੀ (pə̀ʈʈagobí), ਮੱਕੀ (məkki), etc.

4.4 Dialects

Punjabi is spoken all over the world. It has various dialects, but major dialects are Majhi, Doabi and Malwai. Majhi is a prominent dialect which is a benchmark of written Punjabi. Major areas are Amritsar, Gurdaspur (districts of Punjab), various states of Haryana, Uttar Pradesh, etc. Doabi dialect is spoken in Doaba Punjab. Doabi means the region between two rivers, i.e. Beas and Satluj. Major areas are Jalandhar, Kapurthla, Hoshiarpur, etc. Malwi dialect is spoken in Malwa province of Punjab. Major areas are Ferozpur, Fazilka, Muktsar, Faridkot, etc. If "v" is used at the beginning of a word, Doabi use letter "b" instead of "v" or otherwise use "o". Some of the words in the corpus are *ਗਵਾਰ* (gəvar), *ਜਵਾਰ* (jəvar), etc. Doabi speakers use "e" in place of "i". Words in the corpus are *ਇਲਾਚੀ* (ɪlatʃi), *ਇਮਲੀ* (Imli), etc.

4.5 Speakers Variety According to Age, Sex and Location

Isolated words have been recorded by 50 male and 50 female speakers in the age group of 18–35 years. Speakers belong to different cities and villages of various districts of Malwi, Majhi and Doabi dialects.

4.6 Speech Database Annotation

Annotation of speech corpora has been done using wave surfer at the word level. Speech corpora of 125 words spoken by 100 speakers, i.e. 12,500 words, have been segmented and labeled properly in order to study phonetic features further. Durations of segmented data were stored in a lab file. The sample is shown in Figure 2.

4.7 Phonetic Features of Speech Sounds

Man-machine interaction involves the integration of latest technologies for speech input as well as speech output. Speech sounds can be classified into consonants and vowels. Place of articulation refers to restriction of air in the vocal tract. It involves glottis, lip, jaw, teeth, tongue, lungs, larynx and other vocal parts. The manner of articulation refers how close articulators move toward each other, i.e. the way in which air is tailored in the vocal tract. Speech corpora designed so far cover approximately all categories and phonetic features of speech sounds.

5 Speech Databases

Various organizations and institutes are conducting research in the area of speech processing. They have created different speech databases. Databases can be compared on the basis of the following parameters:

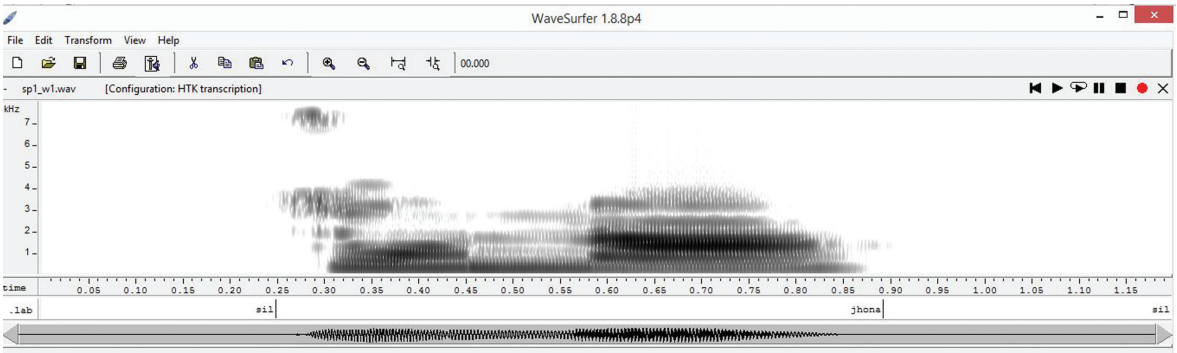

Figure 2: Annotation of Word *ਝੋਨਾ* (tʃõna).

(a) noisy or noise-free environment, (b) number of speakers, (c) recording devices, (d) languages, (e) speech types and (f) purpose [4, 6].

Most of the speech databases have been created in a noise-free environment. But researchers should also create a speech database in a noisy environment to meet the real-life situations and to build a robust automatic speech recognition system [5, 7, 8]. Speech databases that have been created for text-to-speech synthesis are not quite significant for the speech recognition system. Speech databases that have been recorded using landline and mobile phones are sort of deficient due to network errors. The major work in this area has been done for Hindi, Tamil, Bengali and Marathi languages [12–14]. A little work has been done for other Indian languages. In order to build up wider language technologies base for all Indian languages, researchers should do more and more work in this area for all languages.

6 Research Method

The PRAAT tool is used for comparative analysis. Isolated words have been classified into toneme, adhak, nasal (bindi) and nasal (tippi) words. For comparison, these words are spoken by 10 male and 10 female speakers of Malwi, Majhi and Doabi dialects. Pitch, intensity, formant I and formant II values are extracted and compared.

7 Results and Analysis

It has been observed from Table 1 and Figure 3 that the pitch of male speakers is very less than that of female speakers for Malwi, Majhi and Doabi dialects. In case of female speakers, for toneme words and nasal (tippi) words, pitch variation is the same and as follows: Majhi > Malwi > Doabi. For adhak (geminated) and nasal (bindi) words, pitch variation is the same and as follows: Malwi > Majhi > Doabi. In case of male speakers, for all types of words, pitch variation is the same and as follows: Doabi > Malwi > Majhi.

Table 1: Pitch Variations for Different Types of Words in the Database.

Words type	Male/female speakers	Malwi dialect	Majhi dialect	Doabi dialect
Toneme words	Female	258.6585	261.1291	250.9955
	Male	146.3769	128.636	165.7305
Adhak words	Female	252.6602	251.9879	249.4705
	Male	145.521	128.8458	166.6938
Nasal (bindi) words	Female	252.4579	249.7492	244.4647
	Male	141.9681	128.5487	163.6569
Nasal (tippi) words	Female	252.4445	257.3207	245.678
	Male	143.059	129.1502	163.0176

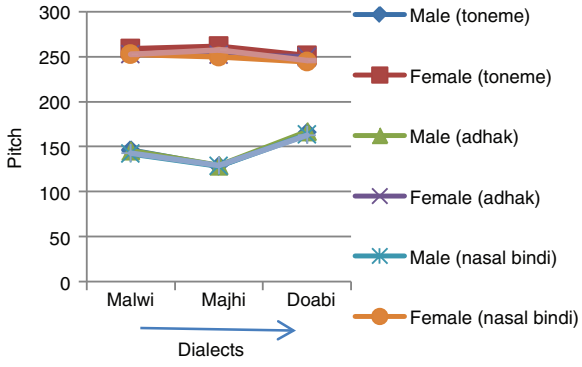

Figure 3: Dialect-Wise Pitch Variation for Tonemes, Adhak, Nasal (Bindi) and Nasal (Tippi) Words.

It has been observed from Table 2 and Figure 4 that the intensity of male speakers is less than that of female speakers for Malwi, Majki and Doabi dialects. In case of female speakers, for toneme, adhak (geminated) and nasal (bindi) words, intensity variation is the same and as follows: Malwi > Majhi > Doabi. For nasal (tippi) words, intensity variation is as follows: Malwi > Doabi > Majhi. In all cases, there is very little variation in Majhi and Doabi dialect speakers.

In case of male speakers, for nasal (bindi) and nasal (tippi) words, intensity variation is as follows: Doabi > Malwi > Majhi. For toneme words, it is as follows: Malwi > Doabi > Majhi. For adhak (geminated) words, it is as follows: Malwi > Majhi > Doabi. In all cases, there is very little variation in Majhi and Doabi dialect speakers.

It has been observed from Table 3 and Figure 5 that the formant I value of male speakers is greater than that of female speakers for Malwi, Majhi and Doabi dialects. In case of female speakers, for all types of words, Formant I variations are as follows: Doabi > Malwi > Majhi. In case of male speakers, for nasal (bindi) words and toneme words, formant I variations are as follows: Malwi > Doabi > Majhi. For nasal (tippi) words and adhak (geminated) words, the variations are as follows: Doabi > Malwi > Majhi.

Table 2: Intensity Variations for Different Types of Words in the Database.

Words type	Male/female speakers	Malwi dialect	Majhi dialect	Doabi dialect
Toneme words	Female	78.07429	74.98018	74.37597
	Male	74.14868	73.77925	73.98708
Adhak words	Female	76.74887	73.22307	71.91624
	Male	73.09751	72.60313	72.3589
Nasal (bindi) words	Female	78.52311	74.69312	74.48076
	Male	73.81041	73.59468	74.00996
Nasal (tippi) words	Female	77.99654	74.06132	74.43187
	Male	73.90844	73.66547	74.07185

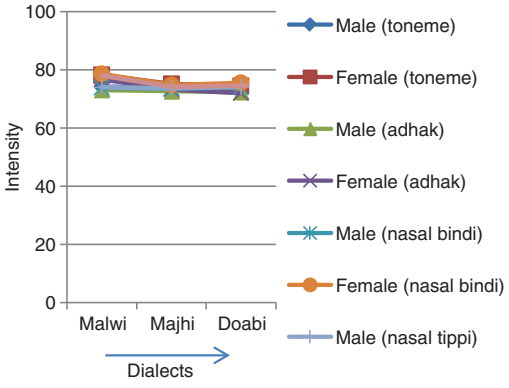

Figure 4: Dialect-Wise Intensity Variation for Tonemes, Adhak, Nasal (Bindi) and Nasal (Tippi) Words.

Table 3: Formant I Variations for Different Types of Words in the Database.

Words type	Male/female speakers	Malwi dialect	Majhi dialect	Doabi dialect
Toneme words	Female	584.3751	544.1478	591.7777
	Male	743.0935	567.2225	723.4262
Adhak words	Female	505.2374	492.3279	510.929
	Male	630.945	513.1304	634.4885
Nasal (bindi) words	Female	527.9302	488.2497	544.5416
	Male	688.9413	541.2241	688.2608
Nasal (tippi) words	Female	473.5769	453.0424	508.958
	Male	622.5815	480.2751	641.8084

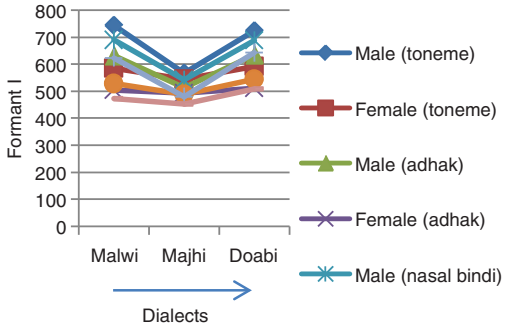

Figure 5: Dialect-Wise Formant I Variation for Tonemes, Adhak, Nasal (Bindi) and Nasal (Tippi) Words.

It has been observed from Table 4 and Figure 6 that the formant II value of male speakers is greater than that of female speakers for Malwi and Doabi dialects but less than in case of Majhi dialect. In case of female speakers, for toneme and adhak (geminated) words, formant II variations are as follows: Malwi > Majhi > Doabi. For nasal (bindi) and nasal (tippi) words, the variations are as follows: Majhi > Malwi > Doabi.

In case of male speakers, for all kinds of words, formant II variations are as follows: Malwi > Doabi > Majhi.

In speech at different levels, dialect-related information is available. Spectral features such as mel-frequency cepstral coefficients and entropy are at the segmental level, and prosodic features such as pitch contour, intensity and duration at the suprasegmental level provide information about dialects. These prosodic features along with spectral features provide discriminating information for dialect identification as well as speaker identification.

Table 4: Formant II Variations for Different Types of Words in the Database.

Words type	Male/female speakers	Malwi dialect	Majhi dialect	Doabi dialect
Toneme words	Female	1594.738	1586.196	1534.068
	Male	1690.307	1430.97	1622.056
Adhak words	Female	1649.454	1646.245	1594.833
	Male	1698.524	1533.741	1697.627
Nasal (bindi) words	Female	1527.631	1528.878	1493.46
	Male	1640.137	1392.773	1605.244
Nasal (tippi) words	Female	1563.765	1577.419	1490.581
	Male	1616.112	1431.992	1653.765

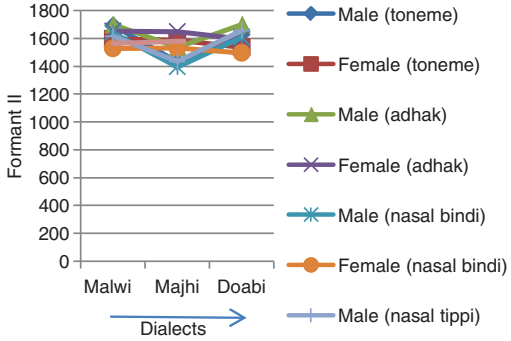

Figure 6: Dialect-Wise Formant II Variation for Tonemes, Adhak, Nasal (bindi) and Nasal (Tippi) Words.

8 Conclusions

In this paper, we have created text and speech corpora of about 12,500 words for Punjabi language. In India, there are various Punjabi dialects, but major dialects are Malwi, Majhi and Doabi. This corpus has been designed in context with these dialects. Tonal words have also been taken into consideration in this corpus. The existing database designs are also discussed and compared. This corpus is of quite importance in speech processing work due to its unique features, which makes this corpus a perfect corpus to develop language technology for Punjabi language. The present study also emphasized on the comparative analysis of prosodic features of Malwi, Majhi and Doabi dialects of Punjabi language. It has been found that the pitch of male speakers is very less than that of female speakers for all dialects. There is little variation in the intensity of male and female speakers. For all dialects, the formant I value of male speakers is greater than that of female speakers, but the formant II value of male speakers is greater than that of female speakers for Malwi and Doabi dialect and less for Majhi dialect. These variations are helpful in speaker identification, dialect identification and language identification.

Bibliography

[1] S. S. Agrawal and K. Samudravijaya, (Chief Editors), Text and speech corpora development in Indian languages, In: *Proceedings of the International Symposium on Speech technology and Processing Systems (ISTEPS-2004 and Oriental COCOSDA-2004)*; Vol. II, CDAC, New Delhi, India, pp. 21–27, November 17–19, 2004.

[2] S. S. Agrawal, K. Samudravijaya and K. Arora, Recent advances of speech database development activities, In: *International Symposium on Chinese Spoken Language Processing (ISCSLP 2006)*, 2006.

[3] S. S. Agrawal, S. Sinha, P. Singh and O. Jesper, Development of text and speech database for Hindi and Indian English specific to mobile communication environment, In: *Proceeding of International Conference on the Language Resources and Evaluation Conference*, LREC, Istanbul, Turkey, 2012.

[4] M. A. Anusuya and S. K. Katti, Front end analysis of speech recognition: a review, *Int. J. Speech Technol.* **14** (2011), 99–145.

[5] S. J. Arora and R. Singh, Automatic speech recognition: a review, *Int. J. Comput. Appl.* **60** (2012), 34–44.

[6] S. J. Arora and R. Singh, Acoustic and phonological analysis of homophones of Punjabi language, *Int. J. Comput. Sci. Eng. Inform. Technol. Res.* **4** (2014), 95–102.

[7] P. Bhaskararao, Sailent phonetic features of Indian languages in speech technology, In: *Sadhana Academy Proceedings in Engineering Sciences*, Indian Academy of Sciences, Bangalore, India, Volume 36, Number 5, pp. 587–599, October 2011.

[8] K. Cini, A survey on speech recognition in Indian languages, *Int. J. Comput. Sci. Inform. Technol.* **5** (2014) 6169–6175.

[9] S. Dhanjal and S. S. Bhatia, Computerization of the Punjabi language, In: *2nd World Punjabi Conference*, Punjab University, Chandigarh, February 24–25, 2009.

[10] S. Dhanjal and S. S. Bhatia, Punjabi Bhasha da Takneekee Bhavikh, In: *Silver Jubilee International Punjabi Development Conference*, Punjabi University, Patiala, February 3–5, 2009.

[11] S. Dhanjal and S. S. Bhatia, A new corpus for the Punjabi speech processing, In: *International Symposium on Frontiers of Research on Music and Speech (FRSM-2012)*, KIIT Gurgaon, India, pp. 223–227, January 18–19, 2012.

[12] M. Dua, R. K. Aggarwal, V. Kadyan and S. Dua, Punjabi automatic speech recognition using HTK, *Int. J. Comput. Sci.* **9** (2012) 359–364.

[13] H. Kaur and R. Bhatia, Speech recognition system for Punjabi language, *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **5** (2015) 566–573.

[14] K. Kumar and R. K. Aggarwal, Hindi speech recognition system using HTK, *Int. J. Comput. Bus. Res.* **2** (2011).

[15] S. Nareshkumar, N. Mariappan and K. Thirumoorthy, Database interaction using automatic speech recognition, *Int. J. Innov. Res. Sci. Eng. Technol.* **3** (2014) 1895–1899.

[16] K. Samudravijaya, P. V. S. Rao and S. S. Agrawal, Hindi speech database, In: *Proceeding International Conference on Spoken Language Processing (ICSLP00)*, Beijing, China, October 2000.