International Journal of Engineering Innovation & Research
Volume 3, Issue 3, ISSN: 2277 – 5668

Marathi Isolated Words Speech Database for Agriculture Purpose

Pukhraj P. Shrishrimal
Email: pukhraj.shrishrimal@gmail.com

Ratnadeep R. Deshmukh
Email: rrdeshmukh.csit@bamu.ac.in

Vishal B. Waghmare
Email: vishal.b.waghmare@ieee.org

***Abstract* – The research in the domain of the language technologies for Indian languages is far behind than the languages of developed nation. The work for the Indo-Aryan language, i.e. Marathi is behind. Development of speech database is the basic need for developing an automatic speech recognition system. The accuracy of speech recognition depends on the quality of the speech data collected and the quality of training set data. This paper describes the progress in the development of isolated words Speech database of Marathi language for agriculture purpose.**

***Keywords* – Speech Database, Speech Recognition, Marathi Language, Isolated Words, Speech Corpus.**

I. INTRODUCTION

There are different means of communication by which human communicate with each other such as writing, speech and Sign Language. The communication between human is dominated by speech. It is the most prominent and common way to pass message between human. There are number of languages that are spoken around the world. The humans have thought about use of speech as a mode of communication between human and computer since long time.

Speech has the potential of being used as a mode of interaction between human and computer. Human beings have long been motivated to create computer that can understand and talk like human. In this direction, researchers have tried to develop system for analysis and classification of the speech signals. Since 1960's the researchers are trying to develop system which can record, interpret and understand human speech [1].

The language technologies may be very useful for a developing country like India. The systems which can understand and interpret speech can prove very efficient in the field of agriculture, health, education and e-governance. The information is today's world is only accessible to those who are technologically literate and the information is in a specific language. The language technologies can be very useful to serve as a natural interface to access the digital content for those who are not having knowledge of the technology.

Hindi is the national language of India and there are 22 languages recognized by the constitution of India. Apart from that there are about 1652 dialects / native languages which are spoken throughout the country. The 23 languages recognized by the constitution of India are: 1) Assamese, 2) Bengali, 3) Bodo, 4) Dogri, 5) English, 6) Gujarati, 7) Hindi, 8) Kannada, 9) Kashmiri, 10) Konkani, 11) Maithili, 12) Malayalam, 13) Manipuri, 14) Marathi, 15) Nepali, 16) Oriya, 17) Punjabi, 18) Sanskrit, 19) Santali, 20) Sindhi, 21) Tamil, 22) Telugu, 23)Urdu [2].

For a multilingual country like India the language technologies can play a vital role. Most of the Indian languages are phonetic in nature. The national language of India i.e. Hindi along with one of the recognized language by constitution of India Marathi is written in devanagari script. If we see the global scenario of speech recognition systems a lot of work has been completed for English and various languages of developed nations around the world. Many research projects have been completed or are under progress for various languages [3].

There is a lot of scope for the development of speech recognition system in Indian languages. The work that is currently under progress is mostly for the national language Hindi later on for Tamil, Telugu Bangla, Assamese and Marathi. However the work for these languages is being carried under the linguistic data consortium for Indian languages (LDC-IL). They are working for development of continuous speech recognition systems [4]. The work for Marathi language is limited as the work is done mostly at IIT Bombay and TIFR, Mumbai.

This paper describes the work for the development of an isolated word speech database in Marathi language. The organization of the paper goes like Section II describes about the Marathi language. The section III describes the development of the text corpus. The details regarding the speech data collection is discussed in the Section IV. Section V describes the recording procedure followed and the problems faced during the development of database is explained in section VI. Section VII focuses on the removal of background noise. The conclusion and the future scope of the work are described in section VIII and IX respectively.

II. MARATHI LANGUAGE

Marathi is one of the 23 recognized languages by the constitution of India. It is written in devanagari script similar to the national language Hindi. The devanagari script is the script used for writing Sanskrit from which these languages are been derived.

Marathi is an Indo-Aryan language, spoken by the Marathi people of western and central India. There were 73 million speakers in 2001 around the world. Marathi has the fourth largest number of native speakers in India [5].

Marathi is spoken in the complete Maharashtra state which covers a vast geographical area which consists of 35 different districts. The major dialects of Marathi are called Standard Marathi and Warhadi Marathi [6]. The other few sub-dialects are like Ahirani, Dangi, Vadvali, Samavedi, Khandeshi and Malwani. However, standard Marathi is the

official language of Maharashtra state. Standard 'Marathi' language is based on dialects used by academics and the print media. The Indic scholars distinguish 42 dialects of spoken Marathi [7, 8].

The dialects bordering other major language areas have many properties in common with those languages, further differentiating them from standard spoken Marathi. The bulk of the variations within these dialects are primarily lexical and phonological (for e.g. accent placement and pronunciation). Although the number of dialects is considerable, the degree of intelligibility within these dialects is relatively high.

Marathi is the official language of Maharashtra and co-official language in the union territories of Daman and Diu [9] and Dadra and Nagar Haveli [10]. In Goa, Konkani is the sole official language; however, Marathi is also used for official purposes. In addition to all universities in Maharashtra, Maharaja Sayajirao University of Baroda (Gujarat), Osmania University (Andhra Pradesh), Gulbarga University (Karnataka), Devi Ahilya University of Indore and Goa University (Panaji) are having special departments for higher studies in Marathi linguistics [11, 12, 13].

III. Development of Text Corpus

The development of text corpus is described in the following section. For developing a speech database the basic need is grammatically correct text corpus which is used for recording the speech samples from various speakers. The text corpus should be correct in terms of typography and grammar. The text corpus for the developed database was developed by visiting various websites related to agriculture having content in Marathi. The words were selected from the various blog articles and forums which were published on the internet. In all hundred words were selected from the various articles, forums and websites [14].

Table I. Developed Text Corpus

Fruits	Grains	Vegetables	Fertilizers	Pesticides	Diseases	Equipments	Cash Crops
चीकू	हरभरा	पालक	शेणखत	निमआर्क	मावा	ट्रॅक्टर	कपाशी/ कापूस
आंबा	चवळी	मेथी	सेंद्रीय खत	मोनोक्रोटोफॉस	तुडतुडे	कापणीयंत्र	ऊस
पेरू	मका	लसूण	हरळीचे खत	मिथील डिमेटॉन	मुळकूज	औत / नांगर	हळद
अंजीर	ज्वारी	मिरची	युरिया	फॉस्फॅमिडॉन	तांबेरा	वखर / कुळव	सूर्यफूल
संत्रे	बाजरी	कोथंबीर	सुपर फॉस्फेट	डायमथोएट	हुमणी	रोपकयंत्र	सोयाबीन
सिताफळ	उडद	वांगी	सिंगल सुपर फॉस्फेट	बेनोमिल	करपा	टिकाव	भुईमुग
कलिंगड	मूग	कांदा	महाधन	कार्बेन्डॅझीन	भिरूड	फावडे	
पपई	मसूर	भेंडी		ट्रायकोडर्मा व्हीरीडी	भुरी	फराट	
मोसंबी	तुर	बटाटे		बावीस्टीन	पिठ्याढेकूण	पाटी / टोपले	
केळी	गहू	कोबी		डायथेन	केवडा	विळा / खुरपे	
डाळिंब		फुलकोबी		क्विनॉलफॉस	खोडमाशी	तिफण	
जांभूळ		भोपळा		लिंडेन	खोडकुज		
		कारली/ कारले		फोरेट	मर		
		वाल		सॅविडॉल	खुजा		
		शेपू		ऑक्झिक्लोराईड	काणी		
				सल्फर पावडर	सिगाटोका		
				कोयलोकोरस निग्रिटस	घाण्या रोग		
				मिथिल पॅराथिऑन	ठीपके		
				डी.डी.व्ही.पी.	व्हाईरस		
					बोकड्या		

The words selected from the websites were classified into different groups. We categorized the words in the following categories of fruits, grains, vegetables, fertilizers, pesticides, diseases, equipments and cash crops. The name of certain grains, cash crops and vegetables were such that they are called / known by more than one name. The developed text corpus contained the all the variations in names also. It was decided to use such combination because few people may speak those words

which are familiar to them. Due to this we can get more variation as at rural areas people still use certain words which are not know to urban people.

The developed text corpus consists of 100 words (isolated) which are specifically related to agriculture. The name of the pesticides in the text corpus was checked and only those pesticides which are approved by the government of India where selected for the text corpus. The words were checked for typographic error.

Table II. Isolated Words in Marathi Language in the Category of Fruits Along With its Transliteration and Ipa

Devanagari	Transliterated (Translated in English)	International Phonetic Alphabet (IPA)
चीकू	Sapodilla	/tʃəiku/
आंबा	Mango	/aməbəa/
पेरू	Guava	/pəerəu/
अंजीर	Fig	/əmdʒəirə/
संत्रे	Orange	/səəmt̪ərəe/
सिताफळ	Custard Apple	/səit̪əapʰələ/
कलिंगड	Watermelon	/kələigəɖə/
पपई	Papaya	/pəpəi/
मोसंबी	Sweet Lime	/məosəəmbəi/
केळी	Banana	/kəeɭəi/
डाळिंब	Pomegranate	/ɖəaɭəibə/
जांभूळ	Syzygium cumini	/dʒəaməbʰəuɭə/

The table I represent the complete text corpus developed. The table II and III represent the words selected for the development of the isolated word speech database in Marathi of category fruits and grains along with the respective transliterations and IPA (i.e. International Phonetic Alphabet).

Table III. Isolated Words in Marathi Language in the Category of Grains Along With its Transliteration and Ipa

Devanagari	Transliterated (Translated in English)	International Phonetic Alphabet (IPA)
हरभरा	Gram	/hərəbʰərəa/
चवळी	Beans	/tʃəuəɭəi/
मका	Maize, Corn	/məka/
ज्वारी	Sorghum bicolor	/dʒəuəarəi/
बाजरी	Pearl Millet	/bəadʒərəi/
उडद	Black Gram/ Black Lentil	/uɖəɖə/
मूग	Mung bean (green gram)	/məugə/
मसूर	Red Lentils	/məsəurə/
तुर	Pigeon Pea / Red Gram	/t̪əurə/
गहू	Wheat	/gəhəu/

IV. Speech Data Collection

This section describes the selection of speakers, data collection policy, recording procedure followed while collection of speech data and data statistics.

A. Selection of speakers

The speech data was collected from the native speakers of the Marathi language. The selected speakers were resident of the villages from Aurangabad district of Marathwada region of Maharashtra state. The speakers were comfortable with reading and speaking the language. The selection of the speaker was done such that it would cover the complete diversity i.e. age group, gender, literacy and language in which they generally speak.

One hundred speakers were selected from all over the district. The Speakers were in age groups ranging from 18-25, 26-32, 32-40, 40-55 and 55 and above.

B. Data Collection Policy

The speakers were asked to speak the selected 100 words with 3 utterance of each word. The Speech data was collected by visiting 10 villages from different Taluka places of Aurangabad district. At every village 10 speakers were selected to speak the words out of which 5 were male and 5 were female. From each village we were collecting speech samples of one male and one female in each age group. The speakers were selected on the basis of the educational qualification and their native language.

C. Data Collection Statistics

The speech data was collected from 10 speakers of 1 village from 10 Taluka places of Aurangabad district. Each speaker was asked to speak 100 words with 3 utterance of every word. Total 300 utterances of the isolated words were collected from every speaker. The database consist 300 utterances of each word. The developed speech database consists of total 30,000 from 100 speakers.

V. Recording Procedure Followed

The isolated words from developed text corpus were recorded from speakers using two different headsets. The headsets used were Sennheiser PC350 and PC360 which are different in terms of their technical specifications. The reason for selecting these specific headsets were that they are having noise cancelling facility and the distance of the both the microphone from mouth of speaker is same.

The data was recorded in normal environment. We used PRAAT software for recording the speech samples. The main strength of PRAAT is its graphical user interface, the functionalities like spectral analysis, pitch analysis, formant analysis, intensity analysis, other functionalities for drawing the cochleagram, spectrogram, speech signal plots and most important that it is open source [15].

The speech data was recorded with a sampling frequency of 16000 Hz, 16 bit in mono audio format. The files were saved with .wav file extension. As the data was recorded in a normal environment the recorded samples consisted of background noise which was later enhanced.

VI. Problems Faced in Database Development

The biggest problem faced during the research work was getting the information. The development of text corpus

was very time consuming and difficult for getting the correct information regarding crops, pesticides and diseases in Marathi language. After the development of the text corpus it was checked for the typographic error; as Marathi is phonetic in nature the typographic error would have resulted is wrong pronunciation of the word which was not acceptable during the database development.

While recording the speech samples it was very difficult to convince the speakers to spend 3 hours from their day to day life for collection of the speech samples. We have to adjust our recording session according to the time slots given by the male and female speakers and many times the time given by speaker would be less so we have to have multiple recording sessions with speakers.

During recording sessions we asked the speaker to speak a single word three times where we tried to capture the correct utterance. The other major problem faced was while speaking the name of pesticides. The speakers were unable to pronounce the name of few pesticides correctly so we have to ask them to repeat the utterance of the word until we got the correct pronunciation of the word.

VII. Speech Signal Enhancement

The recorded speech contained some background noise. The noisy speech samples were processed for the removal of noise. The speech signal enhancement is very important before we can extract the feature for developing the recognition system. The isolated Marathi speech database needs to have good speech samples without background noise. The speech samples should have good quality and good intelligibility for increased recognition accuracy. There are various speech signal enhancement techniques available like: spectral subtraction, subspace based algorithm, adaptive filtering technique (like LMS algorithm, RLS algorithm, Kalman filter) and adaptive comb filtering.

The spectral subtraction speech signal enhancement technique was implemented while enhancing the speech samples for the developed isolated words speech database.

In spectral subtraction, an average signal spectrum and average noise spectrum are estimated in parts of the recording and subtracted from each other, so that average signal-to-noise ratio (SNR) is improved [16]. It is assumed that the signal is distorted by a wide-band, stationary, additive noise, the noise estimate is the same during the analysis and the restoration and the phase is the same in the original and restored signal.

In the signal domain the model can be described as follows:

$$y(n) = x(n) + d(n) \quad (1)$$

Where, x is the speech signal, d is the noise and y noisy speech.

In the frequency domain the noisy speech model equation is expressed as:

$$y(j\omega) = x(j\omega) + d(j\omega) \quad (2)$$

Where, $y(j\omega)$, $x(j\omega)$ and $d(j\omega)$ are the Fourier transforms of the noisy signal y (n), x (n) and d (n) respectively.

As the statistic parameters of the noise are not known, thus the noise and the speech signals are replaced by their estimates:

$$\hat{x}(j\omega) = y(j\omega) - \hat{d}(j\omega) \quad (3)$$

The Noise spectrum estimate $\hat{d}(j\omega)$ is related to the expected noise spectrum E $[|\hat{d}(j\omega)|]$ which is usually calculated using the time-averaged noise spectrum $\hat{d}(j\omega)$ taken from parts of the recording where only noise is present. The noise estimate is given by:

$$\hat{d}(j\omega) = E\left[\left|d(j\omega)\right|\right] \cong \left|\bar{d}(j\omega)\right| = \frac{1}{K}\sum_{i=0}^{K-1}\left|d_i(j\omega)\right| \quad (4)$$

Where $|d_i(j\omega)|$ is the amplitude spectrum of the i^{th} of the K frames of noise. Noise estimate in k^{th} frame may be obtained by filtering the noise using the first order low-pass filter:

$$\hat{d}(j\omega) = \left|\tilde{d}_k(j\omega)\right| = \lambda_n \cdot \left|\tilde{d}_{k-1}(j\omega)\right| + (1-\lambda_n)\cdot\left|d_k(j\omega)\right| \quad (5)$$

Where $\tilde{d}_k(j\omega)$ the smoothed noise estimate in the i^{th} frame is λ_n, is the filtering coefficient. To obtain the noise estimate, the part of the recording containing only noise that precedes the part of containing speech signal should be analyzed [17].

The enhanced speech samples after the removal of background noise using spectral subtraction were stored separately. The original copy of the speech samples were retained after obtaining the noise free speech samples.

The figure 1 represents the waveform of the speech sample for the word आंबा having background noise. Figure 2 represents the spectrogram for the speech sample for the word आंबा having background noise. Figure 3 represents the waveform of the speech sample for the word आंबा after removal of the background noise. Figure 4 represents the spectrogram for the speech sample for the word आंबा after removal of the background noise.

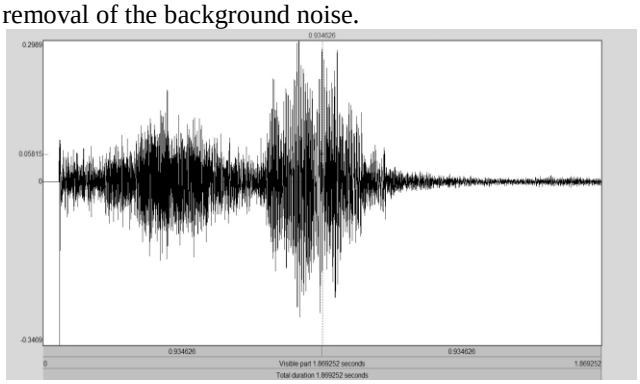

Fig.1. Waveform of the word आंबा having background noise

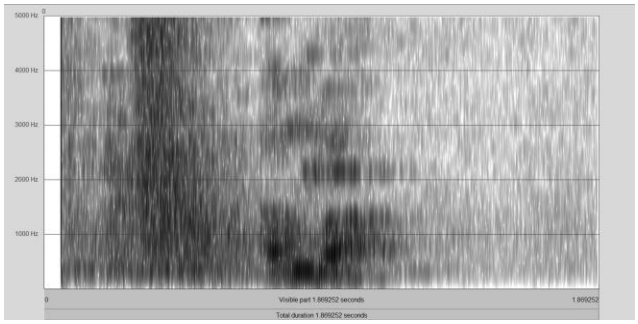

Fig.2. Spectrogram of the word आंबा having background noise

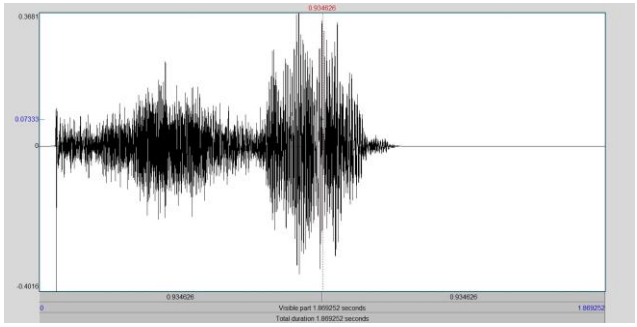

Fig.3. Waveform of the word *आंबा* without background noise

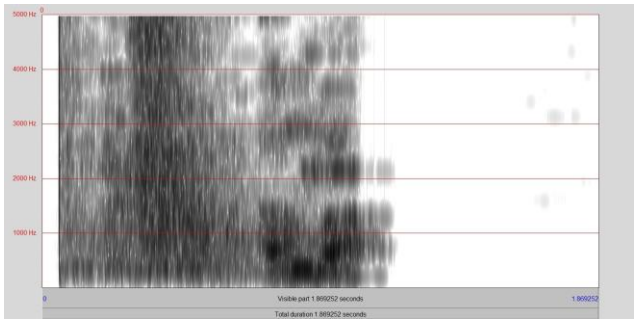

Fig.4. Spectrogram of the word आंबा without background noise

VIII. CONCLUSION

In this paper we have described the development of Marathi isolated words speech database for agriculture purpose. The said database may be useful to study the phonetic variations of Aurangabad district we tried to cover maximum phonetic variations. It will be helpful to develop a robust automatic speech recognition system for agriculture purpose. The developed database can be use for the development of isolated word recognition system for agriculture purpose.

ACKNOWLEDGMENT

This work is supported by University Grants Commission under the scheme Major Research Project entitled as "Development of Database and Automatic Recognition System for Continuous Marathi Spoken Language for agriculture purpose in Marathwada Region". The authors would also like to thank the University Authorities for providing the infrastructure to carry out the research.

REFERENCES

[1] Pukhraj P. Shrishrimal, R. R. Deshmukh, V. B. Waghmare, "Indian Language Speech Database: A Review", International Journal of Computer Application, Vol. 47 No.5, pp. 17-21, June 2012.

[2] Constitution of India, page 330, EIGHTH SCHEDULE, Articles 344 (1) and 351]. Languages.

[3] Sadaoki Furui, "50 Years of Progress in Speech and Speaker Recognition Research", ECTI Transaction on Computer and Information Technology, vol. 1, No. 2, pp. 64-74, Nov. 2005.

[4] Tejas Godambe and Samudravijaya K., "Speech Data Acquistion for Voice based Agricultural Information Retrieval", in Proc. of 39th All India DLA Conference, Punjabi University, Patiala, India, June 2011.

[5] "Abstract of Language Strength in India: 2001 Census". Censusindia.gov.in. Retrieved 2013-05-09.

[6] Dhoṅgaḍe, Rameśa; Wali, Kashi (2009). "Marathi". London Oriental and African language library (John Benjamins Publishing Company) 13: 101, 139. ISBN 9789027238139.

[7] K. Samudravija, "Multilingual Telephony Speech Corpora of Indian Languages", In Proceeding Computer Processing of Asian Spoken Languages. Eds, S. Itahashi Ans C. Tseng, Consideration Books Los Angeles, pp.189-193 (2010)

[8] Agrawal S. S., K.K. Arora, S Arora, Samudravijaya K, "Text and Speech Corpus Development in Indian Languages", ibid, pp. 94-97

[9] The Goa, Daman and Diu Official Language Act, 1987 makes Konkani the sole official language, but provides that Marathi may also be used "for all or any of the official purposes". The Government also has a policy of replying in Marathi to correspondence received in Marathi. Commissioner Linguistic Minorities, 42nd report: July 2003 - June 2004, pp. para 11.3

[10] Marathi is an official language of Dadra and Nagar Haveli Administration's profile. http://dnh.nic.in/deptdoc/vguide.pdf

[11] Dept. of Marathi, M.S. University of Baroda. Msubaroda.ac.in. Retrieved 2014-03-10.

[12] List of statutes http://www.dauniv.ac.in/rules/statute.doc. Retrieved 2014-03-10.

[13] Dept. of Marathi, Goa University. Unigoa.ac.in. Retrieved 2014-03-10.

[14] P. P. Shrishrimal, R. R. Deshmukh, V. B. Waghmare, "Development of Isolated Words Speech Database of Marathi words for Agriculture purpose", Asian Journal of Computer Science and Information Technology (AJCSIT), vol. 2, No. 7, pp. 217-218, July 2012.

[15] http://www.fon.hum.uva.nl/praat/ cited on 09/05/2012 (web reference)

[16] Philipos C. Loizou, Speech Enhancement: Theory and Practice, CRC Press. June 7, 2007.

[17] Saeed V. Vaseghi, Advanced Digital Signal Processing and Noise Reduction, Second edition, John Wiley & sons Ltd, pp. 333-335

AUTHOR'S PROFILE

Mr. Pukhraj P. Shrishrimal

Currently pursuing Ph.D. in Computer Science, Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. Completed M.Phil. (Computer Science), M.Sc. (Compu ter Science) from Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University,

Aurangabad in 2013 and 2010. Completed B.Sc. (Computer Science) from Vivekanand College, Aurangabad in 2008.
Currently working as CHB Faculty and Project Fellow on a Major Research Project Sanctioned by University Grants Commission, New Delhi in Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad.
Mr. Pukhraj P. Shrishrimal is Life Member of Indian Science Congress Association (ISCA), International Association of Engineers (IaENG), Computer Science Teacher Association (CSTA), International Association of Computer Science and Information Technology (IACSIT). He has published more than 10 research papers in various International Journals, International and National Conferences. He is IBM Certified DB2 Fundamental Database Associate
His are of Specialization Digital Signal Processing, Speech Recognition, Computational Auditory Scene Analysis (CASA), Advance Database Management System, Data Warehousing, Data Mining and Human Computer Interaction (HCI).

Dr. R. R. Deshmukh

M.E. (CSE), M.Sc. (CSE) Ph.D. FIETE, Presently working as Professor in Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, MS-India.
He is a Member of Management Council of Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, MS-India. He is Chairman of Ad-Hoc Board of Studies in Computer Science and IT and Bioinformatics, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, MS-India. He is Member of Ad-Hoc Board of Studies in Computer Engineering of Mumbai University, Mumbai. He is Member of Board of Study Computer Science, Solapur University, Solapur.
He is a Fellow of IETE, Senior member of IEEE, Life member of ISCA, CSI, ISTE, ACEEE, IAEng, CSTA and IDES. He has edited of nine books and published more than 70 research papers in reputed Journals, National and international conferences. He is reviewer and editor of several journals at national & international level. He has organized several workshops and conferences. He is nominated as a subject expert on various academic & professional bodies at national level government bodies. He is Faculty member for Engineering, Science & Management Faculty & Member of various committees at university level.
He has two research project projects from UGC and received grants more than 10 Lakhs. His areas of specialization are Human Computer Interaction (HCI), Data Mining, Data Warehousing, Image Processing, Pattern Recognition, Artificial Intelligence, Computational Auditory Scene Analysis (CASA), Neural Networks etc. He won First prize in Inter University State Level Research Festival "AVISHKAR - 2009" under H. L. F. A. category at Teacher level & for the Team Management.

Mr. Vishal B. Waghmare

Currently pursuing Ph.D. in Computer Science, Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. Completed M.Sc. (Computer Science) from Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad in 2008. Completed B.Sc. (Computer Application) from Yogeshwari College, Ambajogai, Beed in 2006. Currently working as CHB Faculty in Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad.
He has received Rajiv Gandhi National Fellowship in the year 2010 and is currently a Senior Research Fellow.
Mr. Vishal B. Waghmare is Student member of IEEE, Life Member of Indian Science Congress Association (ISCA), International Association of Engineers (IAEng), Computer Science Teacher Association (CSTA), International Association of Computer Science and Information Technology (IACSIT). He has published more than 10 research papers in various International Journals, International and National Conferences. His area of Specialization is Digital Signal Processing, Speech Recognition, Web Technologies and Computational Auditory Scene Analysis (CASA).