Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Implementation of a TTS System for Devanagari Konkani Language using Festival

Nilesh B. Fal Dessai
Department of Computer Science & Technology
Goa University, Taleigao Plateau
Goa, India

Gaurav A. Naik
Info Tech Corporation of Goa Limited
IT Hub, Altinho, Panaji
Goa, India

Jyoti D. Pawar
Department of Computer Science & Technology
Goa University, Taleigao Plateau
Goa, India

Abstract: Text to Speech (TTS) Synthesizer is an application that converts text to speech. Development of speech synthesis system is a challenging task as the input text may come in an ambiguous form, different words are pronounced in different ways thus requiring efforts during text pre-processing. This paper discusses the various aspects of Festival and Festvox framework in Linux environment and its use for the implementation of a TTS system for Devanagari Konkani language. Festival does not provide complete language processing support to various languages. The experimental results with a text segment of 100 Konkani sentences shows that 64% word phonetization accuracy is obtained thus indicating scope for improvement in the quality of the output speech if segment of voice used are higher than unit selection voices.

Keywords: TTS; Speech Synthesis; Devanagari; Konkani; Festival; Festvox

I. INTRODUCTION

India is a Multi-lingual country with variety of scripts and hundreds of spoken dialects. It is desired that information along with ICT based services are delivered to a large portion of the population in their own language in the form of voice. Lot of research work is currently carried out in the area of text to speech processing for many Indian languages and the synthesis systems are in great demand for Indian languages. The most important quality of a speech synthesis system is naturalness and intelligibility. Naturalness describes how closely the output sounds like human speech, while intelligibility is the ease with which the output is understood. An ideal speech synthesizer is both natural and intelligible and hence speech synthesis systems usually try to maximize both the characteristics [1]. Intelligibility of the output speech has now reached an adequate level for most applications, especially for the visually challenged and illiterate masses.

Konkani, the official language of the State of Goa in India is also the minority language in the States like Karnataka, Kerala and Maharashtra in India. Konkani is being spoken by about 3.6 million people and is written in both Devanagari and Roman script.

No concrete work is carried out for Konkani in the area of text to speech. The focus of this work is to study the tools, resources and techniques for text to speech processing that have been developed for Indian languages and to implement a TTS system for Devanagari Konkani using Festival in Linux environment. Festival is widely used for the implementation of TTS system for many languages [2] [3] [4].

The paper is organized as follows: Section II and III outline the TTS system and the synthesis techniques. Section IV, V and VI details the Festival framework, its implementation for Konkani and evaluation & discussion of the implemented system respectively, followed by conclusion at the end

II. GENERAL FUNCTIONAL DIAGRAM OF A TEXT – TO – SPEECH SYSTEM

The general architecture of a corpus-based TTS system is depicted in Figure 1. Speech synthesis mainly uses two processing components; the NLP (Natural Language Processing) and the DSP (Digital Signal Processing) modules [5] [6].

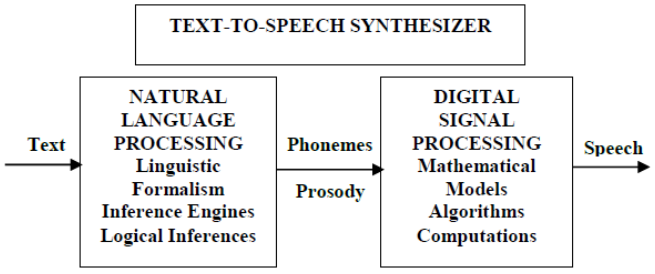

Figure 1: The general architecture of a corpus-based TTS system

This schematic applies for every data driven (i.e. any corpus-based) TTS system, regardless of the underlying technology (e.g., unit selection or parametric). The NLP component accounts for every aspect of the linguistic processing of the input text, whereas the DSP component accounts for the speech signal manipulation and the output generation. For a unit selection TTS, besides the speech units (usually diphones) the speech database contains all the necessary data for the unit selection stage of the synthesis.

In particular, the NLP component is mainly responsible for parsing, analysing and transforming the input text into an

intermediate symbolic format, appropriate to feed the DSP component. Furthermore, it provides all the essential information regarding prosody, i.e. pitch contour, phoneme durations and intensity. It is usually composed of a text parser, a morpho-syntactic analyzer, a text normalizer, a letter-to-sound module and a prosody generator. All these components are essential for disambiguating and expanding abbreviations and acronyms, for producing correct pronunciation and also for identifying prosody related anchor points.

The DSP component includes all the essential modules for the proper manipulation of the speech signal, i.e. prosodic analysis and modification, speech signal representation processing and generation. The DSP component also includes the unit selection module, which performs the selection of the speech units from the speech database using explicit matching criteria.

III. SPEECH SYNTHESIS TECHNIQUES

There are several methods available to synthesize speech with their respective strengths and weaknesses and suits a specific language while does not suit others [7][8]. The commonly used three methods are:

A. Articulatory synthesis

This method is derived from human speech generation and refers to computational techniques for synthesizing speech based on models of the human vocal tract and the articulation processes occurring there. The name of the method is inspired by the term "Articulators" which implies speech organs like jaw, tongue, lips etc.

B. Concatenative synthesis

This method is based on the concatenation (or stringing together) of segments of recorded speech to produce natural speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech. However, differences between natural variations in speech and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches in the output. The three main sub-types of concatenative synthesis are: unit synthesis, diphone synthesis and domain specific synthesis. One of the most important aspects of concatenative synthesis is to the find the correct unit length. [9].

C. Formant synthesis

Formant synthesis does not use human speech samples at runtime. Instead, the synthesized speech output is created using additive synthesis and an acoustic model (physical modeling synthesis). Parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech [10]. This method is sometimes called as rules-based synthesis. However, many concatenative systems also have rules-based components. Many systems based on formant synthesis technology generate artificial, robotic-sounding speech that would never be mistaken for human speech. However, maximum naturalness is not always the goal of a speech synthesis system, and formant synthesis systems have advantages over concatenative systems.

A brief comparison of the speech synthesis techniques is shown in the table I.

Table I. Comparison of Synthesis Techniques

.	*Synthesis Techniques*		
	Articulatory	***Concatenative***	***Formant***
Naturalness	Excellent	Excellent	Satisfactory (Robotic)
Speed of Processing	Moderate	Slower	Faster
Implementation Ease	Difficult	Easier	Easy
Challenging Aspects	Critical in operation of articulator and vocal cords	Choice of unit and space required for its storage	Set of parameters controlling speech

Thus we infer that concatenative synthesis which uses recorded segments of sound units is the best suited method to generate natural and intelligible utterance. In concatenative synthesis the selection of unit length is usually a trade-off between longer and shorter unit segments. With longer units high naturalness, less concatenation points and good control of co-articulation is achieved, but the amount of required units and memory is more. With shorter units, less memory is required, but the sample collecting and labeling procedures become more difficult and complex [1].

IV. FESTIVAL FRAMEWORK

Festival offers a general framework for building speech synthesis systems as well as including examples of various modules [11]. Festival is a speech synthesis system and is developed in CSTR (Center for Speech Technology Research), university of Edinburgh. Festival is compatible to work with all types of voices and supports different platforms. It is an open source framework written in C++, uses the Edinburgh Speech Tools Library for low level architecture and has a scheme based command interpreter for control [12].

The Festival speech synthesis system consists of different modules and they together produce the synthetic speech. The modules are: Tokenization and pronunciation generation, prosody generation and Waveform generation and are depicted in Figure 2.

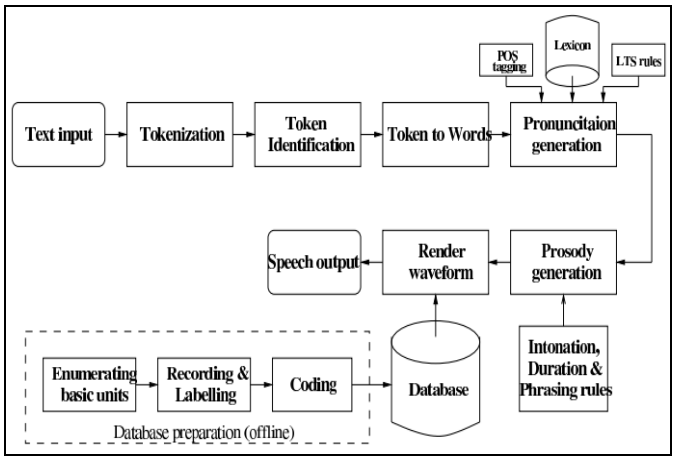

Figure 2: Festival Architecture

In Festival, utterance plays an important role in the generation of synthetic speech. This framework takes the utterance and each of the modules present, manipulates it in some way or the other and passes it to the next module.

Utterance consists of a set of items which are related through a set of relations. Each relation consists of a list or tree of items. Items are used to represent objects like words or segments. Relations are used to link items together in a useful way. An item may have one or more relations.

A. Tokenizaton and Pronunciation Generation

This module carries out the analysis and converts the raw text into acceptable format that can be processed in a reasonable manner. The text analysis block takes the raw input text and produces the pronounceable format. Here all the abbreviations and numbers are expanded with respect to the context in the given text. Following are some of the important steps of the tokenization and pronunciations generation module.

1) Identifying Tokens

The text is converted to tokens depending on the white spaces and punctuation marks. Whitespaces can be viewed as separators and punctuation can also be separated from the raw tokens. Festival converts text into an ordered list of tokens, each with its own preceding whitespace and succeeding punctuation as features of the token.

2) Normalization of Non-Standard Words

In the given input text, all the words which are available in the dictionary are called standard words. Numbers, symbols, abbreviations, etc. which are not available in the dictionary for their pronunciations are called as non-standard words (NSW). NSW can be identified as a separate token by the token identifier rule. Table II shows examples of Konkani NSW and its pronunciations. To identify the token we can use scheme regular expression in Festival

Table II. NSW and its pronunciation

NSW Category	*Written format*	*Pronunciation*	*Translation*
Cardinal Number	९८५०४०९५७९	णव आठ पाच शून्य चार शून्य णव पाच सात णव	*Naav Aath Paach Shunya Char Shunya Naav Paach Saat Naav*
Ordinal Number	७३४	सातशे चौतीस	*Saatshe Chautis*
Decimal	२७.५	सत्तावीस पूंज पाच	*Sattavis Pungh Paach*
Date	०२/०६/२००७	दोन जून दोन हजार सात	*Don June Don Hazzar Saat*
Time	९.२५	णव वरा पंचवीस मिनटां	*Naav Vara Panchvis Minta*
Special Character	रू.	रूपया	*Rupaya*
Abbreviation	चि.	चिरंजीव	*Chiranjiv*

These words are converted to full pronunciation with the following stages.

- Splitter: It will split the token not only with the white space but also with the punctuation.
- Type identifier: It will identify the token type for expansion.
- Token expander: The identified token is expanded depending on the context.
- Language modeling: Is used to select between possible alternative punctuations of the output.

B. Prosodic Generation

The second stage of Festival speech synthesis is linguistic and prosodic processing. The input to this stage is pronounceable words. To convert these pronounceable words into segments with prosody, the system requires phones, durations and tune (F0 contour). The appropriate phone symbols for input words are extracted from the lexicons, which are available in dictionary. Lexicon consists of list of words with their corresponding phone symbols. The words which are not available in dictionary or lexicon list follow the letter to sound rules to extract the phone symbols. Letter to sound rules are very difficult to write, but they are more powerful in the generation of phones from the words.

In Festival, prosodic phrasing can be carried out by two methods, one is phrasing by decision tree and other is phrasing by statistical models. Intonation, durations and post-lexical rules are together called as prosody. Prosody plays an important role in the generation of natural speech as an output. Intonation is nothing but accent and F0 contour. These two parameters are extracted from the existing voice models and these intonation parameters decide the speaker and energy of the output speech.

C. Waveform Generation

This module is the final and most important part of the Festival speech synthesis system. It receives phone information, prosody for synthesis from previous block and existing voice models. By combining all these parameters, it will produce synthetic speech as output. Depending on the voice models, the waveform synthesizer differs in terms of access to the relevant and required information from the voice models to produce synthetic speech.

V. IMPLEMENTATION

In order to implement the TTS system for Devanagari Konkani language, the Festival system has to be trained with the Konkani text and voice to understand the issues and challenges and then accordingly train the system. The system lack few language specific issues and are discussed further in this section.

A. Konkani Phonology

Konkani is an Indo-Aryan language belonging to the Indo-European family of languages and is spoken along the South western coast of India. It is one of the 22 scheduled languages mentioned in the 8th schedule of the Indian Constitution and the official language of the Indian State of Goa [13].

Konkani is written in five scripts: Devanagari, Roman, Kannada, Malayalam, and Perso-Arabic. As shown in table III and IV, the Devanagari Konkani has 12 basic vowels, 36 consonants, 5 semi-vowels, 3 sibilants, 1 aspirate, and many diphthongs like the other Indo-Aryan languages, it has both long and short vowels and syllables with long vowels stressed. Different types of nasal vowels add special features to the Konkani language.

Table III. Konkani Vowels

अ	आ	इ	ई	उ	ऊ
a	*ā*	*i*	*ī*	*u*	*ū*
ए	ऐ	ओ	औ	अं	अः
e	*ai*	*o*	*au*	*aṃ*	*aḥ*

Table IV. Konkani Consonants

क	ख	ग	घ	ङ
ka	*kha*	*ga*	*gha*	*ṅa*
च	छ	ज	झ	ञ
ca	*cha*	*ja*	*jha*	*ña*
ट	ठ	ड	ढ	ण
ṭa	*ṭha*	*ḍa*	*ḍha*	*ṇa*
त	थ	द	ध	न
ta	*tha*	*da*	*dha*	*na*
प	फ	ब	भ	म
pa	*pha*	*ba*	*bha*	*Ma*
य	र	ल	व	
ya	*ra*	*la*	*va*	
ष	श	स	ह	
ṣa	*śa*	*sa*	*ha*	
ळ	क्ष	ज्ञ		
ḷha	*kṣa*	*jña*		

B. Installation Details

1) Linux Platform:

For our system implementation, we have used BackTrack 5 R2, which is an open source Linux distribution. BackTrack provided users with easy access to a comprehensive and large collection of security-related tools. It supports live CD and live USB functionality to allow users to boot BackTrack directly from portable media without requiring installation, though permanent installation to hard disk is possible[14].

2) Installation of Tools:

Edinburgh Speech Tools provides a set of executables, which offer access to speech tools functionality in the form of a standalone program. As far as possible, the programs follow a common format in terms of assumptions, defaults and processing paradigms. Some of the common features of these tools are arguments to functions which can be specified in any order in the command line.

Installation of Festival speech synthesis system tool requires the source of Festival framework and speech tools. Festvox project is another framework to generate the new voice models from the recorded speech database developed at CMU (Carnegie Mellon University) [1][15].

To generate voice models and use these models to synthesize the input text for Indic speech database, current versions of Festival, speech tools and festvox frame work are used. The latest version of Festival and speech tools is available at http://www.cstr.ed.ac.u and the latest version of festvox is available at http://www.festvox.org. We have used the following tools for the implementation of the TTS system:

- festival-2.4-release.tar.gz
- speech _tools 2.4.tar.gz
- festvox-2.7.0-release.tar.gz

Next step is to install a simple test suite with Festival which requires the three basic voices and their respective lexicons pre-installed for it to work. These are available at:

- festvox_kallpc16k.tar.gz
- festlex_POSLEX.tar.gz
- festlex_CMU.tar.gz

Next, extract Festival, speech_tools, and festvox to your home directory and set up the environment variables.

3) Recording and Labeling:

The selected text is recorded with the voice of a native male speaker of Konkani language. The recording is carried out in a studio environment with following characteristics:

- Sampling Rate: 48000 Hz
- Bit Depth: 16 bit
- Channels: Mono

The total recording time for 250 sentences is about 1.20 hrs. After recording the selected sentences, the next phase is to label the syllable sounds in the recorded sound file. This is one of the most important and time consuming task and needs to be done very carefully. For this purpose we have used sound editing software Wavesurfer 1.8.8 and the sentences have been labeled manually one by one after carefully listening and analyzing the word sounds [16].

Assuming that Festival works perfectly, we then build the voices using festvox. The format to pass data is shown in table V:

Table V. Sample Training Data

Sr. No.	Sentence in Konkani
1	(data_00001 "*रमाबाय पणजे शारांत वाडिल्ली.*") (data_00001 "Ramabai Panaje Sharat Vadilee.")
2	(data_00002 "*शारांचें वातावरण तिजेर परिणाम करूंक पावलें नाशिल्लें.*") (data_00002 "Sharache Vatavaran Tejer Parinam Karunk Pavle Nashelle.")
3	(data_00003 "*तांचीं कारणांय उणी नाशिल्लीं.*") (data_00003 "Tachi Karnai Une Nashelle.")
4	(data_00004 "*राज्य गोंय्यांचें.*") (data_00004 "Rajya Goranche.")
5	(data_00005 "*तातूंत आनी राजधानयेचें शार.*") (data_00005 "Tatut Ani Rajdhanyeche Shar.")

The format is "(" followed by a filename, root followed by the text for that sentence, followed by a ")" each on separate line. This text when converted to a phone sequence by Festival should match (as closely as possible) the phone sequence of the speech.

VI. EVALUATION AND DISCUSSION

Considering that the writing script for Konkani and Marathi is Devanagari, and the availability of enough Marathi speech data for training, the system was first tested for a sample text of approximately 100 sentences of Konkani language trained on Marathi voice data. However it was observed that the generated Konkani speech was more towards the Marathi accent. Hence we used Konkani speech training data to implement the TTS system using Festival Framework.

Voice quality testing is performed using subjective test. In subjective tests, human listeners of different age groups e.g. college students and teachers, are made to listen and rank the quality of processed voice files according to a certain scale. The most common scale is called MOS (Mean Opinion Score) and is composed of five scores of subjective quality, 1- Bad, 2- Poor, 3- Fair, 4- Good, 5- Excellent [17]. The MOS score is the average of all the scores voted by different listeners for the different voice file used in the experiment.

The system is implemented for Marathi and Konkani trained text & voices. The MOS depicting the output Konkani speech obtained using Marathi and Konkani voice training respectively is graphically represented in Figure 2.

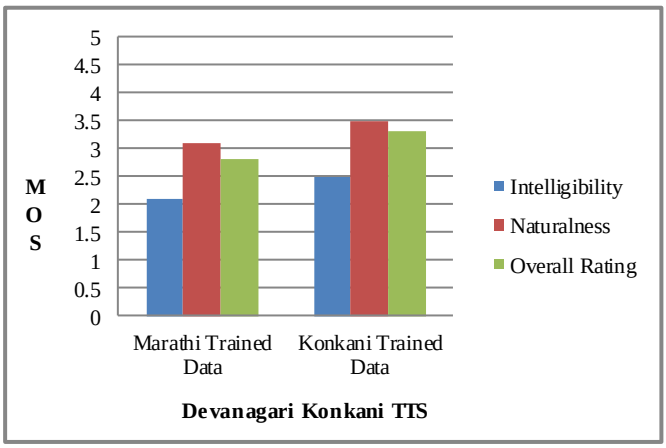

Figure 3: MOS Ratings

Based on the outcome of the listeners test, it is observed that TTS system using Festival for Konkani generates an acceptable output for synthesis. The Konkani output speech generated with Marathi voice data training was with Marathi accent. On analyzing the test cases and listener's feedback, in terms of some Konkani Phonology rules when applied to certain types of sentences [18][19][20], following are the some of the important observations:

- Recorded Syllables are best suited for generation of speech as compared to the speech obtained from syllables formed by concatenation diphones/ phonemes.
- Syllables formed using diphones/phonemes did not differentiate for presence of inherent schwa for consonants. For example in the word 'घर' (ghar) is spoken as 'घर'(ghar**a**). All consonants are uttered straight forward.
- Syllables formed using diphones/phonemes do not recognize or classify words for presence of nasal sounds represented by 'ं' over consonant or vowel. For example in the word 'झुंबर' (jhu**m**ber), the nasal sound is represented as 'म्' (m). e.g. झुंबर → झु+म्+बर
- Syllables formed do not differentiate the effect of vowel harmony for specific words. For example the phrase 'ती तेंफळ' uttered as 'ती तेंफळ' (tee teefal) represents the tree and the phrase 'तें तेंफळ' uttered as 'तें तॅफळ' (te tefal) represents the fruit of the tree. It is significant to note that, the word is written as 'तेंफळ' in both the phrases.
- Syllables formed using diphones/phonemes do not consider diphthong wherein sound units of type इव (Ev), इय (Ey), उव (Uv), उय (Uy), एव (Iv), एय (Iy), ओव (Ov) and ओय (Oy) are differently uttered. For example in the word 'दिवज' (divaj), there is stress on the first vowel and 'व' (V) remains silent, resulting in the utterance of sound unit 'दिवज् (इव)'
- Not adequately considered for sentence and words level stress as prosody control in synthesis.

Thus in context of the above observations, we infer that in order to obtain better naturalness, the synthesizer requires addressing the language specific issues like phonology for implementation in Festival.

VII. CONCLUSION

A speech synthesis system is designed and implemented for Devanagari Konkani language using Festival. The output synthesized speech exhibits quality and naturalness based on the subjective quality test results. The system can be implemented and tested for more Devanagari languages to draw comparative results. Further, the voice quality can be improved by implementing language specific phonological rules.

VIII. REFERENCES

[1] M. Nageshwara Rao, Samuel Thomas, T. Nagarajan, and Hema A. Murthy, "Text-to-Speech Synthesis using syllable-like units", Proceedings of National Conference on Communication (NCC), 2005.

[2] S. Kayte, B. Gawali, "A Text-To-Speech Synthesis for Marathi Language using Festival and Festvox", International Journal of Computer Applications (0975 – 8887), Volume 132 – No.3, December 2015.

[3] F. Y. Sadeque, S. Yasar, M. Islam, "Bangla Text to Speech Conversion: A Syllabic Unit Selection Approach", 978-1-4799-0400-6, 2013.

[4] Abu Naser, Devojyoti Aich, Md. Ruhul Amin, "Implementation of Subachan: Bengali Text To Speech Synthesis Software", 6th International Conference on Electrical and Computer Engineering, ICECE 2010, 18-20 December 2010, Dhaka, Bangladesh.

[5] M. Waseem, C. N. Sujatha, "Speech Synthesis System for Indian Accent using Festvox", International Journal of Scientific Engineering and Technology Research, Issue 34, November 2014, Pages: 6903-6911.

[6] N. S. Krishna, H. A. Murthy, T. A. Gonsalves, "Text-to-Speech in Indian Languages", Proceedings of International Conference on Natural Language Processing, ICON 2002, Mumbai, pp. 317.326.

[7] Fatima Chouireb M. G., "Towards a high quality Arabic speech synthesis system based on neural networks and residual excited vocal tract model", Springer 2008.

[8] Lemmetty S., "Review of Speech Synthesis Technology", Master Thesis, Helsinki University of Technology, 1999.

[9] P. Chaudhury, M. Rao, K. V. Kumar, "Symbol based concatenation approach for Text to Speech System for Hindi using vowel classification technique", World Congress on Nature and biologically Inspired Computing, pp.1082 – 1087, 2009.

[10] A. Chauhan, V. Chauhan, S. P. Singh, A. K. Tomar, H. Chauhan, "A Text to Speech System for Hindi using English Language", International Journal of Computer Science and Technology, Vol. 2, Issue 3, pp. 322-326, 2011.

[11] S. N. Kayte, M. Mundada, C. Kayte, "Speech Synthesis System for Marathi Accent using FESTVOX", International Journal of Computer Applications (0975 – 8887), Volume 130 – No.6, November 2015.
[12] The Festival Speech Synthesis System http://www.cstr.ed.ac.uk/projects/festival/
[13] Konkani Langauge https://en.wikipedia.org/wiki/Konkani_language
[14] Back Track Linux http://www.backtrack-linux.org/downloads/
[15] F. Alam, "Text To Speech Synthesis for Bangla language", Thesis Submitted to the Department of Computer Science of BRAC University.
[16] Wavesurfer Tool https://en.wikipedia.org/wiki/WaveSurfer
[17] P. C. Loizou, "Speech Quality Assessment", University of Texas-Dallas, Department of Electrical Engineering, Richardson, TX, USA.
[18] Shanai Goibab, "Konkani Nadshahtra", 1940, Vol. 1, pp 392-439.
[19] Priyadarshani Tadkodkar, Asha Mandgutkar, "Konkani Parichai", pp 5-21.
[20] V. M. Dhume,"Konkani Shudhlekanache Nhem-Goa Konkani Academy", pp 6-11.