IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 5, Issue 5, Ver. I (Sep - Oct. 2015), PP 76-81
e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
www.iosrjournals.org

Di-phone-Based Concatenative Speech Synthesis Systems for Marathi Language

Sangramsing Kayte[1], Monica Mundada[1], [2]Dr. Charansing Kayte
Department of Computer Science and Information Technology
Dr. Babasaheb Ambedkar Marathwada University,Aurangabad
[2]Department of Digital and Cyber Forensic, Aurangabad Maharashtra, India

Abstract:*The main objective of this paper is to provide a comparison between two di-phone-based concatenative speech synthesis systems for Marathi language. In concatenative speech synthesis systems, speech is generated by joining small prerecorded speech units which are stored in the speech unit register. A di-phone is a speech unit that begins at the middle of one phoneme and extends to the middle of the following one. Di-phones are commonly used in concatenative text to speech (TTS) systems as they have the advantage of modeling co-articulation by including the transition to the next phone inside the unit itself. The first synthesizer in this comparison was implemented using the Festival TTS system and the other synthesizer uses the MARY TTS system. In this comparison, the differences between the two systems in handling some of the challenges of the Marathi language and the differences between the Festival TTS system and the MARY TTS system in the DSP modules are highlighted. Also, the results of applying the diagnostic rhyme test (DRT) on both of the synthesizers are illustrated.*

Keywords:*-Text-to-Speech Di-phone-based Concatenation, Festival TTS, MARY TTS, Digital Signal Processing Diagnostic Rhyme Test*.

I. Introduction

Text-to-speech synthesis enables automatic conversion of a sequence of type-written words into their spoken form. This paper deals with text-to-speech synthesis of Marathi language. A few attempts have been made in the past to cover different aspects of a possible TTS system for Marathi language [1][2]. However, no-one has succeeded in building a complete system providing high quality synthesized speech. We have worked on text-to-speech synthesis for a year. Here we describe the current version of our Marathi TTS system. Our improvements in the future will be based on this system. The synthesis task is performed here through the following two steps: analyzing text and producing speech. Each of these steps includes several modules, operating sequentially, as shown in Fig.1. At the first step, input text is normalized in the Text processing module. The tasks of this module cover sentence tokenization, non-standard words and homograph disambiguation. In the phonetic analysis module, letter-to-sound rules are used for finding pronunciations of the normalized words. Their intonation, which includes accent, boundaries, duration and F0 is produced in the Prosodic analysis module. At the second step, the synthesizer creates a speech waveform from the complete phonetic and prosodic description. The di-phone concatenative synthesis is used to generate a waveform from a sequence of phones by selecting and concatenating units from a prerecorded database of di-phones. At the end of the paper, results of evaluation of the system are given and discussed and some promising directions forDiphone synthesis techniques require less database as compared to the unit selection synthesis. It uses two adjacent phones to make the speech waveform. But this techniques suffers through the problem of coarticulation. Diphone synthesis is one of the most popular methods used for creating a synthetic voice from recordings or samples of a particular person; it can capture a good deal of the acoustic quality of an individual, within some limits.

II. Text Analysis

The General block diagram of concatenative text-to-speech synthesis is illustrated in Fig. 1.

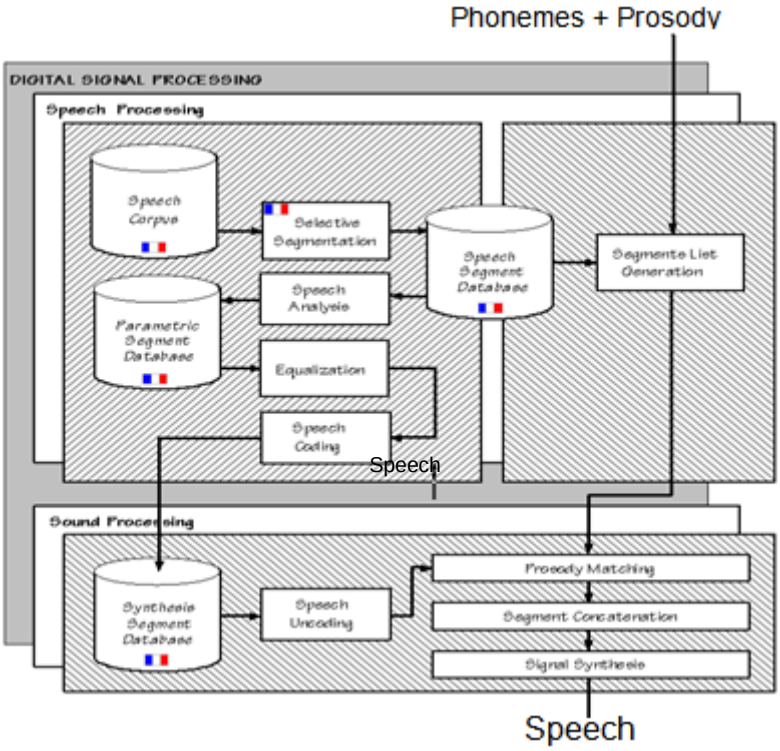

Figure 1: General block diagram of concatenative text-to-speech Synthesis

2.1 Text Normalization

The first step in all text-to-speech systems is preprocessing or normalizing the input text in many ways. First, the input text has to be divided into sentences. Then each sentence has to be divided into a collection of tokens (such as words, numbers, dates and other types). Non-natural language tokens such as abbreviations and acronyms must be transformed to natural language tokens. In the following subsections, we will explain the steps of text normalization in more details [3].

2.1.1 Sentence Tokenization

The first mission in text normalization is sentence tokenization. This step has some complications because sentences terminations are not always designated by periods and can sometimes be designated by other punctuations like colons. To identify sentence boundaries, the input text is divided into tokens separated by whitespaces and then any token including any of these characters (! , . , or ?) is chosen and a machine learning classifier can be employed to find out whether each of these characters inside these tokens indicate an end-of-sentence or not [4].

2.1.2 Non-Standard Words

The second mission in text normalization is normalizing non-standard words such as numbers, dates, abbreviations or acronyms. These tokens have to be transformed to a sequence of natural words so that a synthesizer can utter them in the proper way. The complexity with non-standard words is that they are often ambiguous. For example, abbreviations and acronyms must be expanded with the assistance of an abbreviation dictionary [5].

2.1.3 Homograph Resolution

The final mission in text normalization is homograph resolution. Homographs are words that have the same sequence of characters but differ in their pronunciation. For example, the two forms of the word "use" in the subsequent sentence " पुणेशहरातीलएकमध्यवर्तीठिकाण", have different accents. The proper pronunciation of each of these forms can easily be identified if the part-of-speech is known. The first form of the word use in the earlier sentence is a noun whereas the second one is a verb. Liberman and Church showed that the part-of-speech can disambiguate many homographs in 44 million words . A word sense disambiguation algorithm can be used to resolve homographs in the cases in which homographs cannot be resolved by their part-of-speech [6].

2.2 Accent

The next step after normalizing the input text is to find the proper accent for each word. The most important component in this phase is a large accent lexicon. The accent lexicon alone is not sufficient, because the input text can include words such as names that cannot be found in the lexicon. For this reason, many text-to-speech systems utilize a name- accent lexicon in addition to the principal accent lexicon [7]. Since the pronunciation of many names can be produced by analogy, this name- accent lexicon is not needed to be very large. The accent of unknown words that are not found in the accent lexicon can be produced through the use of grapheme-to-phoneme conversion methods [8][16][17].

2.2.1 Grapheme-to-Phoneme Conversion

A grapheme-to-phoneme algorithm generates a sequence of phones from a sequence of characters. The earliest of G2P algorithms were rule based techniques. These are called letter-to-sound rules. The results of LTS rules are quite reasonable for languages with a shallow orthography such as Marathi and Spanish. On the other hand, LTS rules produce pitiable results for languages like English and French. For this reason, most modern text-to-speech systems for such languages apply data driven and statistical techniques [7]. The general case is that there is no one-to-one correspondence between letters and phonemes. Multiple phonemes can be aligned by a single character two letters can align to a single phoneme, or a letter may align to no phonemes at all (for example, the e in make). Before applying the data driven G2P algorithm, characters to phonemes alignment has to be made.

2.2.2 Prosodic Analysis

The final step of text analysis is prosodic analysis. Prosody refers to the characteristics that make sentences flow naturally [14]. Without these characteristics, speech would sound like a reading of a list of words. The three main components of prosody are phrasing, stress, and pitch. Phrasing has many effects on speech synthesis; the final vowel of a phrase is longer than the previous vowels and there is often a drop in the fundamental frequency from the start of a phrase to its end. Phrasing prediction can be based on deterministic rules. Modern techniques for phrasing prediction are data-driven techniques. Decision trees can be used in phrase break prediction [9]. Some machine learning algorithms have been applied for phrasing prediction such as neural networks [10]. Stress is used to indicate the strength of a word, syllable or phrase when it is used in a sentence. Variation in stress is used to distinguish between nouns and verbs. For example, if you say two words "पहिल्याandदुसऱ्या"[6], you should find that the stress in nouns is on the first syllable whereas in verbs it is on the last syllable. The stressed syllables may be characterized by increased loudness, longer duration or increased fundamental frequency. Improper location of stress may result in a change in the word meaning. Pitch is the pattern of fundamental frequency variation over a sentence. Variation in pitch may have an effect on the whole sentence. A noticeable example of intonation is the distinction between sentences and yesno questions in Marathi. It is possible to change the meaning of a sentence from a statement to a question by raising the pitch of the last syllable [6][18][19].

III. Marathi Language Specific Difficulties

Marathi is an Indo-Aryan language spoken by about 71 million people mainly in the Indian state of Maharashtra and neighbouring states. Marathi is also spoken in Israel and Mauritius. Marathi is thought to be a descendent of Maharashtri, one of the Prakrit languages which developed from Sanskrit. Marathi first appeared in writing during the 11th century in the form of inscriptions on stones and copper plates. From the 13th century until the mid-20th century, it was written with the Modi alphabet. Since 1950 it has been written with the Devanāgarī alphabet [2][18][19].

Table: 1 Vowels [2]

अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	अं	अः	अँ	आँ
a	ā	i	ī	u	ū	ṛ	e	ai	o	au	aṅ	aḥ		
[ə]	[a]	[i]	[i]	[u]	[u]	[ru]	[e]	[əi]	[o]	[əu]	[əⁿ]	[əh]	[æ]	[ɔ]
प	पा	पि	पी	पु	पू	पृ	पे	पै	पो	पौ	पं	पः		
pa	pā	pi	pī	pu	pū	pr	pe	pai	po	pau	paṅ	pah		

Table: 2 Consonants [2]

क	ka	[kə]	ख	kha	[$k^hə$]	ग	ga	[gə]	घ	gha	[$g^ɦə$]	ङ	ṅa	[ŋə]
च	ca	[tsə/tʃə]	छ	cha	[$ts^hə$]	ज	ja	[ʄə/zə]	झ	jha	[$ʄ^ɦə$/$z^ɦə$]	ञ	ña	[ɲə]
ट	ṭa	[ʈə]	ठ	ṭha	[$ʈ^hə$]	ड	ḍa	[ɖə]	ढ	ḍha	[$ɖ^ɦə$]	ण	ṇa	[ɳə]
त	ta	[t̪ə]	थ	tha	[$t̪^hə$]	द	da	[d̪ə]	ध	dha	[$d̪^ɦə$]	न	na	[nə]
प	pa	[pə]	फ	pha	[$p^hə$/fə]	ब	ba	[bə]	भ	bha	[$b^ɦə$]	म	ma	[mə]
य	ya	[jə]	र	ra	[rə]	ऱ	ṛa	[ɽə]	ल	la	[lə]	व	va	[və/wə]
श	śa	[ʃə]	ष	ṣa	[ʂə]	स	sa	[sə]						
ह	ha	[ɦə]	ळ	ḷa	[ɭə]	क्ष	kṣa	[kʃə]	ज्ञ	jña	[ʤɲə]	श्र	śra	[ʃrə]

3.1 Vowelization

Vowelization is the process of adding vowels to an unmarked text [2]. The Marathi text written in newspapers, scientific or literature books does not contain vowels and other markings needed to pronounce the text correctly. Vowels are added to the text only in the cases where ambiguity appears and cannot be resolved from the context; otherwise writers assume that the reader has enough knowledge of the language that enables him to infer the correct vowels [2].The Festival-based synthesizer dealt with the discretization problem by the use of a Romanized version of Marathi text [2]. Since Marathi is written in a different alphabet than any language written with Latin alphabet, it is difficult for people with no familiarity of the Marathi alphabet to comprehend Marathi texts. So that it is helpful to transliterate this alphabet into Latin alphabet [2]. On the other hand, the MARY-based synthesizer requires that the input text be fully diacritized. Some synthesizers implement a module for automatic vowelization. But, since the accuracy of automatic vowelization is not high and speech synthesis requires high accuracy, the MARY-based synthesizer followed a user-dependent approach.

3.2 Dialects

Marathi is an Indo-Aryan language spoken predominantly by Marathi people of Maharashtra. It is the official language and co-official language in Maharashtra and Goa states of Western India respectively, and is one of the 23 official languages of India. There were 73 million speakers in 2001; Marathi ranks 19th in the list of most spoken languages in the world. Marathi has the fourth largest number of native speakers in India.[11] Marathi has some of the oldest literature of all modern Indo-Aryan languages, dating from about 900 AD. The major dialects of Marathi are Standard Marathi and the Varhadi dialect.[12] There are other related languages such as Khandeshi, Dangi, Vadvali and Samavedi. Malvani Konkani has been heavily influenced by Marathi varieties. This diversity of the Marathi dialects is considered a problem for speech synthesis form many reasons. First, what dialect is to be generated? A choice must be done between generating Modern Standard Marathi and one of the other dialects. Second, MSA is understood by people with a high level of education so that its listener base is limited. Both the Festival-based synthesizer and the MARY-based synthesizer used MSA.

3.3 Differences in Gender

The Marathi speech is influenced by the gender of the person the speech is directed to or is about. For example, the feminine form of a verb augments the masculine form by the feminine " पहिल्या" or by the feminine "दुसऱ्या" according to the verb tense [6]. As a consequence to that when the speech is the final product of a system such as a translation system or a synthesis system, inappropriate gender marking is more obvious and unsatisfactory than it is when the system generates only text.

IV. An Marathi Di-Phone Speech Synthesizer In Festival

The Festival Speech Synthesis System was developed at the center for Speech Technology Research at the University of Edinburgh in 1990 [20]. The system has been developed by Paul Taylor, Alan Black and Richard Caley. Festival is implemented in C++ and scheme. Despite the fact that it would be gorgeous to implement the system in a single language, practical reasons made the use of more than one language a necessity. Festival is not used only as a research platform but also it is used as a run-time system so that speed is obligatory. Because of this, it is essential to have significant amounts of the code written in a compiled low-level language such C++. For particular types of operations, such as the array processing often used in signal processing, a language such as C++ is much faster than higher-level alternatives. However, it is too preventive to use a system that is completely compiled, as this prevents essential run-time configuration [13][18]. In order to give parameters and specify flow of control, Festival also offers a scripting language based on the Scheme programming language. Scheme has a very simple syntax but is at the same time powerful for specifying parameters and simple functions. Scheme is selected because it is restricted and is considered as a small

language and would not increase the size of the Festival system [14]. In the Festival-based synthesizer, the first task required was to construct the natural language modules for the Marathi language. This step requires two types of analysis. The first type is the language specific analysis such as phones, lexicon, tokenization and others. The second type is the speaker specific analysis, where prosodic analysis such as duration and intonation are the main issues [6][18][19]. The phone-set definition is the first text analysis module in which all phonemes have to be defined and a specification of articulatory features for each phoneme has to be included. The second text analysis module is the lexicon module which is responsible for finding the pronunciation of a word. This is done either by a lexicon, i.e. a large list of words and their pronunciations or by some letter to sound rules. Since the Marathi language has a simple and well-defined syllabic structure and can be unambiguously derived from a phone string, the use of letter to sound rules is a suitable choice. The second task required was to construct a di-phone database for that new language. In their system for the Marathi language, they used fabricated words where only one occurrence of each di-phone is recorded. The use of fabricated words does not demand searching for natural examples that have the desired di-phone. In the Festival-based synthesizer, the synthesis modules have used linear predictive coding (LPC) which is used to represent the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model. It is one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters.

V. An Marathi Di-Phone Speech Synthesizer Using Mary

MARY stands for Modular Architecture for Research on speech synthesis and it is a tool used for research, development and teaching in the field of text-to-speech [15][18][19]. Two tasks are required to add support for a new language to MARY TTS; the first task is to construct a minimal set of natural language processing (NLP) components for the new language. In this task some kind of script is applied on a voluminous body of encoded text in the target language, such as a XML dump of the Wikipedia in the target language to extract the actual text without markup and the most frequent words, and then a pronunciation lexicon has to be built up. Using MARY transcription tool, which supports a semi-automatic procedure for transcribing new language text and automatic training of letter-to-sound rules for that language, many of the most frequent words has to be manually transcribed then a ' trainpredict ' button in the GUI is used to automatically train a simple letter-to-sound algorithm and predict pronunciations for the untranscribed words in the list. To be able to use the MARY transcription tool, a XML file describing the allophones that can be used for transcription and providing for each allophone the phonetic features that are to be used for characterizing the phone later is needed. The second task is the creation of a voice in the target language and the ' redstart' voice recording tool can be used for this purpose [16]. The synthesizer module in MARY uses multi-band resynthesize overlap add (MBROLA) for di-phone synthesis. MBROLA is both a di-phone synthesis technique and an actual system that constructs waveforms from segment, duration and F0 target information. MBROLA is a time-domain algorithm, as pitch synchronous overlap add (PSOLA), which implies very low computational load at synthesis time. Unlike PSOLA, however, MBROLA does not require a preliminary marking of pitch periods. MBROLA's synthesis produces high quality speech without requiring too much effort to design the di-phone database. This quality is due to that it is based on a preprocessing of di-phones (imposing constant pitch and harmonic phases), which enhances their concatenation while only slightly degrading their segmental quality [17][18][19].

VI. Experiments And Results

The two most important criteria that are measured when evaluating a synthesized speech are the intelligibility and the naturalness of the speech. The intelligibility of the speech means whether or not the synthesizer's output could be understood by a human listener while the naturalness of the speech means whether or no the synthesized speech sounds like the human speech [6]. The feeling of naturalness about speech is based on a complex set of features. In the Diagnostic Rhyme Test (DRT) which measures the intelligibility test, a list of word pairs that differ only in a single consonant are uttered and the listeners are asked to mark on an answer sheet which word of each pair of the words they think is correct. Some examples of word pairs used in the DRT are shown in Table 1. The result of applying this test on the Festival based synthesizer was approximately 89%, while the results of applying that test on the MARY-based synthesizer was 92% which means 92% of the test words were correctly understood. The following table lists some examples of the word pairs used in the DRT applied on the MARY-based synthesizer. In another type of test that measures the naturalness, the listeners were asked to rank the voice quality using a five level scale. Concerning the Festival-based synthesizer, 55% of the listeners considered the synthesized speech natural and 32% considered it

Table 1: Some examples of word pairs used in the DRT [6].

mar_001	कारणआपल्याकडेतीपद्धतनाही
mar_002	त्यानेपहिल्यावदुसऱ्यामहायुद्धात
mar_003	तुम्हालात्यातूनबराचअंदाजयेईल
mar_004	सिंहरासंसिंहनावाचीखगोलीयरास
mar_005	अशाप्रकारेअमेरिकेचायायुद्धातप्रवेशझाला
mar_006	रत्नागिरीजिल्ह्याच्यामाहितीसाठीयेथेटिचकीद्या
mar_007	पुणेशहरातीलएकमध्यवर्तीठिकाण
mar_008	जलंधरशहरातीलक्रिकेटचेमैदानआहे
mar_009	आपणचर्चाकरतानाकृपयाचारवापरूनसहीकरावी
mar_010	केवळप्रबंधकहीपानेबदलूशकतातकाय

Unnatural. On the other hand, more than 67% of the listeners considered the speech of the MARY-based synthesizer natural.

VII. Conclusion

The results of the DRT which measures the intelligibility for both the Festival-based synthesizer and the MARY-based synthesizer have been presented. They can be considered acceptable results. On the other hand, the results of test that measures the naturalness of the synthesized speech for both the synthesizers needs to be enhanced. The naturalness measure results are not so good because the di-phone database contains only one instance of each speech unit. So, when concatenating these units to each other, the prosodical characteristics of these units might be different and this degrades the quality of the resulting speech.

References

[1]. Sangramsing Kayte, Dr. Bharti Gawali "Marathi Speech Synthesis: A review" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 3 Issue: 6 3708 – 3711

[2]. Sangramsing Kayte, Monica Mundada "Study of Marathi Phones for Synthesis of Marathi Speech from Text" International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-10) October 2015

[3]. Richard Sproat and Steven Bedrick (September 2011). "CS506/606: Txt Nrmlztn". Retrieved October 2, 2012.

[4]. John Holmes and Wendy Holmes, "Speech Synthesis and Recognition", Second edition published by Taylor & Francis, 2001.

[5]. Sangramsing N.kayte "Marathi Isolated-Word Automatic Speech Recognition System based on Vector Quantization (VQ) approach" 101th Indian Science Congress Jammu University 03th Feb to 07 Feb 2014.

[6]. Sangramsing Kayte, Monica Mundada, Santosh Gaikwad, Bharti Gawali "PERFORMANCE EVALUATION OF SPEECH SYNTHESIS TECHNIQUES FOR ENGLISH LANGUAGE " International Congress on Information and Communication Technology 9-10 October, 2015

[7]. Marchand Y. and Damper R., "A multistrategy approach to improving pronunciation by analogy",Computational Linguistics, vol. 26, no. 2, pp.195–219, 2000.

[8]. Monica Mundada, Bharti Gawali, Sangramsing Kayte "Recognition and classification of speech and its related fluency disorders" International Journal of Computer Science and Information Technologies (IJCSIT)

[9]. Wang M. Q. and Hirschberg J., "Automatic classification of intonational phrase boundaries", Computer Speech and Language, vol. 6, pp. 175–196, 1992.

[10]. Fackrell J. W. A. et al., "Multilingual prosody modeling using cascades of regression trees and neural networks", in Proceedings of Eurospeech Budapest, Hungary, pp. 1835-1838, 1999.

[11]. Abstract of Language Strength in India: 2001 Census". Censusindia.gov.in. Archived from the original on 10 February 2013. Retrieved 2013-05-09.

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

[13]. M. Schroder and J. Trouvain, "The German textto-speech synthesis system MARY: A tool for research, development and teaching", International Journal of Speech Technology, vol.6, pp.365–377, 2003.

[14]. M. Schroder et al., "Multilingual MARY TTS participation in the Blizzard Challenge 2009", In Blizzard Challenge, Edinburgh, UK, 2009.

[15]. Etxebarria B. et al., "Improving Quality in a Speech Synthesizer Based on the MBROLA Algorithm", Proceedings of Eurospeech, Budapest, pp. 2299-2302, 1999.

[16]. Monica Mundada, Sangramsing Kayte "Classification of speech and its related fluency disorders Using KNN" ISSN2231-0096 Volume-4 Number-3 Sept 2014

[17]. Monica Mundada, Sangramsing Kayte, Dr. Bharti Gawali "Classification of Fluent and Dysfluent Speech Using KNN Classifier" International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 9, September 2014

[18]. Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "Di-phone-Based Concatenative Speech Synthesis System for Hindi" International Journal of Advanced Research in Computer Science and Software Engineering -Volume 5, Issue 10, October-2015

[19]. Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "A Review of Unit Selection Speech Synthesis International Journal of Advanced Research in Computer Science and Software Engineering -Volume 5, Issue 10, October-2015