Computer Engineering and Intelligent Systems www.iiste.org
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)

Marathi Speech Synthesized Using Unit selection Algorithm

Kalyan D.Bamane (Corresponding author)

D.Y.Patil College of Enginerring, Akurdi,Pune-44

University of Pune

Tel: 020-27653054 E-mail: kalyandbamane@gmail.com

Kishor N.Honwadkar

D.Y.Patil College of Enginerring, Akurdi,Pune-44

University of Pune

Tel: 020-27653054 E-mail: knhonwadkar@yahoo.co.in

Abstract

In this paper, we present the concatenative text-to-speech system and discuss the issues relevant to the development of a Marathi speech synthesizer using different choice of units: Words, Di phone and Tri phone as a database. Quality of the synthesizer with different unit size indicates that the word synthesizer performs better than the phoneme synthesizer. The most important qualities of a speech synthesis system are naturalness and intelligibility. We synthesize the Marathi text and perform the subjective evaluations of the synthesized speech. As a result 85% of speech synthesized by the proposed method was preferred to that by the conventional method; the results show the effectiveness of the proposed method. In this paper we are going to focus on a Dip hone and Trip hone through which will get a 95% quality voice.

Keywords: Speech synthesis, concatenation, unit selection, corpus, DSP.

1. Introduction

Text-to-speech systems have an enormous range of applications. Their first real use was in reading systems for the blind, where a system would read some text from a book and convert it into speech. These early systems of course sounded very mechanical, but their adoption by blind people washardly surprising as the other options of reading Braille or having a real person do the reading were often not possible. Today, quite sophisticated systems exist that facilitate human computer interaction for the blind, in which the TTS can help the user navigate around a windows system .The mainstream adoption of TTS has been severely limited by its quality. Apart from users who have little choice as in the case with blind people, people's reaction to old style TTS is not particularly positive. While people may be somewhat impressed and quite happy to listen to a few sentences, in general the novelty of this soon wears off. In recent years, the considerable advances in quality have changed the situation such that TTS systems are more common in a number of applications. Probably the main use of TTS today is in call-centre automation, where a user calls to pay an electricity bill or book some travel and conducts the entire transaction through an automatic dialogue system Beyond this, TTS systems have been used for reading news stories, weather reports, travel directions and a wide variety of other applications.

ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)

1.1.1 THE Goals of TTS.

One can legitimately ask, regardless of what application we want a talking computer for, is it really necessary that the quality needs to be high and that the voice needs to sound like a human?

Wouldn't a mechanical sounding voice suffice? Experience has shown that people are in fact very sensitive, not just to the words that are spoken, but to the *way* they are spoken. After only a short while, most people find highly mechanical voices irritating and discomforting to listen to. Furthermore tests have shown that user satisfaction increases dramatically the more natural sounding the voice is. Experience and particularly commercial experience shows that users clearly want natural sounding that is human-like systems.

Hence our goals in building a computer system capable of speaking are to first build a system that clearly gets across the message, and secondly does this using a human-like voice. Within the research community, these goals are referred to as **intelligibility** and **naturalness** .A further goal is that the system should be able to take any written input; that is, if we build an English text-to-speech system, it should be capable of reading any English sentence given to it. With this in mind, it is worth making a few distinctions about computer speech in general. It is of course possible to simply record some speech, store it on a computer and play it back. We do this all the time; our answer machine replays a message we have recorded, the radio plays interviews that were previously recorded and so on. This is of course simply a process of playing back what was originally recorded. The idea behind text-to-speech is to "play back" messages that weren't originally recorded. One step away from simple playback is to record a number of common words or phrases and recombine them, and this technique is frequently used in telephone dialogue services. Sometimes the result is acceptable, sometimes not, as often the artificially joined speech sounded stilted and jumpy. This allows a certain degree of flexibility, but falls short of open ended flexibility. Text-to-speech on the other hand, has the goal of being able to speak anything, regardless of whether the desired message was originally spoken or not. there are a number of techniques for actually generating the speech. These generally fall into two camps, which we can call bottom-up and concatenative.

In the bottom-up approach, we generate a speech signal "from scratch", using our knowledge of how the speech production system works. We artificially create a basic signal and then modify it, much the same way that the larynx produces a basic signal which is then modified by the mouth in real human speech. In the concatenative approach, there is no bottom-up signal creation perse; rather we record some real speech, cut this up into small pieces, and then recombine these to form "new" speech. Sometimes one hears the comment that concatenative techniques aren't real speech synthesis in that we aren't generating the signal from scratch. This point may or may not be relevant, but it turns out that at present concatenative techniques far out perform other techniques, and for this reason concatenative techniques currently dominate.

1.1.2 The Engineering Approach

We take what is known as an engineering approach to the text-to-speech problem. The term engineering is often used to mean that systems are simply bolted together, with no underlying theory or methodology. Engineering is of course much more than this, and it should be clear that great feats of engineering, such as the Brooklyn bridge were not simply the result of some engineers waking up one morning and banging some bolts together. So by "engineering", we mean that we are tackling this problem in the best traditions of other engineering; these include, working with the materials available and building a practical system that doesn't for instance take days to produce a single sentence. Furthermore, we don't use the term engineering to mean that this field is only relevant or accessible to those with traditional engineering backgrounds or education. As we explain below, TTS is a field relevant to people from many different backgrounds.

One point of note is that we can contrast the engineering approach with the scientific approach. Our task is to build the best possible text-to-speech system, and in doing so, we will use any model, mathematics, data, theory or tool that serves our purpose. Our main job is to build an *artefact* and we will use any means possible to do so. All artifact creation can be called engineering, but *good* engineering involves more: often we wish to make good use of our resources we don't want to use a hammer to crack a nut we also in general want to base our system on solid principles. This is for several reasons. First, using solid say mathematical principles assures use are on well tested ground; we can trust these principles and don't have to experimentally verify every step we do. Second, we are of course not building the last ever text-to-speech system; our system is one step in a continual development; by basing our system on solid principles we hope to help others to improve and build on our work. Finally, using solid principles has the advantage of helping us diagnose the system, for instance to help us find why some components do perhaps better than expected, and allow the principles of which these components are based to be used for other problems.

Speech synthesis has also been approached from a more scientific aspect. Researchers who pursue this approach are not interested in building systems for their own sake, but rather as models which will shine light on human speech and language abilities. As such, the goals are different, and for example, it is important in this approach to use techniques which are at least plausible possibilities for how humans would handle this task. A good example of the difference is in the concatenative waveform techniques which we will use predominantly; recording large numbers of audio waveforms, chopping them up and gluing them back together can produce very high quality speech. It is of course absurd to think that this is how humans do it. We bring this point up because speech synthesis is often used or was certainly used in the past as a testing ground for many theories of speech and language.

ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)

I. SPEECH SYNTHESIS SYSTEM

A Text-to-Speech (TTS) Synthesizer is a computer based system that should be able to read any text aloud, whether it was introduced in the computer by an operator or scanned and submitted to an Optical Character Recognition (OCR) system. The objective of a text to speech system is to convert an arbitrary given text into a spoken waveform. Main components of text to speech system are: Text processing and Speech generation.

i).SCRIPTS OF INDIAN LANGUAGES

The basic units of the writing system in Indian languages are Aksharas, which are an orthographic representation of speech sounds. An Akshara in Indian language scripts is close to a syllable and can be typically of the following

Form: C, V, CV, CCV, VC and CVC where C is a Consonant and V is a vowel

ii) FORMAT OF INPUT TEXT

The scripts of Indian language are stored in digital Computers in ISCII, UNICODE and in transliteration scheme of various fonts. The input text could be available in any of these formats could be conveniently separated from the synthesis engine. An Indian language have a common phonetic base, the engines could be built for one transliteration scheme that can represent the script of all Indian language.

iii) MAPPING OF NON-STANDARD WORDS

TO STANDARD WORDS

In practice, an input text such as news article consists of standard words and non-standard words such as initials, digits, symbols and abbreviations. Mapping of nonstandard words to a set of standard words depends on the context, and it is a non-trivial problem.

iv) STANDARD WORDS TO PHONEME SEQUENCE

Generation of sequence of phoneme units for a given standard word is referred to as letter to sound rules. The complexity of these rules and their derivation depends on the nature of the language.

V) SPEECH GENERATION COMPONENT

Given the sequence of phonemes, the objective of the speech generation component is to synthesize the

acoustic wave form. Speech generation has been attempted by concatenating the recorded speech segments art speech synthesis generates natural sounding speech by using large number of speech units. Storage of large number of units and their retrieval in real time is feasible due to availability of cheap memory and computation power. The approach of using an inventory of speech units is referred to as unit selection approach. It can also be referred to as data-driven approach or example based approach for speech synthesis. The issues related to the unit selection speech synthesis system are: 1) Choice of unit size, 2) Generation of speech database, 3) Criteria for selection of a unit. coverage of all possible words, phrases, proper nouns, and other foreign words may not be ensured. (Note 1)

There are two issues concerning the generation of unit selection databases. They are: 1) Selection of utterances which has the coverage of all possible units, 2) Recording of these utterances by a good voice talent.

II CONCATENATIVE SYNTHESIS

In this approach synthesis is done by using natural speech. This methodology has the advantage in its simplicity, i.e. Concatenative synthesis is based on the concatenation or stringing together of segments of recorded speech. Generally, concatenative synthesis produces the most natural sounding synthesized speech. There are three main subtypes of concatenative synthesis: Unit selection synthesis, Diaphone synthesis, Domain-specific synthesis.

A. UNIT SELECTION SYNTHESIS

Unit selection synthesis uses large databases recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, syllables, morphemes, words, phrases, and sentences. Typically, the division into segments is done using a specially modified speech recognizer set to a forced alignment mode with some manual correction Afterward, using visual representations such as the Waveform and spectrogram. An index of the units in the speech database is then created based on the segmentation and acoustic parameters like the fundamental frequency pitch duration, position in the syllable, and neighboring Phones. Unit selection provides the greatest naturalness, Because it applies only small amounts of digital signal

1.1 Research Methods:

In this we are focusing on dip hone and trip hone from the word and following methods are used for the same:

1.1.1Research Plan for Marathi TTS

Speech Database

1. Marathi Text corpus specially designed for TTS considering following facts.

I. It should contain all phonetic element of Marathi language i.e. Corpus containing phonetically balanced sentences.(e.g c,cv,vc,v,ccv etc)

II. It should contains all features required for Unit selection approach (i.e. Linguistic & acoustic

features)

2. Recording of Speech corpus.

I. Recording of designed corpus can be done by professional new reader or anchor having better pronunciation as per language rules.
II. Cleaning of recorded data is performed at studio by sound expert (i.e. removing unwanted noise).
III. Segmentation of recorded data is performed according to pages and files.
IV. Re-sampling of Segmented wave files according standards of TTS (Normally PCM 22Khz 16bit Mono)

1.2 Annotation of the Speech data

Tagging of speech data into following units.

A. words B. Syllables (bha:rət=bha:+rət)

C . Tri-phone(syə,stri mainly containing /y/) D. Di-phone(kə,sa:,ri etc) E.Phones(ə,I, k etc)

Database is generated for unit selection approach and Indexing of can be done for performance improvement of the system

Tri-phone and di-phone convertor

1.3 TTS Engine

1. Natural Language Processing Module

Preprocessing Module (Abbreviation, Acronyms & some special character ())

I. Devenagri text parser Module(Date &Number identification etc)
II. Grapheme to Phoneme Convertor(IPA)
- Schwa handling-We cannot provide inherent schwa with every alphabet or devnagri characters)
- Nasalization's- Handles the problem of (Bindi in Hindi) answer.
- Morphophonemic analyzer-resolve the issue of compound words like rajyasabha, loksabha)
- Syllabification- break the word in to syllables like (bha:rət=bha:+rət)
-

2. Unit Selection Module-

Algorithm to find out best unit for concatenation by considering linguistic and acoustic features of units (our approach decision tree)

3. DSP Module

I. Concatenation of selected sound units.
II. Speed Modifier module
III. Multimedia module (play pause and stop functionality)

1.4 Analysis Result

ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)

Proposed System:

The concatenative text-to-speech system and discuss the issues relevant to the development of a Marathi speech synthesizer using different choice of units:

Words, dip hone and trip hone as a database. Here we are using IPA method through which we could come to know that which ia s previous Unit which one is a current unit and next unit.

Quality of the synthesizer with different unit size indicates that the word synthesizer performs better than the phoneme synthesizer. The most important qualities of a speech synthesis system are naturalness and intelligibility. We synthesize the Marathi text and perform the subjective evaluations of the synthesized speech. As a result,(1) 85% of speech synthesized by the proposed method was preferred to that by the conventional method, The results show the effectiveness of the proposed method we are going to focus on a Dip hone and Tri phone through will get a 95% quality voice.

Grapheme to phoneme convertor for Marathi.

Input: Marathi text (Unicode)

Output : IPA (Unicode)

The following Rules are used for the analysis :

1.G2P rule-one to one mapping e.g 1ka->kə

2.Nasalization rule: It is used for silent na which is used in Marathi Language.

3.Schwa handling rule: It is used to resolve the jurk from the word.

4.Di-phone/tri-phone rule: It is used to handle dip hone and tri phone from the word.

To find a best match target unit following mathematical formula is used

1. Unit cost

When source unit Su is given, system should match the target unit Tu through the following

Formula:

$$P = \begin{cases} 1 & \text{for all} \\ 0.8 & \text{Different unit end with same c or v} \\ 0.5 & \text{In same class} \end{cases}$$

Linguistic Unit cost= $\sum$ get penalty(P)

$$\begin{cases} 1 & \text{for pitch diff} < \pm 100 \end{cases}$$

P= 0.5 >=±100 <=±500

0

Acoustic unit cost = ∑ get penalty(P)

Target Unit = Max(linguistic Unit Cost) Set + Max (Acoustic unit cost) Set

Above formula is used for the best match to the target unit. When match is found then Clear voice of the text is to be obtained.

Two types of features have been used:

1. Linguistic feature:

This feature used for dip hone. In this we use previous unit –current unit-next unit approach has been used.

In this feature defined class for previous and next unit is developed. But our research is for voiced stop and for one class. (Note1)

Above is a example of a linguistic feature in one class and not been used before.

Here International phonetic alphabet for Nasalization and dental features are used.(Note 2)

2 Acoustic Feature:
3

In this feature pitch and duration is been involved. Our approach to decide value of them.

When any dip hone is set it is directly saved in database.Follwing is a Acoustic feature of the(Note3).

References:

[1] Dutoit T., “An Introduction to Text-To-Speech Synthesis”, Kluwer Academic Publishers, 1996.

[2]Allen J., Hunnicut S., Klatt D., “From Text To Speech, The MITTALK System”, Cambridge University Press, 1987.

[3]Hunnicut S., "Grapheme-to-Phoneme rules: a Review", Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden, QPSR 2-3, pp. 38-60.

[4]Belrhari R., Auberge V., Boe L.J., "From lexicon to rules: towards a descriptive method of French text-to-phonetics transcription", Proc. ICSLP 92, Alberta, pp. 1183-1186.

[5] Kager R., “Optimality Theory”, Cambridge University Press, 1999 “Hindi Bangla English – Tribhasa Abhidhaan”, Sandhya Publication, 1st Edition March 2001