# BUILDING A TEXT-TO-SPEECH SYSTEM FOR PUNJABI LANGUAGE

Rupinderdeep Kaur, Mr. R.K. Sharma and Mr. Parteek Kumar

Computer Science and Engineering department,
Thapar University, Patiala, Punjab

***ABSTRACT***

*A Text-To-Speech (TTS) system is a computer application that is capable of converting typed text into speech. This paper contains description of working of a TTS system along with architecture of the system and various available TTS systems for Indian Languages with comparison of these systems on the basis of the methods used by them for speech synthesis. TTS system generally involves two steps, text processing and speech generation. Synthetic speech may be used in several applications, like, telecommunications services, language education, aid to handicapped persons, fundamental and applied research etc. TTS involves many challenges during the process of conversion of text to speech. The most important qualities expected from speech synthesis system are naturalness and intelligibility. The general architecture of a TTS system and different waveform generation methods are discussed in this paper. A scheme for developing a TTS system for Punjabi Language is also included in this paper.*

***KEYWORDS***

*TTS, Phonetics, Prosody, Speech Synthesis*

## 1. INTRODUCTION

Spoken words (speech) play a great role in the lives of people. Speech represents the spoken form of a language and is also one of the important means of communication. Over the past few decades, a good amount of research is being done in the field of converting text into speech. These efforts have resulted in important advances with many systems being able to generate the sound close to a real natural sound. These advances in speech synthesis also pave the way for many new speech related applications. Some of these applications are audio books, announcement systems, speech systems for visually handicapped *etc.*

## 2. TEXT-TO-SPEECH SYSTEM

The function of a Text-To-Speech (TTS) system is to convert an arbitrary text to a spoken waveform. This broadly involves two steps, namely, text processing and speech generation. Text processing is used to convert the given text to a sequence of synthesis units, while speech generation includes the generation of an acoustic wave form corresponding to each of these units in the sequence [1] [2].

Natarajan Meghanathan et al. (Eds) : ACSIT, SIPM, FCST, CoNeCo, CMIT - 2017
pp. 71– 87, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70806

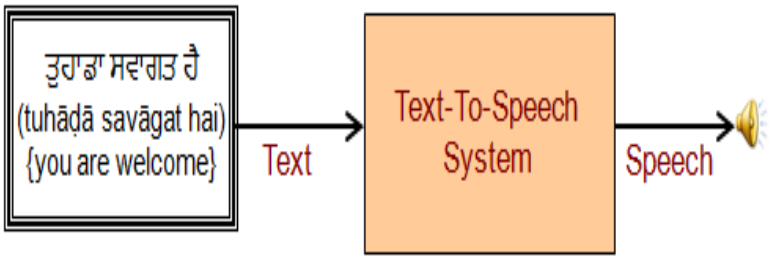


Figure 1. Text-To-Speech System

## 3. OBJECTIVES OF TEXT-TO-SPEECH SYSTEM

People are very sensitive, not just to the words that are spoken, but also to the way these are spoken. So the primary and the most important objective of any TTS system should be to produce a sound which looks like a natural sound. The sound should not be felt highly mechanical and should be comforting when listened. The important qualities of a speech synthesis system are naturalness and intelligibility. Naturalness means how close the output sounds is to human speech, while intelligibility is the ease with which the output is understood by a listener. One more important objective of a TTS system is that it should be able to take any written input, *i.e.,* any format or font and converts this to speech.

## 4. THE BASIC TERMINOLOGY

The basic terms used in speech synthesis include Phonology, Phonemics, Phonetics and Prosody. These are briefly explained in following sub sections.

### 4.1. Phonology

Phonology is the branch of linguistics that deals with systematic organization of sounds in a language. It has traditionally focused on the study of systems of phonemes in languages. It may also cover a linguistic analysis either at a level beneath the word or at all levels of language where sound is considered to be structured for conveying linguistic meaning [22]. This phoneme system and linguistic analysis is a key process for developing a TTS system for the language under consideration. Analysis of sounds for a given language helps in developing an efficient TTS system for that language.

### 4.2. Phonemics

A phoneme is a basic unit of a language's phonology, which is combined with other phonemes to form meaningful units such as words or morphemes. Phoneme can be described as "the smallest contrastive linguistic unit which may bring about a change of meaning"[22]. For example, the words ਬਾਪ (bāp) /baap/ {father}, ਪਾਪ (pāp) /paap/ {sin} and ਮਾਪ (māp) /maap/ {measure} can be obtained by replacing the phoneme /p/ for the phoneme /m/ and /b/. These words, which differ in meaning through a contrast of a single phoneme, are called minimal pairs. In order to build an efficient TTS system, a deep study of these phonemes is very necessary.

### 4.3. Phonetics

Phonetics is the branch of linguistics that comprises the study of sounds of human speech [22]. The International Phonetic Alphabet (IPA) is used as the basis for the phonetic transcription of speech. It is based on Latin alphabet and is able to transcribe most features of speech such as consonants, vowels, and suprasegmental features. Every documented phoneme available within the known languages in the world is assigned its own corresponding symbol. Figure 2 contains the IPA chart given by International Phonetics Association. According to the IPA symbols present in this chart, the Punjabi sentence ਪੁਰਾਣੀ ਕਹਾਵਤ ਹੈ ਕਿ ਹਾਰੀਏ ਨਾ ਹਿਮੰਤ ਵਿਸਾਰੀਏ ਨਾ ਰਾਮ (purāṇī kahāvat hai ki hārīē nā himant visārīē nā rām) can be transcribed as pʊɾáːɳikəháːvə_thæ̀_keː__háːɾienáːhɪmə_tə̆__bɪsaːɾɨynáɾàːm.

It is worth mentioning here that phonetics removes the language barrier and converts the set of phonemes to corresponding phonetic symbols. Phonetics plays major roles in TTS system because it not just concatenates the symbols to form a word, as done in phonology, but also represents the symbols for supra-segmental and tonal features of human sounds. For languages which are highly a tonal language, after phonology, conversion of phonemes to phonetics is very necessary for TTS system to be natural.

### 4.4. Prosody

In linguistics, Prosody is the rhythm, stress, and intonation of speech. Prosody may reflect various features of the speaker or the utterance: the emotional state of the speaker; the form of the utterance (statement, question, or command); the presence of emphasis, contrast, and focus. Prosodic features are supra-segmental. They are not confined to any one segment, but occur in some higher level of an utterance. Prosodic units are marked by phonetic cues, such as a coherent pitch contour – or the gradual decline in pitch and lengthening of vowels over the duration of the unit, until the pitch and speed are reset to begin the next unit. Breathing, both inhalation and exhalation, seems to occur only at these boundaries where the prosody resets [22]. Phonetics helps to incorporate these prosodic features into a TTS system to make it natural and intelligent.

## 5. ARCHITECTURE OF TEXT-TO-SPEECH SYSTEM

The conversion of Text-To-Speech is not a single step process. Many things need to be considered in the process of conversion of Text-To-Speech. The process of converting text into speech breaks down into a number of stages. The general architecture of converting Text-To-Speech is shown in Figure 3. O‟Malley [3] described the process of converting text into speech as a four stage process. These stages are:

- Text Pre-processing and Text Normalization
- Linguistic Analysis
- Prosodic Prediction
- Waveform Generation

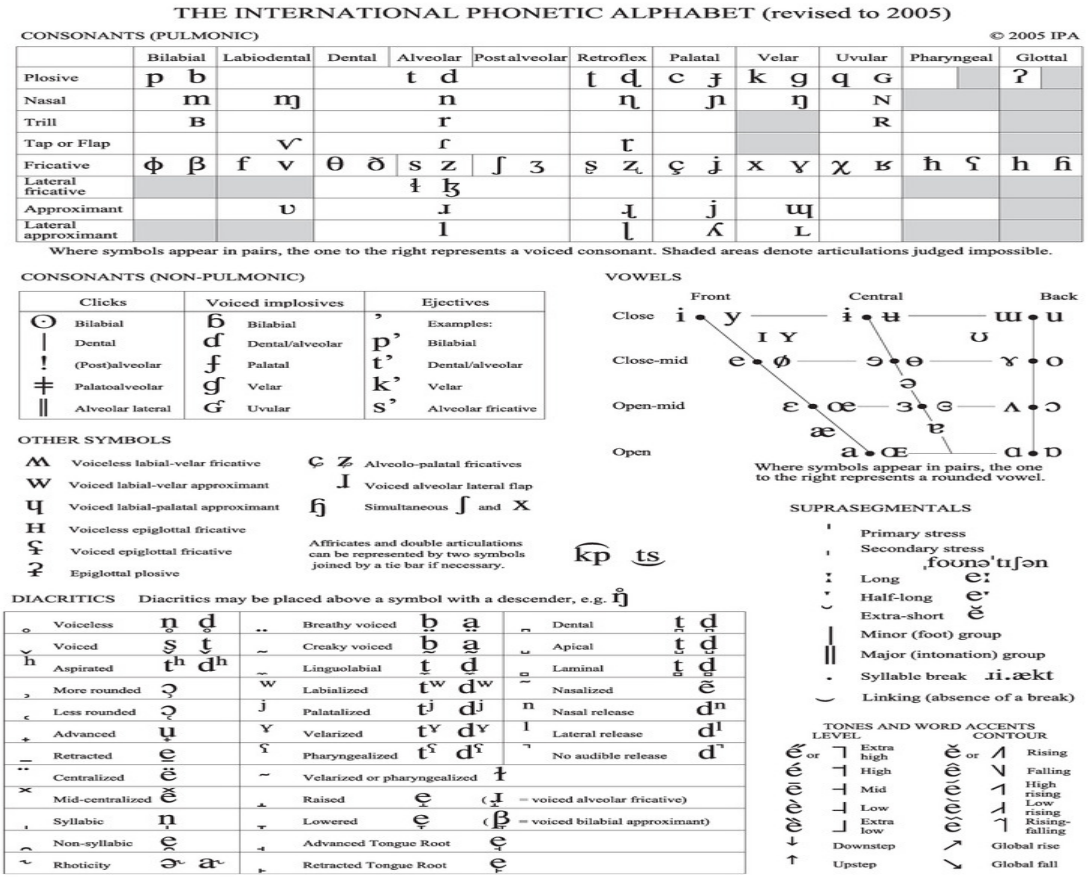


Figure 2. IPA Chart

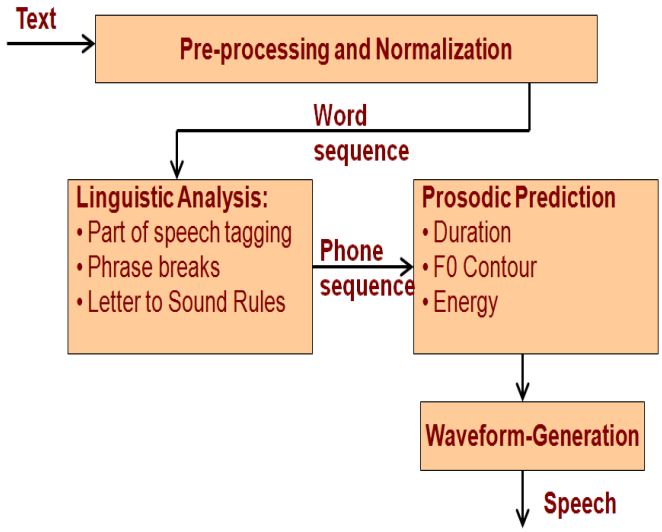


Figure 3. Architecture of Text-To-Speech System

## 5.1. Text Pre-Processing and Text Normalization

The typical input to a Text-To-Speech system is the text as available in electronic documents, news papers, blogs, emails *etc*. The text available in real world is a sequence of words available in standard dictionary. The text contains several non-standard words such as numbers, abbreviations, homographs and symbols built using punctuation characters such as exclamation '!', smileys ':-)' *etc*. The goal of text pre-processing module is to process the input text and generate appropriate phone sequences for each of the words.

The text in real world consists of words whose pronunciation is typically not found in dictionaries or lexicons such as "ਸੋਮ", "ਮੰਗਲ", and "ਡਾ." *etc.* Such words are referred to as Non-Standard Words (NSW). Various categories of NSW are as follows:

- Numbers, whose pronunciation changes depending on whether they refer to currency, time, telephone numbers, zip code *etc*.
- Abbreviations, contractions, acronyms such as "ਡਾ.", "ਲੈਫਟੀ.", "ਰੁ." *etc.*
- Punctuations 3-4, +/-, and/or, 4), dates, time, units and URLs.

The other issue of NSW is that, most of NSWs are homographs, *i.e*., words with same written form but different pronunciation. For example, IV which can vary as "four" in (*ਆਰਟੀਕਲ IV*), the "fourth" in (*ਹੇਨਰੀ IV*), and in the same as three or four digit numbers which could be dates and ordinary numbers (*26 ਜਨਵਰੀ 1950, 1950, 50 ਟਨ*).

Machine learning models such as Classification and Regression Trees (CART) have been used to predict the class of NSW which is typically followed by rules to generate appropriate expansion of a NSW into a standard form [4]. Rule-based systems for NSW operate on local grammars containing abstract contexts for within-sentence occurrences and sentence boundaries [28]. Mikheev's rule-based segmentation [29] is preceded by capitalized word disambiguation. There are also some language modelling and machine learning approaches for normalization subtasks. For example, in Sproat *et al.*, [4] word normalization is amongst others formulated in terms of maximizing the conditional probability of a normalized word sequence given an observed token sequence.

## 5.2. Linguistic Analysis

Linguistic analysis refers to scientific analysis of a language sample. It involves at least one of the five main branches of linguistics, which are, phonology, morphology, syntax, semantics, and pragmatics. Linguistic analysis can be used to describe unconscious rules and processes that speakers of a language use to create spoken or written language, and this can be useful to those who want to learn a language or translate from one language to another. For TTS system, linguistic analysis is done by part of speech tagging, prosodic phrase break marking and by defining letter to sound rules.

### 5.2.1. Part of Speech tagging

It is a process of assigning a part-of-speech to each word in a sentence. Part-of-Speech (POS) tagging is the task of determining the correct parts of speech for a sequence of words. Words are

often ambiguous with respect to part of speech. The word ਵਾਰ (waar)*, e.g.*, is a noun in the sentence "ਰਾਮ ਨੇ ਸ਼ਾਮ ਤੇ ਵਾਰ ਕੀਤਾ" (ram ne sham te waar kita), a finite verb in the sentence "ਰਾਮ ਨੇ ਸ਼ਾਮ ਨੂੰ ਵਾਰ ਵਾਰ ਕੇਹਾ ਕੇ ਓਹ ਚਲਾ ਜਾਵੇ" (ram ne sham nu waar waar keha ke ohh chala jave). This ambiguity is tackled by POS and results produced with predefined tags.

Rule-based approaches like ENGTWOL [30] operate on dictionaries containing word forms together with associated POS labels and morphologic and syntactical features and context sensitive rules to choose appropriate labels during application. In statistical approaches [31], the most probable tag sequence given the observed word sequence is generally estimated. In transformation-based tagging [32] a hybrid approach can be found, where disambiguation rules are derived by statistical means.

### 5.2.2. Phrase breaks

Phrase breaks incorporate the silence region in a speech. In text, it is marked with the help of POS. For Example consider the Punjabi sentence given in (1) and (2).

ਮੰਤਰੀ ਜੀ ਨੇ ਕਿਹਾ ਕਿ ਅੱਜ ਸਾਡੇ ਦੇਸ਼…. ...(1)

(mantrī jī nē kihā ki ajj sāḍē dēsh)

ਮੰਤਰੀ ਜੀ ਨੇ ਕਿਹਾ ਕਿ [pause] ਅੱਜ ਸਾਡੇ ਦੇਸ਼.... ...(2)

(mantrī jī nē kihā ki [pause] ajj sāḍē dēsh)

Here in (1), 'ਕਿ' is the preposition marked by POS and (2) tells about a phrase break, *i.e.,* while converting this text into speech, there will be a small pause in this region [27].

### 5.2.3. Letter to sound rules

Text to phoneme conversion is to derive the pronunciation for each word based on its spelling. Word pronunciations may be conveniently described as sequences of phonemes, which are units of sound in a language that serve to distinguish one word from another. There are mainly two approaches for text to phoneme conversion. These are dictionary based approach and rule based approach.

In Dictionary based approach, a large dictionary containing all the words of a language and their correct pronunciations is stored in the database. To determine the correct pronunciation of each word, it is looked up in the dictionary and the spelling is replaced with the pronunciation specified in the dictionary. This approach is quick and accurate, but it fails if looked up word is not present in the dictionary. Also, the size of dictionary should be considered carefully, as the memory space requirements of the synthesis system depends on the size[7].

In Rule based approach, pronunciation rules are applied to words, to determine their pronunciations based on their spellings. This approach is useful for the languages having very regular writing system, therefore, the prediction of the pronunciation of words based on their spellings is quite successful. In contrast, this approach works on any input, but the complexity of the rules grows substantially as the system takes into account irregular spellings or pronunciations [11].

## 5.3. Prosody Prediction

Prosodic prediction analysis deals with modeling and generation of appropriate duration and intonation contours for the given text. This is inherently difficult since prosody is absent in text. For example, the sentences – where are you going?; where are you GOING? and where are YOU going?, have same text-content but can be uttered with different intonation and duration to convey different meanings. To predict appropriate duration and intonation, the input text needs to be analyzed. This can be performed by a variety of algorithms including simple rules, example-based techniques and machine learning algorithms. The generated duration and intonation contour can be used to manipulate the context-insensitive diphones in diphone based synthesis or to select an appropriate unit in unit selection voices [21]. The prosody can be predicted by duration, F0 contour and energy contour.

### 5.3.1. Duration

Durations of the syllables are analyzed with respect to positional and contextual factors. For detailed duration analysis, syllables are categorized into groups based on size of the word and position of the word in the utterance, and the analysis is performed separately on each category. From the duration analysis, it is observed that durations of sound units depend on several factors at various levels, and it is very difficult to derive precise rules for accurate estimation of durations. Therefore, there is a need to explore nonlinear models to capture the duration patterns of sound units from features [23].

The factors affecting the duration of the basic sound units can be broadly categorized into phonological, positional and contextual. The vowel is considered as the nucleus of a syllable, and consonants may be present on either side of the vowel. The duration of a syllable is influenced by the position of the vowel, the category of the vowel and the type of the consonants associated with the vowel. The contextual factors include the preceding and the following syllables. In addition, the gender of the speaker, psychological state of the speaker, age *etc.* also affects the duration. For Example, consider the Punjabi sentence ਮੈਂ ਠੀਕ ਹਾਂ (maiṃ ṭhīk hāṃ)। The duration of the word “ਠੀਕ” (ṭhīk) can vary depending upon the mood of the speaker.

### 5.3.2. F0 Contour

F0 is usually defined, for voiced speech, as the rate of vibration of the vocal folds. F0 is considered one of the most important features for the characterisation of emotions and is the acoustic correlate of the perceptive pitch. Fundamental frequency, also known as pitch, is usually the lowest frequency component, or partial, which relates well to most of the other partials. In a periodic waveform, most partials are harmonically related, meaning that the frequency of most of the partials is related to the frequency of the lowest partial by a small whole-number ratio. The frequency of this lowest partial is the fundamental frequency of the waveform and how this fundamental frequency changes over the period of time is determined by the F0 contour [23].

Figure 4, shows the example of F0 contour of the word “ਸ਼ਲਗਮ ” (shalgam)

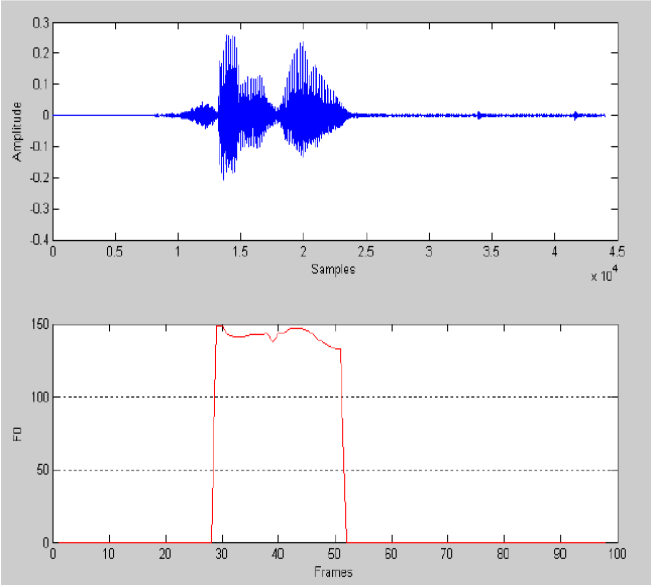


Figure 4. Signal and F0 of "शलगम " (shalgam)showing zero value of F0 in unvoiced region

### 5.3.3. Energy Contour

In linguistics, speech synthesis and music, the energy contour of a sound is a function or curve that tracks the perceived pitch of the sound over time. Energy contour may include multiple sounds utilizing many pitches, and can relate to frequency function at one point in time to the frequency function at a later point. It is fundamental to the linguistic concept of tone, where the pitch or change in pitch of a speech unit over time affects the semantic meaning of a sound. It also indicates intonation in pitch accent languages.

One of the primary challenges in speech synthesis technology, is to create a natural-sounding energy contour for the utterance as a whole. Unnatural energy contours result in synthesis that sounds lifeless or emotionless to human listeners, a feature that has become a stereotype of speech synthesis in popular culture.

## 6. METHODS OF SPEECH SYNTHESIS

Synthesized speech can be produced by several different methods. All of these have some benefits and deficiencies. The methods are usually classified into three groups [5]. These are as follows:

- Formant synthesis, which is done by exciting a set of resonators by a voicing source or noise generator to achieve the desired speech spectrum.
- Concatenative synthesis, which uses different length pre-recorded samples derived from natural speech.
- Statistical parametric synthesis, which is based on Hidden Markov Models (HMM).

Figure 5 shows the various methods for Wave-Form Generation. The formant and concatenative methods are the most commonly used in most synthesis systems. The Statistical parametric synthesis is still too complicated for high quality implementations, but may arise as a potential method in the future [6].

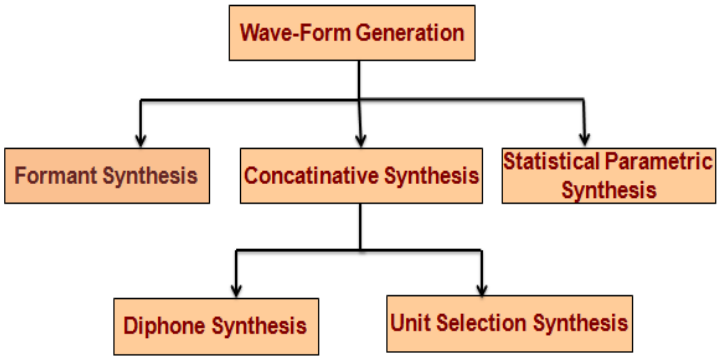


Figure 5. Wave-Form Generation Methods

## 6.1. Formant Synthesis

The most widely used synthesis method during last decades has been formant synthesis which is based on the source-filter-model of speech. It consists of artificial reconstruction of the formant characteristics to be produced. This is done by exciting a set of resonators by a voicing source or noise generator to achieve the desired speech spectrum. Parameters such as fundamental frequency, voicing and noise levels are varied over time to create a waveform of artificial speech [7]. This method is sometimes called rules-based synthesis. Figure 6 shows the typical architecture of formant synthesis.

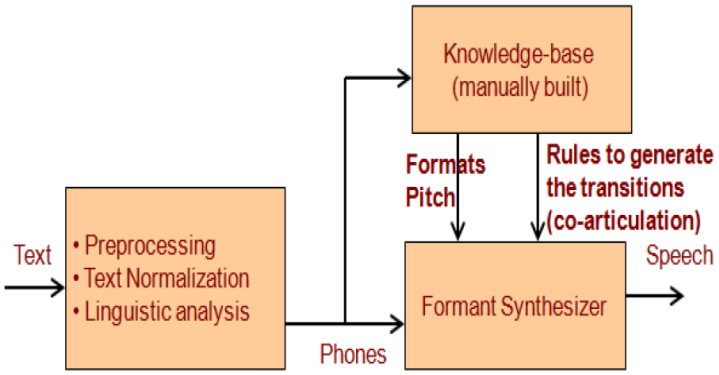


Figure 6. Architecture of Formant Synthesis

## 6.2. Concatenative Synthesis

Concatenative synthesis is based on the concatenation of segments of recorded speech. In this method, waveform segments are stored in a database. For a given text, these segments are joined based on some joining rules. Connecting pre-recorded natural utterances is probably the easiest

way to produce intelligible and natural sounding synthetic speech. One of the most important aspects in concatenative synthesis is to find correct unit length. The selection is usually a trade-off between longer and shorter units. With longer units, high naturalness, less concatenation points and good control of coarticulation are achieved, but the amount of required units and memory is increased. With shorter units, less memory is needed, but the sample collecting and labelling procedures become more difficult and complex [8].

For example, for Punjabi language, if longer units contains triphones, *i.e.,* CVC pairs, then memory will increase as there would be the combination of all the consonants firstly with all the vowels and then those combinations again with all the consonants, so memory requirement for saving those units will increase as compared to the shorter units wherein when single consonants and vowels are saved as individual units, it will take a very less memory. There are two main sub-types of concatenative synthesis [9][10], as given below:

- Unit selection synthesis
- Diphone synthesis

Figure 7 shows the architecture of concatenation based synthesis.

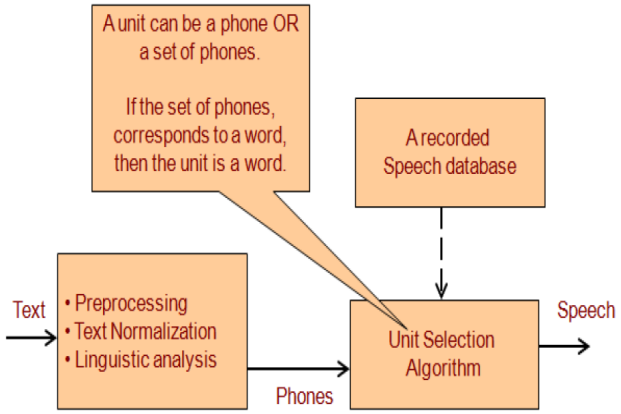


Figure 7. Architecture of concatenation based synthesis

In unit selection synthesis large databases of recorded speech are used [9]. The primary motivation for the use of large databases is that with a large number of units available with varied prosodic and spectral characteristics it should be possible to synthesize more natural-sounding speech than that can be produced with a small set of controlled units [10]. During database creation, each recorded utterance is segmented into individual phones, diphones, half-phones, syllables, morphemes, words, phrases, and sentences [11]. Figure 8, shows the architecture of unit selection process.

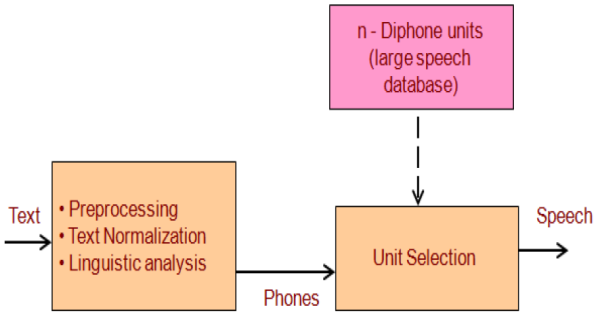


Figure 8. Architecture of unit selection process

Diphone synthesis uses a minimal speech database containing all the diphones occurring in a language. In diphone synthesis, only one example of each diphone is contained in the speech database. The quality of resulting speech is generally worse than that of unit-selection systems, but more natural-sounding than the output of formant synthesizers [7]. Figure 9 shows the snapshot of diphone.

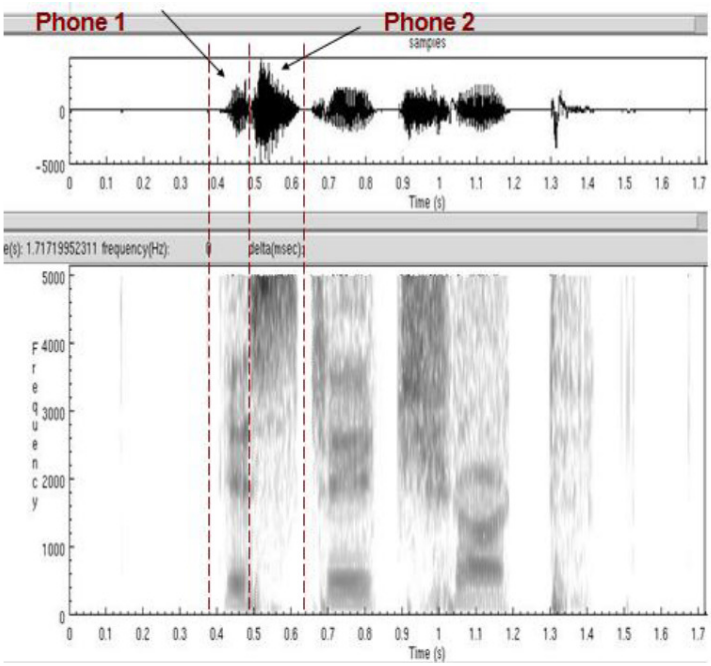


Figure 9. Snapshot of diphone

## 6.3. Statistical Parametric Synthesis

Statistical Parametric Synthesis (SPS) is one of the latest trends in TTS system. The SPS methods produce speech from a set of parameters learned from the speech data. Unlike traditional parametric synthesis methods which require manual specification and hand-tuning of the parameters, the SPS methods use statistical machine learning models such as Classification and Regression Tree (CART), HMMs, *etc*. to estimate the parameters of speech sounds and their

dynamics. The SPS methods offer simplicity in storage by encoding the speech data in terms of a compact set of parameters, and also provide mechanisms for manipulation of prosody, voice conversion *etc*. The SPS methods are found to produce intelligible and consistent speech as compared to natural and often inconsistent speech by unit selection techniques [12].

## 7. TEXT-TO-SPEECH SYSTEMS FOR INDIAN LANGUAGES

In order to help visually impaired and vocally disabled persons and to cater to the needs of day to day increasing applications of TTS system, development of innovative TTS systems has become a necessity in India. Different TTS systems have been built for Indian languages. Some of them are described below.

### 7.1. Dhvani- Indian Language Text-To-Speech System

Dhvani is a Text-To-Speech System specially designed for Indian languages. It has been developed by Simputer trust headed by Dr. Ramesh Hariharan at Indian Institute of Science Bangalore in year 2000. This system is available online at *http://dhvani.sourceforge.net/.* It uses diphone concatenation algorithm. Currently this system has Hindi, Malayalam, Kannada, Bengali, Oriya, Punjabi, Gujarati, Telugu and Marathi modules. All sound files stored in the database are compressed files. It has different modules for every language. It is based on observation that a direct grapheme to phoneme mapping exits for all Indian languages in general. It is an attempt in India to cover all Indian languages under a single framework. In this system, each language requires a Unicode parser [14] [15].

### 7.2. Shruti: An Embedded Text-To-Speech System for Indian Languages

Mukhopadhyay *et al.* have developed ***Shruti*** system in year 2006 at Indian Institute of Technology Kharagpur. It is a Text-To-Speech system, which has been developed using a concatenative speech synthesis technique. This is the Text-To-Speech system built specifically for two of the Indian languages, namely Bengali and Hindi [16][13].

### 7.3. HP Labs India TTS System

A Hindi TTS system was developed at HP Labs India on a generic TTS framework that it created based on Festival [24]. Festival is available online at http://www.cstr.ed.ac.uk/projects/festival/. This involved extending the Natural Language Processing module of Festival by providing tools that were more appropriate for handling non-European languages like Hindi. It also incorporated methodology and tools for creating TTS speech databases, from best practices for choosing a voice talent to tools for automatic segmentation and annotation of the database [17][18]. The TTS framework is illustrated in Figure 10 [25].

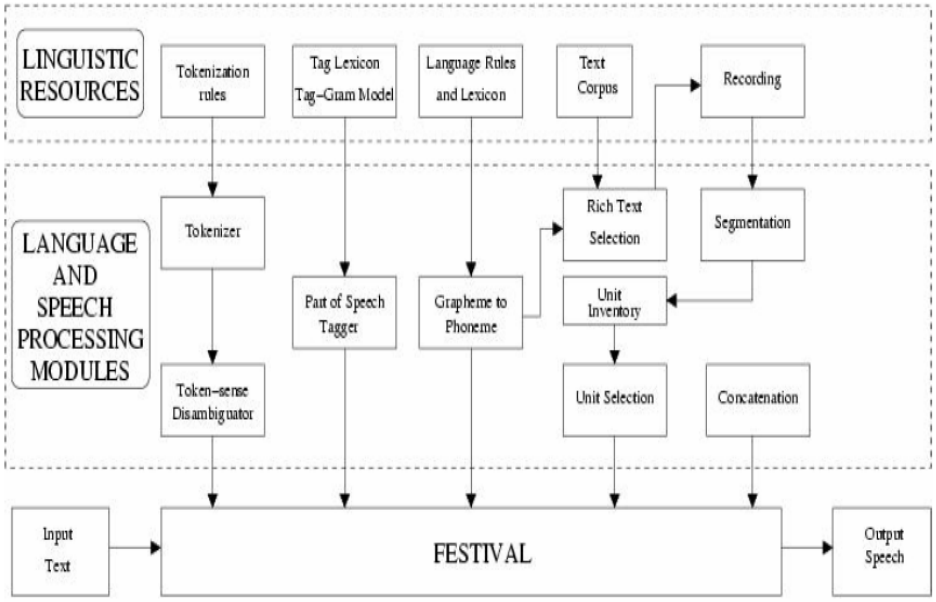


Figure 10. TTS framework created by HP Labs, India

### 7.4. Vani: An Indian Language Text-To-Speech Synthesizer

Vani is an Indian Language Text-To-Speech synthesizer developed at IIT Bombay, India. This system is available online at http://www.cse.iitb.ac.in/vani/. Generally, all existing TTS systems allow user to specify what is to be spoken but does not give any control on how it has to be spoken. In ***Vani***, a new encoding scheme has been introduced called *vTrans*. A *vTrans* file makes a person to encode what text he wants to be spoken and also the way that text to be spoken. A signal processing module is then used to bring out this speech by making appropriate variations to the sound database.

*vTrans* is an XML document that contains a head and a body. In the head part of XML document the parameters and styles are defined. These parameters are pitch, volume and duration. The body section contains several tags. They may be nested, but text is put only in the innermost of the tags. The text within the tags is ITrans encoded. The tags may be any of the parameters defined in the head section. The attributes assigned to these tags determines which style to use and allow the user to scale and translate the function to be used as required [19] [20].

## 8. COMPARISON OF INDIAN LANGUAGES TTS SYSTEMS

In the above sections, four TTS systems have been discussed, namely **Dhvani**, ***Shruti***, ***HP Lab system*** and ***Vani*** system. These systems are compared in Table I on the basis of languages it support, size of the speech units, methodologies used to code, store and synthesize the speech and prosody [26].

## 9. CONCLUSION

This paper describes the process of development of a TTS system for Punjabi Language. A general Architecture and various components, namely, Text Pre-processing, Text Normalization, Linguistic Analysis, Prosody Information and Waveform Generation of this system have been discussed.

Three Waveform generation approaches, namely, Formant based approach, Concatinative based approach and statistical parametric based synthesis have also been presented in this work. An illustration on various existing TTS systems for Indian Languages has also been included in this paper and these systems have been compared on a few parameters. The further scope of this work shall include the development of various components of a TTS system for Punjabi Language and integrating them in a fully functional TTS system.

Table 1. Comparison of Indian Languages TTS systems

| S. No. | Name of the system | Language supported | Synthesis Technique | Database/Unit | Text processing/ Tools | Prsosody | Phonetics |
|---|---|---|---|---|---|---|---|
| 1 | *Dhvani*- Indian Language Text-To-Speech system | Hindi, Malayalam, Kannada, Bengali, Oriya, Punjabi, Gujrati, Telegu. | Diphone Concatination | Syllable Database | Parsing rule for Phonetization | Not supported | Not Used |
| 2 | *Shruti:* an embedded Text-To-Speech system for Indian Languages | Bengali and Hindi | Concatination synthesis | Phoneme Database | Parsing rules for Phonetization | Prosodic and intonational rules are applied | Used |
| 3 | TTS System by HP Labs, India. | Hindi | Concatination synthesis | Phoneme or syllable Database | Festival Based- FestVox tools | Not supported | Not Used |
| 4 | *Vani:* an Indian Language Text-To-Speech synthesizer | Hindi | Concatination Sunthesis | Fract Phoneme Database | Encoding Scheme called vTrams | Parameters to control speech like pitch, volume and duration is given by user | Not Used |

## REFERENCES

[1] D. Klatt, “Review of Text-To-Speech Conversion for English”, Journal of the Acoustical Society of America, JASA, no. 3, vol. 82, 1987, pp. 737-793.

[2] S.P. Kishore, R. Kumar, R. Sangal, "A data-driven synthesis approach for Indian languages using syllable as basic unit", in Proc. International. Conference on Natural Language Processing (ICON), 2002, pp. 311-316.

[3] M.H. O'Malley, "Text-To-Speech conversion Technology", IEEE Computer, vol. 23, 1990, pp. 17-23.

[4] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards., "Normalization of non-standard words", Computer Speech and Language, no. 3, vol. 15, 2001, pp. 287–333.

[5] S. Lemmetty, "Review of Speech Synthesis Technology", MS Thesis, Electrical and Communications Engineering, Helsinki University of Technology, 1999.

[6] S. Thomas, "Natural sounding Text-To-Speech synthesis based on Syllable-like units", MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology, Madras, 2007.

[7] 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, no. 3, vol. 2, 2011, pp. 322-326.

[8] P. Chaudhury, M. Rao, K.V. Kumar, "Symbol based concatenation approach for Text-To-Speech System for Hindi using vowel classification technique", in Proc. World Congress on Nature and biologically Inspired computing, 2002, pp. 1082 – 1087.

[9] S.P. Kishore, A.W. Black, "Unit Size in Unit Selection Speech Synthesis", in Proc. EUROSPEECH 2003, Geneva, Italy, 2003.

[10] A.J. Hunt, A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", in Proc. IEEE international Conference on Acoustics, Speech and Signal Processing, Atlanta, GA , USA , vol. 1, 2011, pp. 373 – 376.

[11] N.S. Krishna, H.A. Murthy, T.A. Gonsalves, "Text-To-Speech in Indian Languages", in Proc. International Conference on Natural Language Processing, ICON-2002, Mumbai, 2002, pp. 317-326.

[12] A. W. Black and K. Tokuda, "The Blizzard Challenge - 2005: Evaluating corpus based speech synthesis on common datasets", in Proc. INTERSPEECH, Lisbon, Portugal, 2005, pp. 77–80.

[13] A. Basu, D. Sen, S. Sen and S. Chakraborty, "An Indian Language Speech Synthesizer – Techniques and Applications", in Proc. National Systems Conference, NSC 2003, Indian Institute of Technology, Kharagpur, India, 2003, pp. 217-223.

[14] S. Thottingal, (december 2012) "Dhvani Indian Language Text-To-Speech System", http://foss.in/2007/register/slides/Dhvani/, 2007.

[15] R. Hariharan, (April 2013) [Online], http://dhvani.sourceforge.net/, 2007.

[16] A. Mukhopadhyay, S. Chakraborty, M. Choudhury, A. Lahiri, S. Dey, A. Basu, "Shruti- an Embedded Text-To-Speech System for Indian Languages", in Proc. IEEE Proceedings on Software Engineering, no. 2, vol. 153, 2006, pp. 75–79.

[17] A.G. Ramakrishnan, K. Bali, "Tools for the Development of a Hindi Speech Synthesis System", in Proc. 5th ISCA Speech Synthesis Workshop, Pittsburgh, 2004, pp. 109-114.

[18] A.K. Singh, "A Computational Phonetic Model for Indian Language Scripts", in Proc. Fifth International Workshop on Writing Systems, Nijmegen, Netherlands, 2006.

[19] H. Jain(April 2013) [Online]. http://www.cse.iitb.ac.in/vani/, 2004.

[20] H. Jain, V. Kanade, K. Desikan, "Vani-An Indian Language Text-To-Speech synthesizer", Project Report, Department of Computer Science and Engineering, IIT, Bombay, 2004.

[21] A. W. Black and P. Taylor, "Assigning Intonation Elements and Prosodic Phrasing for English Speech Synthesis from high level Linguistic input", in Proc. of ICSLP, Yokohama, Japan, 1994, pp. 715–718.

[22] P. Ladefoged and K. Johnson, A course in Phonetics, Cengage learning, 2010.

[23] P. Sharma, "Automatic Identification of Silence, Voiced and Unvoiced Chunks in Speech", ME Thesis, SMCA, Thapar University, 2012.

[24] Alan black and Rob Clark, (April 2013) [Online] http://www.cstr.ed.ac.uk/projects/festival/, 2004.

[25] Kalika Bali, N. Sridhar Krishna, Sameer Badasker, KSR Anjaneyulu, "Enabling IT Usage through the Creation of a High Quality Hindi Text-to-Speech System", Project Report, HP Laboratories, India, 2007.

[26] S. Gupta, "Hindi Text To Speech System", ME Thesis, Computer Science and Engineering Department, Thapar University, Patiala, 2012.

[27] P. Singh and G.S. Lehal, "Corpus Based Statistical Analysis of Punjabi Syllables for Preparation of Punjabi Speech Database", International Journal of Intelligent Computing Research (IJICR), No. 3, Volume 1, June 2010.

[28] L.L. Cherry and W. Vesterman., "Writing tools - the STYLE and DICTION programs", 4.3 BSD UNIX System Documentation, University of California, Berkeley, 1991.

[29] A. Mikheev., "Periods, capitalized words, etc." Computational Linguistics, No. 3, Volume 28, 2002, pp. 289-318.

[30] A. Voutilainen., "A syntax-based part of speech analyser." in Proc. of the Seventh Conference of the European Chapter of the Association for Computation al Linguistics, Dublin. Association for Computational Linguistics, 1995, pp. 157.164.

[31] F. Jelinek., "Markov source modeling of text generation." The Impact of Processing Techniques on Communications, NATO ASI series, 1985, pp. 569.598.

[32] E. Brill., "Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging." Computational Linguistics, No. 4, Volume 21, 1995, pp. 543-566.

## AUTHORS

Rupinderdeep Kaur is Lecturer in Computer Science and Engineering Department at Thapar University, Patiala. She has more than 6 years of academic experience. She has done her B. Tech from Chandigarh Engineering College and M.E. in Software Engineering from Thapar University. She is Persuing her Ph.D. in "Prosody based Text-To-Speech System for Punjabi Language" from Thapar University.

R.K. Sharma is professor in the Computer Science and Engineering Dapartment at Thapar University, Patiala. He has more than 23 years of teaching experience. He has earned his Ph.D. degree in 1993 from IIT Roorkie on “Computer-Aided Simulation Studies Exploring Efficient Mixing-Type Estimators”. He has published more than 90 research papers in Journals, Conferences and Magazines of repute. He has Guided 15 students for PhD work.

Parteek Kumar is Associate Professor in the Computer Science and Engineering Dapartment at Thapar University, Patiala. He has more than 18 years of academic experience. He has earned his B.Tech from SLIET, MS from BITS Pilani, Ph.D from Thapar University. He has published more than 50 research papers and articles in Journals, Conferences and Magazines of repute.