TEXT TO SPEECH SYSTEM OF INDIAN LANGUAGE: REVIEW

Manoj M.Khandare[1] , Shaikh Abdul Hannan[1], R.R. Manza[2] R. J. Ramteke[3]
[1]Vivekanand College, Aurangabad, India (hannan_7us@yahoo.com)
[2]Department of Computer Science and Information Technology, Dr. BAMU, Aurangabad (rameshmanza@yahoo.com)
[3]Department of Computer Science and IT, Reader, Jalgaon (rakeshramteke@gmail.com)

Abstract:

In this paper, we introduce to technique for Text To Speech in Indian Languages and factor involve in Offline and Online phases of TTS system. OFFLINE phase includes pre-processing, segmentation and pitch marking. ONLINE phase includes text analysis and synthesis. This paper also involves the different method for pre-processing text and Speech Synthesis methods. This paper present the salient features and Duration methodology used in TTS for Indian languages.

Introduction:

The objective of a text to speech system is to convert an arbitrary given text into a corresponding spoken waveform. Text processing and speech generation are two main components of a text to speech system. The objective of the text processing component is to process the given input text and produce appropriate sequence of phonemic units. These phonemic units are realized by the speech generation component either by synthesis from parameters or by selection of a unit from a large speech corpus. For natural sounding speech synthesis, it is essential that the text processing component produce an appropriate sequence of Phonemic units corresponding to an arbitrary input text.[2][3][4][5]

Keywords:

Speech synthesis, Decision rule, Duration modeling Indian language scripts, Akshara mapping.

Speech Synthesis

The techniques employed for synthesizing speech from text may be broadly classified into three categories:

I. Formant-based

II. Parameter-based

III. Concatenation-based

In the formant-based approach, we can synthesize a signal by passing the global periodic waveform through a filter with the formant frequencies of the Vocal tract. It makes use of the rules for modifying the pitch, formant frequencies and other parameters. However, the technique fails to produce good quality, natural sounding speech, as appropriate rules cannot be

derived for unlimited speech. As the model uses a number of resonators, it is computationally expensive. On the other hand, in parameter-based synthesizers, the waveforms are modeled using Linear Prediction (LP) coefficients. These LP coefficients fail to model nasals perfectly. The linear prediction model is an all-pole model which models vowels exceptionally well, but fails to model the nasals and silence (stops). In **concatenate synthesis** the speech units are usually words, syllables, demi-syllable, phonemes, and sometimes even triphones. In the present system partnemes are mainly used as units. Partnemes are waveforms which are transitory from a consonant to a vowel (CV), from a vowel to a consonant (VC) or only consonant or only vowel. A partneme that is a vowel, has only one pitch period stored, called the epoch, since only one epoch can completely characterize a vowel. The partnemes are spliced out from prerecorded utterances taking into consideration the combination of all vowels and consonants. Given a word, sentence or any text the synthesized speech is generated through a number of steps which are described in the remaining part of this section.[3][4][5][6][7][8]

Algorithmically converting an input text into speech waveforms and some previously coded speech data. Speech synthesizers can be characterized by the size of the speech units as well as by the method used to code, store and synthesize the speech.[5][6]

In the first step, whatever text is required to be spoken is passed to **a Natural Language Processor (NLP**). The NLP analyses the text and performs grapheme to phoneme conversion. A grapheme is the actual text whereas the phoneme is a token that directly maps to a signal unit in the voice database or the partneme dictionary. This grapheme to phoneme conversion requires morphological and phonological analysis . So given text as input to the NLP the output obtained consists of a string of phonemes. Given the name of the city of

'Aurangabad' as input to the NLP, the resulting output has been shown the input grapheme string shown in English exactly corresponds to the

grapheme string shown in Hindi or Marathi above it. Note that the vowel 'a' does not have a separate grapheme representation.[3][6][7]

The **TTS output** thus obtained so far from the algorithms described previously is flat and expressionless even though it is natural. This is because the way of utterance, i.e. the intonation and prosodic variations, are an integral part of natural speech. So the system must also take intonation and prosody into consideration. Thus before concatenating the units the base frequencies, amplitude and duration are modified depending on previously observed patterns of utterances. These patterns are based on certain rules which have been extracted by studying natural speech.[8][9][10][11]

Speech synthesis in Indian Languages:

In Indian languages, both CONCATENATIVE and FORMANT synthesis techniques are used. A text processor first accepts text as input and outputs phonemes and prosody markers. Next speech is generated by concatenating di-phone like segments of recorded natural speech in the former, and by a production model based on "formant"(resonance)frequencies in the latter. A major problem of concatenate method is discontinuity at segment boundaries. In formant synthesis it is difficult to capture the stop consonant to vowel transitions by rules. Ongoing efforts can stick to the respective areas of expertise.

The speech synthesis also involves the major issues of speech like, segment inventory and selection boundaries of the context that can be selected. Manual segmentation of a large database needs considerable skilled man hours and may delay the effort; there should be some objective criteria to select the 'most suitable 'segment, at a given context. The HHM (Hidden Markov Model) is used to automatic labeling of given segment. HHM can also be used to automatically select the optimal set of diphone and polyphone unit to be used by checking for the spectral stability at the segment boundaries. Discontinuity at Segment Boundary is a problem in concatenation synthesis; sudden feeling of discontinuity can reduce the smoothing. The Wave Interpolation (WI) is used for removing the hardness of speech, a few frame at the boundaries (1-3 pitch on each side) are removed and reconstructed by WI. This ensures smooth transition between segments. Prosody has three elements: Pitch, Intensity and Duration. Modulating duration and intensity is easy in both concatenation and formant syntheses. But in concatenation synthesis with time domain approach, pitch change is tricky. An easy way is to 'stretch' or compresses each pitch cycle in inverse ratio of pitch.[8][9][10][11]

Methods and Technologies used in TTS:

Nature of Indian Language Scripts:

The scripts in Indian languages have originated from the ancient Brahmin script. The basic units of the writing system are referred to as *Aksharas*. The properties of Aksharas are as follows:

(1) An Akshara is an orthographic representation of a speech sound in an Indian language;

(2) Aksharas are syllabic in nature;

(3) The typical forms of Akshara are V, CV, CCV and CCCV, thus have a generalized form of C*V. The shape of an Akshara depends on its composition of consonants and the vowel, and sequence of the consonants. In defining the shape of an Akshara, one of the consonant symbols acts as pivotal symbol (referred to as semi-full form). Depending on the context, an Akshara can have a complex shape with other consonant and vowel symbols being placed on top, below, before, after or sometimes surrounding the pivotal symbol (referred to as half-form). Thus to render an Akshara, a set of semi-full or half-forms have to be rendered, which in turn are rendered using a set of basic shapes referred to as *glyphs*. Often a semi-full form or half-form is rendered using two or more glyphs, thus there is no one-to-one correspondence between glyphs of a font and semi full or half-forms. [16][18][24]

Digital Storage of Indian Language Scripts

There is a chaos as far as the text in Indian languages in electronic form is concerned. Neither can one exchange the notes in Indian languages as conveniently as in English language, nor can one perform search easily on texts in Indian languages available over the web. This is because the texts are being stored in ASCII font dependent glyph codes as opposed to Unicode. The glyph coding schemes are typically different for different languages and within a language there could exists several font-types with their own glyph codes (as many as major news portals in a language).[10][15][16][18][24]

A Phonetic Transliteration Scheme for Digital storage of Indian Language Scripts

To handle diversified storage formats of scripts of Indian languages such as ASCII based fonts, ISCII (Indian Standard code for information

Interchange) and Unicode etc, it is useful and becomes necessary to use a meta-storage format. A transliteration scheme maps the Aksharas of Indian languages onto English alphabets and it could serve as met storage format for text-data. Since Aksharas in Indian languages are orthographic represent of speech sound, and they have a common phonetic base, it is suggested to have a phonetic transliteration scheme such as **IT3**. Thus when the font-data is converted into IT3, it essentially turns the whole effort into font-to-Akshara conversion.[10][15][16][18][24]

Identification of Font -Type:

To identify the font-type of a given test font-data, the steps involved are as follows: 1) Generate the terms (glyph sequences) of the test font-data2) Compute the relevancy scores of the terms and for each of the document (font-type) using the corresponding TF-IDF weights of the terms 3) The test font-data belongs to the document (font-type) which produces a maximum relevancy score. The performance of TF-IDF approach for identification of font-type was evaluated on 1000 unique sentences and words per font-type. We have added English data as also one of the testing set, and is referred to as English-text. The performance of font-type identification system using different terms *single* glyph, *current and next* glyphs, *previous, current and next* glyphs and it could be observed that the use of *previous, current and next* glyphs as a term provided an accuracy of 100% in identification of font-type even at the word level.

Font –to –Akshara mapping:

Font-data conversion can be defined as converting the font encoded data into Aksharas represented using phonetic transliteration scheme such as IT3. As we already mentioned that Aksharas are split into glyphs of a font, and hence a conversion from font-data has essentially to deal with glyphs and model how a sequence of glyphs are merged to form an Akshara.It has two phases, in the first phase we are building the base-map table for a given font-type and in the second phase forming and ordering the assimilation rules for a specific language.

Building a Base-MapTable for a Font-type: (phase 1)

The base-map table provides the mapping basic between the glyphs of the font-type to the Aksharas represented in IT3 transliteration scheme. The novelty in our mapping was that the shape of a glyph was also included in building this mapping table.

Building Pronunciation Models for Aksharas:

A framework based on machine learning techniques where pronunciation of Aksharas could be modeled using machine learning techniques and using a small set of supervised training data..

Use of Contextual Features

Contextual features refer to the neighbor phones in a definite window-size/level. Using the contextual features, experiments were performed for various Contextual Levels (CL).A decision forest was built for each phone to model its pronunciation. A decision forest is a set of decision trees built using overlapping but different sub-sets of the training data and it employs a majority voting scheme on individual prediction of different trees to predict the pronunciation of a phone. Table shows the results of pronunciation model for Hindi, Bengali and Tamil using various level of contextual features. We found that that a context level of 4 (i.e., 4 phones to the left and 4 phones to the right) was sufficient to model the pronunciation and moving beyond the level of 4, the performance was degraded.

Languages	Context Level				
	1	2	3	4	Mean
Hindi	90.2%	91.44%	91.78%	91.61%	91.25
Bengali	82.77%	84.48%	84.56%	83.56%	83.84
Tamil	98.16%	98.24%	98.10%	98.05%	98.13

Table 1 . Pronunciation Model with Contextual features

Decision Rules for selection of Allophones of Marathi Affricates:

Decision tree learning methodology is used to identify factors that influence the choice of appropriate allophone. We have to predict the place of articulation with high accuracy to develop automatic speech recognition for TTS system for Indian language. Two main stages in the operation of an unlimited vocabulary text to speech system are conversion of

1. Text to Phoneme
2. Phoneme to speech

Devnagari script is used by some modern technologies such as Hindi and Marathi. There is a near to one to one correspondence between

grapheme and phoneme. The exceptions to these correspondences are primarily due to schwa deletion. In order to synthesize speech that is acceptable to native speakers, a text to phoneme module to take into account allophones of the language in addition to such a phonological rules. For example, the pronunciation of the phoneme/a/ in the Hindi word "phlao " is different from its canonical pronunciation.[1][2]

There are four affricates and the corresponding graphemes in most Indian languages. The place of the articulations of the affricate phonemes is palatal. The affricates are categorized by the binary values of the two distinctive features. 1) Voicing 2) Aspiration
Here we focus on the rules of pronunciation of Marathi unvoiced affricate using data driver approach.[2][17][18][28]

Here we set rules that will add in identifying the place of articulation of affricates in Marathi word when its orthography is given. We hypothesis that allophone of Marathi affricate is chosen such that the place of articulation of the allophone is chosen to adjacent phonemes. According to reference (2) palatal affricates occur before the vowels I, ii, e, ai, and au where dental affricates occur before the vowels u, uu and o ; there is no such rule when affricate précised the vowel a and aa. An example of the letters case is contrast between the palatal affricates in the word "caar (four)" and the dental affricate in the word "caara (fodder)", affricates in the both cases are followed by /a:/ and both occurs at the word initial positions. Since the linguistic solution to this problem is not available. In this paper we describe the method and the software used for generating rules set for lexical attributes of the phoneme used for decision and text corpus.

The inputs to the system are values of set of attribute (articulatory and lexical properties) of a word and correct place of articulation (the truth value). A detection tree is well suited for this purpose to takes input as set of properties of objects and outputs as yes/no decision. In some cases there is confusion among the native speakers as to whether to use dental or palatal affricates. Moreover, the decision tree can be re-cased as sets of if-the rules. This property of decision module is very useful for incorporation into text to phone module of text to speech system.[14][29][30]

1) Decision tree tool kits: A public domain decision tree constriction tool "c4.5" was used. It uses Quinlan's ID3 algorithm for construction a decision tree. This construction needs input into the form of two input file, first file contains a list of attributes and the second (data) files contains the value of these attributes.

Database: In this method, TDIL Marathi text corpus is used. This database has 465 files containing text drawn from diverse sources. This corpus has about 60,000 Marathi unique words containing unvoiced unaspirated /c/ represented by the script "ca". The TDIL Marathi text corpus (in ISCII format) we Romanized for processing bycomputer in Linux environment. This facilitated the process of

a) Selection of words containing the phoneme /c/ and
b) Creation of words set with desired characteristics.

Decision Rules:
In this method, "C 4.5 "decision tree construction tool can generate both unpruned and pruned decision tree for a given training data. The size of (numbers of nodes) of pruned tree is (41) is much smaller than that of unpruned tree (83).
The place of the articulation of the allophone is Dental if (the phone /c/ is followed by/e/, /a/, or /A/) otherwise Palatal. When data set 2 is used for constricting the decision tree, the single affricate rule is "The place of the articulation of the allophone is dental if the following phoneme is not a front vowel", this is articulatory principle.

Duration knowledge for text to speech conversion system for Telugu

Naturalness of speech is achieved by prosodic features (incorporating supra segmental features) It involves the duration of basic units intonation patterns and stress. The duration factor is the one of the important factor of naturalness of speech.
Duration information can be analyzed by the basis of two factors.

1) Positional factors
2) Contextual factors

For deriving the duration information position factors have more effect than the contextual factors.
Following attributes is responsible for lack of naturalness in synthetic speech.
Inappropriate modeling of the physical acoustic properties of the vocal tract. (Flangan, 1972)

- Incorrect modeling of articulatory and coarticulatory properties of natural speech. (Strevens and Bickley, 1991)
- Failures in modeling the prosodic structure of natural speech. (Akers and Lennig, 1985).

- Above factors play the major role in prosodic information?

Variation in duration Pitch and stress are presented prosodic features (Supro segmental features) synthesis speech.

There are various factors which affect duration of basic units this are mainly classified into positional and contextual factors.

Positional factors affect the duration of basic units according to position of the unit in the text.

Different position of units having different effect on duration factor. The duration units are

1) Word final position
2) Syllable boundary
3) Phrase boundary
4) Sentence ending position
5) Word initial position

Duration of the basic units depending on the context in which the unit is present. Contextual factor include the effect of the nature of the preceding and the following units on the present unit. [K.Kiran Kumar].

There are other factors which also affect on duration. The gender of speaker is one of the factor affecting an duration (male, female). The psychological state of the speaker (happy, fear, anger, sad, natural) and the age factor is also considered.

Measurement of Duration

The two factors are mainly affected on duration of the basic units.1) Positional factor

2) Contextual factor[18][28]

For measuring the duration of basic unit we have to measure the base duration of units, A effect of positional factors, duration modified due to effect of contextual factor.

Measurement of Base-Duration

Reference duration which we called as base duration to represent the changes occurring in duration due to any factor. Base duration should not have any effect of any factor like positional or contextual.

Basic Unit	Word	Basic Unit classes	DB (MS)	BD (MS)	PC
/DU/	/naTuDu/	Voiced unit	78	67	16
/Ka/	/Sainika/	Unvoiced unit	107	86	25
/la/	/digumatyla/	Liquids	103	70	47
/ma/	/tama/	Nasals	99	76	30
/rA/	/dvArA/	Trills	126	101	24

Table 2: Duration of basic unit, base duration percentage change abbreviations DB, BD, PC resp.

with off position factor of units.

Measurement of the duration modified due to the effect of positional factors. The effect of positional factors on the basic units for different manners of articulation and voicing is not the same. For that we grouped the basic unit into six classes, unvoiced, voiced, liquids, nasal, trills and semi vowels. The duration of basic unit can measure when they are affected by any of the positional factor. The basic unit of all the above classes showing a similar tread when they are in word initial position. When a basic unit in a word re in final position only then it will considered separately.

When a basic unit is in middle of ta word then the change in duration is considered by contextual factors.[1][2][17][18][28]

Measurement of the duration modified due to the effect of contextual factors

The effect of positional factor on duration is more than that of contextual factor. The effect of both represented as a set of IF-THEN rules, the activation knowledge during synthesis is achieved by means of a rule interpreter. [1][2][17][18][28][29][30]

Speaker	Basic Unit	Word	DB	Units Types	Difference in two speaker	average
S1	/DU/	/natudu/	78	Vo	03	76.5
S2	/Du/	/mUDu/	75			
S1	/Ka/	/Sainika/	107	Un	07	110.5
S2	/ka/	/naitika/	114			
S1	/la/	/digmatula/	103	Li	49	
S2	/lA/	/idilA/	152			127.5
S1	/na/	/chEcina/	91	Na	12	97
S2	/na/	/Jarigina/	103			
S1	/ra/	/itara/	83	Tri	04	85
S2	/ra/	/amara/	87			

Table. 3 : Analysis of Duration of some basic unit due to effect of contextual and positional factor speaker 1 and speaker 2 on same unit.

Conclusion: In this paper, we introduce to technique for Text To Speech in Indian Languages and factor involved in Offline and Online phases of TTS system. OFFLINE phase includes pre-processing, segmentation and pitch marking. ONLINE phase includes text analysis and synthesis. Also we describe Methods and Technologies used in TTS like Nature of Indian Language Scripts, Digital Storage of Indian Language Scripts, and A Phonetic Transliteration Scheme for Digital storage of Indian Language Scripts, Identification of Font –Type, Font –to – Akshara mapping, Building a Base-Map Table for a Font-type. Further the introduction to Duration knowledge for text to speech conversion system for Telugu and contextual factor affected on duration factor is described.

References:

1.Gopinath, D.P.; Divya Sree, J.; Mathew, R.; Rekhila, S.J.; Nair, A.S. Information Technology, "Duration Analysis for Malayalam Text-To-Speech Systems", ICIT apos;06. 9th International Conference on Volume , Issue , 18-21 Dec. 2006 Page(s):129 – 132.

2.Farbod Razzazi, Abolghassem Sayadiyan. – "Soft Segment Modeling, a Robust Duration Modeling for speech recognition". [International Conference on Knowledge Based Computer Systems].

3.G.L.Jayavardhana Rama, A.G. Ramakrishna Thirukkural – "A text to speech synthesis system".

4.Scottish Gaelic "Approach to speech synthesis"

5. R .Marshall, S.Furui and M.M.Sondhi (Eds) "Training, Multi-lingual Transcription and Linguistic Identity" [International Conference on Knowledge Based Computer Systems] .[Furui and Picone,1989] .Advances in speech signal processing .Marcel Dekker, Inc. NY, 1991.

6. Aniruddha Sen, "Text to Speech Conversion in Indian English", *Proc. KBCS 2000,* pp.564-575, Mumbai, Dec 2000.

7. L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition"**,** *Proceedings of the IEEE*, Vol. 77, No. 2. February 1989.

9. F. Razzazi and A. Sayyadian, "Data Refining HMM, A New Approach to HMM Based Speech Recognition System Improvement", *Proceedings of Forum Acusticum, Seville*, August 2002.

10. L.R. Rabiner, "A Tutorial On Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of The IEEE, Vol. 77, No. 2, Feb. 1989, PP. 257-286.

11. M. Ostendorf, V. Digalakis, O. Kimball, "From HMM To Segment Models: A Unified View of Stochastic Modeling of Speech Recognition", IEEE Transactions on Speech and Audio Processing, Vol. 4, No. 5, Sep. 1996, PP. 360-378.

14. F. Razzazi and A. Sayyadian, "Soft Segment Modeling, A New Approach in Duration Modeling for Speech Recognition**,"** Proceedings of ninth Australian Conference of speech Science and Technology, Melbourne, December 2002.

15.Carison, R. and B. Granstrom, "A search for durational
rules in real speech database.," Phonetica, vol. 43, pp. 140–154, 1986.

16. van Santen, J. P. H., “Contextual effects on vowel durations., ”Speech Communication, vol. 11, pp. 513–546, 1992.
17.Bartkova, K. and C. Sorin, “A model of segmental duration for speech synthesis in french.,” Speech Communication, vol. 6, pp. 245–260, 1987.
18. Simoes, A.R.M., “Predicting sound segment duration in connected speech: An acoustical study of brazilianportugese.,” In Workshop on Speech Synthesis, ESCA, Autrans., pp. 173–176, 1990.
19.Riley, M.D., “Tree-based modeling for speech synthesis.”In: G. Bailly, C. Benoit, and T. Sawallis Eds.),Talking machines: Theories, models and designs,, pp.
265–273, 1992.
20.Hyunsong Chung and Mark A. Huckvale, “Linguistic factors affecting timing in korean with application to speech synthesis,” in Euro speech, Denmark, 2001.
21.van Santen, J.P.H., “Assignment of segmental durationin text-to-speech synthesis.,” *Computer Speech andLanguage*, vol. 8, pp. 95–128, 1994.
22.Campbell, W., “Syllable-based segmental durations.,”In: G. Bailly, C. Benoit, and T. Sawallis (Eds.), Talking machines: Theories, models and designs,, pp. 43–60,1992.
23. Mitchell, T.M., Machine Learning, McGraw-Hill, New York, 1997.
24.Malfrere, F. and T. Dutoit, “High quality speechsynthesis for phonetic speech segmentation.,” in Eurospeech,Rhodes, Greece, 1997, pp. 2631–2634.
25.Cassidy, S., The EMU Speech Database System,http://www.shlrc.mq.edu.au/emu/, 2002.
26. Lee, S. and Y.H. Oh, “Tree-based modeling of prosodic phrasing and segmental duration for korean
tts systems.,” Speech Communication, vol. 28, pp.283–300, 1999.
27.Taylor, P., R. Caley, and A.W. Black,The Edinburgh Speech Tools Library,1.2.1 edition, University of Edinburgh,http://www.cstr.ed.ac.uk/projects/speechtools.html,2002.
28.Simoes, A.R.M., “Predicting sound segment durationin connected speech: An acoustical study of Brazilian Portugese”, In Workshop on Speech Synthesis,ESCA, Autrans, pp. 173-176, 1990.
29.Riley, M.D., “Tree-based modeling for Speech synthesis”, In: G. Bailly, C. Beno it, and T. Sawallis (Eds.), Talking machines: Theories, models and designs,
pp. 265-273, 1992.
30. Hyunsong Chung andMark A. Huckvale, “Linguistic factors affecting timing in Korean with application
to speech synthesis”, in Eurospeech, Denmark, 2001.
31.van Santen, J.P.H., “Assignment of segmental duration in text-to-speech synthesis”, Computer Speech and Language, vol. 8, pp. 95-128, 1994.
32.Campbell, W., “Syllable-based Segmental Durations”, In: G. Bailly, C. Beno it, and T. Sawallis (Eds.), Talking machines: Theories, models and designs, pp. 43-60, 1992.
33. Mitchell, T.M., Machine Learning, McGraw-Hill, New York, 1997.
34.Boersma, P. and D. Weenik., “Praat: A System for Doing Phonetics by Computer”, (http://www.praat.org/), 2001.
35.Ramakrishnan, A.G. et.al., “Tools for the Development of a Hindi Speech Synthesis System”, In 5th ISCA Speech Synthesis Workshop, Pittsburgh, pp. 109–114, 2004.
36. Roger Tucker, “Local Language Speech Technology Initiative (LLSTI)”, (http://www.llsti.org),
2003.
37. Lee, S. and Y.H. Oh, “Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems”, Speech Communication, vol. 28, pp. 283-300, 1999.
38. Taylor, P., R. Caley, and A.W. Black,
“The Edinburgh Speech Tools Library”,
1.2.1 edition, University of Edinburgh,
http://www.cstr.ed.ac.uk/projects/speechtools.html,2002.