www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume – 4 Issue - 12 December, 2015 Page No. 15303-15311

Overview of Hidden Markov Model for Test-To-Speech Synthesis Methods

Sangramsing N. Kayte , Monica Mundada
Department of Computer Science & IT Dr. Babasaheb Ambedkar Marathwada University,Aurangabad.
bsangramsing@gmail.com , monicamundada5@gmail.com

Abstract: Speech synthesis is the process of production of artificial speech. The system used for generation of speech from text is called as text-to-speech (TTS) system. In TTS system, text and voice models for a particular language or multiple languages are given as input to the system, which generates speech as output corresponding to the provided voice models. Speech synthesis systems can be extremely useful to people who are visually challenged, visually impaired and illiterate to get into the mainstream society. More recent applications include spoken dialogue systems and communicative robots. HMM (Hidden Markov Model) based Speech synthesis is the emerging technology for TTS. HMM based speech synthesis system consists of training phase and synthesis phase. In the training part, phone and excitation parameters are extracted from speech database and modeled by context dependent HMMs. In synthesis part, the system will extract the suitable phone and excitation parameters from the previously trained models and generates the speech.
Keywords: HMM TTS.

1. Introduction

The primary goal of Text-to-Speech (TTS) synthesis is to convert an arbitrary input text into intelligible and natural sounding speech. A TTS system trained for a particular language can be used to synthesize arbitrary text in that language. The accuracy of speech synthesis system depends on the quality of the recorded speech data base and the speaker. TTS uses linguistic analysis for correct pronunciation, prosody (pitch, duration etc.,) and acoustic representations of speech to generate waveforms.

TTS system includes two main components: the front-end and the back-end. The front-end is the part of the system closer to the text input which is responsible for text analysis where conversion of ambiguous symbols like dates to their equivalent word format and grapheme to phoneme conversion takes place. The back-end is the part of the system that is closer to the speech output which converts the output of the front-end (phonetic transcriptions and prosodic information) to the corresponding waveform.

Naturalness And Intelligibility

There are two properties for any TTS system. One is naturalness which is the degree to which the synthesized speech sounds close to speech uttered by humans. The other is intelligibility which is the degree of ease with which people understand the synthesized speech. Understandability is sometimes used in the place of intelligibility. These properties can also be added by three other concepts called flexibility, pleasantness and similarity to original speaker [1].

Flexibility refers to how well the system handles symbols which need translation. For example, time phrases, out-of vocabulary words etc. Pleasantness deals with the desirability and pleasure that one associates with listening to the synthesized voice sound. Similarity to original speaker deals with how close the synthesized voice compares to that of the original speaker [2].

The quality of a speech synthesis system is often determined by using these parameters. From the parameters described above, it is clear that they are not independent from one another and the accuracy of one parameter influence remaining other parameters. If the intelligibility is less, this will be perceived as less natural and the pleasantness of the synthesized speech becomes worse.

3 Speaker Dependent and Speaker Independent Systems

Speech synthesis systems can be trained for speaker independent, speaker dependent or adaptive platforms. A speaker dependent system is one that is trained on data from one particular speaker. A

speaker independent system is trained on data from several speakers and can be used to synthesize text using any of the trained voices. An adaptive TTS synthesis system is one that allows a new speaker to be adapted based on trained data of a speaker independent system using only minimal data from the target speaker.

3.1 Limited Domain and Open Vocabulary

The TTS synthesis systems can be developed for either limited domain or open vocabulary platforms. Limited domain speech synthesis systems are those trained using data from a particular domain for example medicine to be used only for purposes relating to that domain. Such systems have been proven to exhibit high performance, naturalness, intelligibility, and low Word Error Rates (WERs). Although, limited domain systems do not require a huge database, they cannot synthesize words which are not in their lowercase database. An open vocabulary TTS system is one that is trained on general purpose data from a particular natural language to be used for general purpose applications. Unlike limited domain systems, open vocabulary systems are flexible in that they can synthesize even words not in their database. Open vocabulary systems, however, often require a huge database, more training data produce less natural speech than that produced by limited domain systems, and their Word Error Rates (WERs) are often higher than those of limited domain systems. In case of unit selection systems, to achieve good quality synthesis, the speech unit database should have good unit coverage. With more data, it is more likely that a database will contain a unit that is closer to the target unit. However, it is relatively difficult to collect and segment large amount of speech data for different languages and storing of large database in small memory devices is not possible [1] [2].

3.2 Text-to-Speech (TTS) Architecture

The architecture of TTS system consists of mainly four modules as a whole. It includes text normalization module, linguistic analysis module, and prosodic analysis module and waveform synthesizer. The first three modules i.e., text normalization and linguistic analysis modules are considered as front-end of the TTS system and the last module is considered as back-end of TTS system. The front-end provides a symbolic linguistic representation of the text in terms of phonetic transcription and prosody information. The back-end often referred to as the "synthesizer" converts the symbolic linguistic representation into sounds [3].

3.3 Text Normalization Module

Text-to-Speech synthesizer works internally by synthesizing words. The input text not only contains ordinary words but also Non Standard Words (NSWs) which include numbers (ex.15), date (ex. 3/4/2010), acronyms (ex.USA), abbreviations (ex. Mr. and Dr.) symbols (ex. $), etc. We cannot find NSWs in a dictionary. All these NSWs must be first converted to actual words and then synthesized. This conversion takes place internally within the synthesizer. Such conversion is called „Text Normalization" [4].

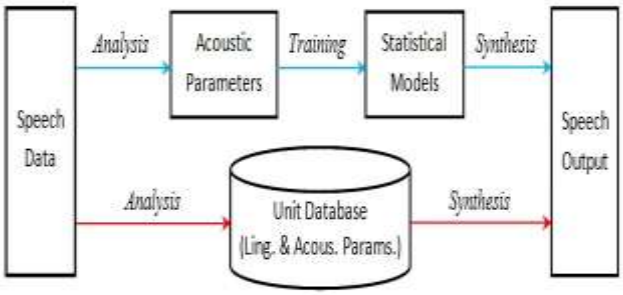

Figure 1: Block diagram of Text-to-Speech (TTS) system

Text normalization module identifies numbers, abbreviations (Mr. and Dr.), acronyms (UNESCO, IBM), dates (12/12/12), special symbols (&, %) etc., and transforms them in to full text. For example Dr. Smith is converted as „Doctor Smith" and the special symbol „&" as „and". The TTS system should not misinterpret the dot after „„in"" in the example „„the table is 36.5 in. long"" as the end of the sentence. In addition, punctuation and other special characters can be part of a time (e.g., 7:30 pm), or date (e.g., 5/25/2004), or currency expression ($10 =„„ten dollars""). Text normalization is difficult because it is context sensitive (e.g., $1.5 million = „„one point five million dollars"") [2]. NSWs are different from standard words. For abbreviations such as Pvt (private), one needs, in effect, to recover the missing letters and then pronounce the resulting word. NSWs have a much higher tendency than ordinary words to be ambiguous. This ambiguity affects not only what the NSWs denote, but also how they are read. The correct reading of „IV" could be four, fourth or I.V (intravenous). 1750 could be read as seventeen fifty or seventeen hundred (and) fifty (or one thousand seven hundred (and) fifty) as a cardinal number. For TTS systems, the primary consideration is how the

NSW is pronounced. Bulk of work on "text normalization" in most TTS systems is done using hand constructed rules. For example, in various applications of the AT&T Bell Labs TTS system, it was important to be able to detect and pronounce (U.S. and Canadian) telephone numbers correctly. Expansion of non-standard words is accomplished by some combination of rules. Hence, a telephone number detector was included as part of the text preprocessing portion of the system. Ambiguous expansions, for example „St." as Saint or Street are usually handled by rules that consider features of the context. In this case, if the following word begins with a capital letter then it is quite likely that the correct reading is „Saint" (Saint John). If the previous word starts with a capital letter, the correct reading is quite likely „Street". Sense disambiguation techniques are developed to handle ambiguous words like crane (a bird, vs. a piece of construction equipment). These previous approaches say that these are often impractical, especially when one is moving to a new text domain. In some domains, such as in real estate classified ads: the example below is taken from the New York Times real estate ads for January 12, 1999: 2400" REALLY! HI CEILS, 18" KIT, MBR/Riv vu, mds, clsts galore! $915K [5][6][7].

Here we find CEILS (ceilings), KIT (kitchen), MBR (master bedroom), Riv vu (river view), mds (maids (room) (?)) and clsts (closets), none of which are standard abbreviations, at least not in general written English.

Taxonomy of nonstandard word is developed to cover the different types of non-standard words that we observe. Different categories were chosen for expanding tokens to a sequence of words. A „token" is a sequence of characters separated by white space. Taxonomy of non-standard words used in the text normalization models is in the Table 2.1. In this, four different categories are defined for tokens that included alphabetic characters: expand to full word or word sequence (EXPN), say as a letter sequence (LSEQ), say as a standard word (ASWD) and misspelling (MSPL). The ASWD category includes both standard words that are simply out of the vocabulary of the dictionary used for NSW detection and acronyms that are said as a word rather than a letter sequence (e.g. NATO). The EXPN category is used for expanding abbreviations such as fplc or fireplace, but not used for expansions of acronyms or abbreviations to their full name. For example, IBM is typically labeled as LSEQ, while NY is labeled as EXPN (New York). Similarly, gov"t is labeled as an expansion. Several categories are defined for tokens involving numbers. Four main ways are identified to read numbers: as a cardinal (e.g. quantities), an ordinal (e.g. dates), a string of string of digits (e.g. phone numbers), or pairs of digits (e.g. years). Some categories can optionally be spoken in different ways. For example, a street address can be read as digits or pairs. Categories are defined for the most frequent types of numbers encountered.

3.4 Taxonomy of Non Standard Words (NSWs)

Table 1: Taxonomy of Non Standard Words

Alpha	EXPN	Abbreviation	Adv, N, Y, mph, gov"t
	LSEQ	Letter sequence	CIA, D.C, CDs
	ASWD	read as word	CAT, proper names
	MSPL	Misspelling	Geography
N U M B E R S	NUM	Number (cardinal)	12, 45,1/2, 0.6
	NORD	Number (ordinal)	May 7,3rd,Bill Gates III
	NTEL	Telephone (or part of)	212 555-4523
	NDIG	Number as digits	Room 101
	NIDE	Identifier	747, 386, 15, 3A
	NADDR	Number as street address	5000 Pennsylvania, 4523 Forbes

	NZIP	Zip code or PO Box	91020
	NTIME	a (compound) time	3:20
	NDATE	a (compound) date	2/2/99, 14/03/87 (or US) 03/14/87
	NYEAR	year(s)	1998, 80s, 1900s, 2003
	MONEY	money (US or other)	$3.45, HK$300, Y20,000, $200K
	BMONEY	money tr/m/billions	$3.45 billion
	PRCT	Percentage	75%, 3.4%
M I S C	SPLT	mixed or "split"	WS99,x220,2-car
	SLNT	not spoken word boundary	word boundary or emphasis character: M.bath, KENT*RLTY
	PUNC	not spoken, phrase boundary	Non-standard punctuation:"***"in $99.9K***Whites, "…" in DECIDE…Year
	FNSP	funny spelling	Sllooooow:sh*t
	URL	url, pathname or email	http://apj.co.uk,/usr/local,phj@tpt.com
	NONE	should be ignored	ascii art, formatting junk

For Roman numerals, to have a special category is not chosen but instead to label them according to how they are read i.e. as a cardinal (NUM, as in World War II) or an ordinal (NORD, as in Louis XIV or Louis the XIV). In the expansion of numbers in to a word sequence, the complicated case is money, where $2 billion is spoken as two billion dollars. So, the dollars moves beyond the next token. Allowing words to move across token boundaries complicates the architecture and is only necessary for this special case. So, a special tag is defined to handle these cases (BMONEY). Sometimes a token must be split to identify the pronunciation of its subparts. For example, WinNT consists of an abbreviation Win for Windows and the part NT to be pronounced as a letter sequence. To handle such cases, the SPLT tag at the token level is introduced [16-19].

4. Linguistic Analysis Module

Linguistics is the study of language. Linguistic Analysis is nothing but phonetic analysis. The main function of this module is to convert the sequence of words in to sequence of phonemes. Each and every word in the text is represented by its equivalent phonetic representation. Phone is the smallest sound element which can be segmented. It represents the typical kind of sound. Phonetics plays a major role in the generation of synthesized speech. Phonetics can be used to differentiate homographs. Homographs are the words that are spelled same but pronounced differently. For example, like present and past tense form of verb „to read". If we know the name and how to pronounce the word Califano in our dictionary, we can make the pronunciation of Balifano. Because both the names share the suffix „alifano". The pronunciation of Balifano can be done by removing the phoneme /k/ corresponding to the letter C in Califano and replacing it with the phoneme /b/. The grapheme to phoneme module finds the correct pronunciation for the input words. The simplest approach for grapheme to phoneme conversion is to use a dictionary based approach. In this method, a large dictionary contains all the words and their pronunciations. Determining the correct pronunciation of each input word is simple since it only involves looking up each word in the dictionary and replacing it with the pronunciation defined in the dictionary [8]. Another approach is a rule-based method where rules for the pronunciation of the words are applied to words to find out their pronunciations based on their spelling. Both of these two approaches have their advantages and

disadvantages. The dictionary based method is very simple and accurate. However, it fails completely if the word is not found in the lookup dictionary. Also, as the dictionary size grows, the memory requirements of the system also become more demanding. The rule based approach is able to work with any input but the rules easily become very complex. The manual creation of the rules is also a very time consuming and language dependent process [16-19].

4.1 Prosodic Analysis Module

Prosody is the rhythm, stress and intonation of speech. Intonation is variation of spoken pitch and it indicates the attitudes and emotions of speaker. Prosody reflects various features of the speaker or the utterance. That means the emotional state of the speaker and the form of utterance that might be statement, question or command is reflected. Intonation on a particular could differentiate between sentence moods. As an example, two sentences are given below which shows the difference in utterance of an interrogative and imperative sentence respectively [9][10].

you are finISHED (interrogative sentence)
you are FINIshed (imperative sentence)

The input given to the prosodic analysis module is sequence of phonemes. The output of the prosodic analysis module is sequence of phonemes with pitch and duration. The main function of this prosodic analysis module is to get the output of synthesizer just like human conversation.

5. Architecture for Prosody Generator

The various modules of prosody generator are described in detail given below:

1. Speaking Style

Prosody depends not only on the linguistic content of a sentence. Different people generate different prosody depending on his or her mood.
The various parameters which influence the speaking style are given belw:
a**. Character:** Character refers to extra linguistic properties of a speaker such as membership in a group and individual personality. It also includes socio syncratic features such as a speaker‟s region and economic status. In addition, idiosyncratic features such as gender, age, speech defects etc. affect speech and physical status is also a determiner of prosodic character.
b. **Emotion:** Temporary emotional conditions such as amusement, anger, contempt, grief, sympathy, suspicion, etc. have an effect on prosody. TTS systems need to provide information on the simulated speaker's state of mind. These are relatively unstable properties, somewhat

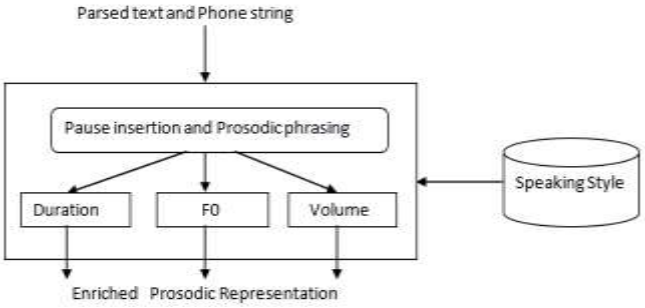

Figure 2: Architecture of Prosody Generator

Some basic emotions that have been studied in speech include:

i. Anger: Generally correlated with wide and raised pitch range

ii. Joy: This is related with increase in pitch and pitch range, with increase in speech rate. Smiling generally raises F0 and formant frequencies

iii. Sadness: Generally has normal or lower than normal pitch realized in a narrow range, with a slow rate and tempo. It may also be characterized by slurred pronunciation and irregular rhythm.

iv. Fear: This is characterized by high pitch in a wide range, variable rate, precise pronunciation, and irregular voicing perhaps due to disturbed respiratory pattern.

6. Symbolic Prosody

Symbolic prosody deal with two factors: breaking the sentence into prosodic phrases which are separated by pauses and assigning labels to different syllables or words within each prosodic phrase. The term 'juncture' refers to prosodic phrasing: where words cohere and where prosodic breaks (pauses or special pitch movements) occur.
The primary phonetic means of signaling juncture are:

- Silence insertion
- Characteristic pitch movements in the phrase final syllable.
- Increasing the duration of a few phones in the phrase final syllable.

- Irregular voice quality such as vocal fry

The block diagram of the pitch generator decomposed in symbolic and phonetic prosody is shown in the Figure 3 the various components are described in the following section:

i) Pause

The main aim to insert pause in running text is to structure the information which is generated in the form of voice output. In typical systems, the reliable location which indicates the insertion of pause is pronunciation symbols. In predicting pauses it is necessary to consider their occurrence and their duration, the simple presence or absence of a silence (of greater than 30 ms) is the most significant decision, and its exact duration is secondary, based partially on the current rate setting and other extraneous factors. The goal of a TTS system should be to avoid placing pauses anywhere that might lead to ambiguity, misinterpretation, or complete breakdown of understanding. Fortunately, most decent writing apart from email incorporates punctuation according to exactly this metric: no need to punctuate after every word, just where it aids interpretation.

ii) Prosodic Phrases

Based on punctuation symbols present in the text, commercial TTS systems are using the simple rules to vary the pitch of text depending on the prosodic phrases. For example if comma symbol appears in the text, the next word will be in the slightly higher pitch than the current pitch. The tone of particular utterance is set by using standard indices called as ToBI (Tone and Break Indices). These are standard for transcribing symbolic intonation of American English utterances, and can be adapted to other languages as well. The Break Indices part of ToBI specifies an inventory of numbers expressing the strength of a prosodic juncture. The Break Indices are marked for any utterance on their own discrete break index tier or layer of information, with the BI notations aligned in time with a representation of the speech phonetics and pitch track [23]. On the break index tier, the prosodic association of words in an utterance is shown by labeling the end of each word for the subjective strength of its association with the next word, on a scale from 0 to 4.

Four types of phrase breaks were labeled: 0 for no break, 1 for minor break, 2 for major break and 3 for punctuation mark break. No break means that no pause is inserted at word boundaries. The minor break is a very short pause between two words and the major break is a longer pause caused by respiratory effects. Punctuation mark break is labeled at the position of punctuation mark.

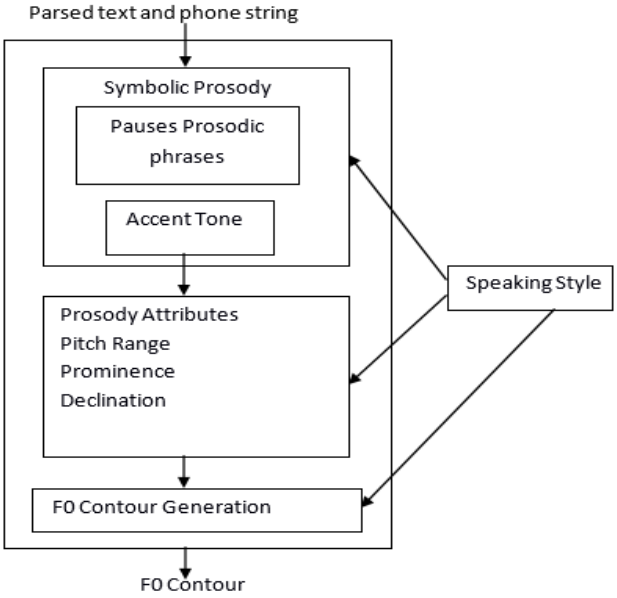

Figure 3: Pitch generator decomposed into symbolic and phonetic prosody

6.1 Duration Assignment

There are various factors which influence the phoneme durations. The common factors are:

a. Semantic and pragmatic conditions.
b. Speech rate relative to speaker intent, mood and emotion.
c. The use of duration or rhythm to possibly signal document structure above the level of phrase or sentence.
d. The lack of a consistent and coherent practical definition of the phone such that boundaries can be clearly located for measurement.

One of the commonly used methods for duration assignment is called as Rule based method. This method uses table lookup for minimum and inherent duration for every phone type. The duration is rate dependent, so all phones can be globally scaled in their minimum duration for faster or slower rates. The inherent duration is raw material and using the specified rules, it may be stretched or contracted by pre-specified percentage attached to each rule type as specified and then it is finally added back to the minimum duration to yield a millisecond time for a given phone.

6.3 Pitch Generation

Since generating pitch contours is an incredibly complicated problem, pitch generation is often divided into two levels, the first level called as symbolic prosody and the second level, generating

pitch contours from this symbolic prosody. It is useful to add several other attributes of the pitch contour prior to its generation, which is discussed below:

i) Pitch Range

Pitch range refers to the high and low limits within which all the accent and boundary tones must be realized: a floor and ceiling, so to speak, which are typically specified in Hz. This may be considered in terms of stable, speaker specific limits as well as in terms of an utterance or passage.

ii) Gradient Prominence

Gradient prominence refers to the relative strength of a given accent position with respect to its neighbors and the current pitch range setting. The simplest approach, where every accented syllable is realized as a high tone, at uniform strength, within an invariant range, can sound unnatural.

iii) Phonetic F0: Micro prosody

Micro prosody refers to those aspects of the pitch contour that are unambiguously phonetic and that often involve some interaction with the speech carrier phones.

6.3 Waveform Synthesizer

Waveform synthesizer is the final part of speech synthesis system. It generates synthetic speech as output by receiving phone information, prosody from the previous block i.e., prosodic analysis module and existed voice models. This can be done either based on a parametric representation in which phoneme realizations are produced by machine or by selecting speech units from a database.

In the latter method, a sophisticated search process is performed in order to find the appropriate phoneme, di-phone, tri-phone, or other unit at each time. Whichever method is chosen, the resulting short units of speech are joined together to produce the final speech signal. One of the biggest challenges in the synthesis stage is actually to make sure that the units connect to each other in a continuous way so that the amount of audible distortion is minimized.

7 Speech Synthesis Methods

There are different kinds of synthesis methods that can be used when building a TTS synthesis system. Some of these methods require a set of rules to drive the synthesizer whereas others depend on parameters exercised from the recorded speech database [11]. These classifications are called rule driven and data driven or corpus based synthesis respectively [12]. Examples of rule driven synthesis include articulatory synthesis and formant synthesis. On the other hand, examples of data driven synthesis include concatenative synthesis and HMM based synthesis [13].

7.1 Articulatory Synthesis

Articulatory synthesis is a technique for synthesizing speech based on human speech production model directly. It is motivated by how the human articulators such as vocal tract, nasal tract, lungs and larynx generate speech. The synthetic speech produced by this model is most natural but it is the most difficult method and computationally very expensive. It is based on the articulatory theory of speech production. In an articulatory model, the tube corresponding to the vocal tract is usually divided into many small sections, and each section is approximated by an electrical transmission line. Typical Articulatory synthesis system uses seven to eleven parameters to adequately describe motion of various articulators. The synthesizer requires one parameter for controlling velum opening, one for lip rounding, one for lip closure, two each for the tongue body and tongue tip as they have both vertical and horizontal degrees of freedom, one each for jaw height, pharynx width, and larynx height.

7.2 Formant Synthesis

This is the oldest method for speech synthesis, and it dominated the synthesis implementations for a long time. Formant synthesis is based on the well-known source filter model which means that the idea is to generate periodic and non-periodic source signals and to feed them through a resonator circuit or a filter that models the vocal tract. The principles are thus very simple, which makes formant synthesis flexible and relatively easy to implement. Formant synthesis can be used to produce any sounds. On the other hand, the simplifications made in the modeling of the source signal and vocal tract inevitably lead to somewhat unnatural sounding result.

7.3 Concatenative Synthesis

Unit selection synthesis and di-phone synthesis are the two well-known Concatenative synthesis strategies. Unit selection synthesis is the so called cut and paste synthesis in which short segments of speech are selected from a prerecorded database and concatenated one after another to produce the desired utterances. The short segments of speech may range from phones to phrases. The longer the selected units are, the fewer problematic concatenation points will occur in the synthetic speech. The greatest drawback is that it requires very

large speech database for training the system which is very hard to collect. Another limitation in this synthesis technique is the strong dependency of the output speech on the chosen database. Phonemes and diphones are most commonly chosen selected units because they are short enough to attain sufficient flexibility and to keep the memory requirements reasonable. Using longer units, such as syllables or words, is impossible or impractical The use of di-phones in the concatenation provides rather good possibilities to take account of co-articulation because a di-phone contains the transition from one phoneme to another and the latter half of the first phoneme and the former half of the latter phoneme. Consequently, the concatenation points will be located at the center of each phoneme, and since this is usually the steadiest part of the phoneme, the amount of distortion at the boundaries can be expected to be minimized. While the sufficient number of different phonemes in a database is typically around 40–50, the corresponding number of di-phones is from 1500 to 2000 but a synthesizer with a database of this size is generally implementable. In both phoneme and di-phone concatenation, the greatest challenge is the continuity. To avoid audible distortions caused by the differences between successive segments at least the fundamental frequency and the intensity of the segments must be controllable. The creation of natural prosody in synthetic speech is impossible with the present day methods but some promising methods for getting rid of the discontinuities have naturally been developed. Compared to unit selection synthesis technique, di-phone synthesis uses a minimal speech database containing all the di-phones occurring in a given language. In di-phone synthesis, only one example of each di-phone is contained in the speech database. At runtime, the target prosody of a sentence is superimposed on these minimal units by means of digital signal processing techniques such as linear predictive coding (LPC). Di-phone synthesis usually suffers from the sonic glitches at concatenation points and the quality of the resulting speech is generally not as good as that from unit selection but more natural sounding than the output of formant synthesis [12].

8. HMM Based Speech Synthesis

HMM based speech synthesis is also an acoustic model based synthesis method employing HMMs. It is the emerging technology for TTS. HMM based synthesis is also called Statistical Parametric Synthesis. In this system, the frequency spectrum (vocal tract), fundamental frequency (vocal source) and duration (prosody) of speech are commonly modeled simultaneously by HMMs. Speech waveforms are generated from HMMs themselves based on maximum likelihood criterion [1]. Although many speech synthesis systems can synthesize high quality speech, they still cannot synthesize speech with various voice characteristics such as speaker individualities, speaking styles, emotions, etc. To obtain various voice characteristics in speech synthesis systems based on the selection and concatenation of acoustical units, a large amount of speech data is necessary. However, it is difficult to collect and store such huge amount of speech database for different languages. Moreover, storing big database in devices having only a small amount of memory is not possible. So, in order to construct speech synthesis systems which can generate various voice characteristics, the HMM based speech synthesis system (HTS) was proposed. HMM based speech synthesis system mainly consists of two parts. One is the training part and the other is synthesis part. In the training part, spectrum and excitation parameters are extracted from speech database and modeled by context dependent HMMs. In the synthesis part, context dependent HMMs are concatenated according to the text to be synthesized. Then spectrum and excitation parameters are generated from the HMM. Finally, the excitation generation module and synthesis filter module synthesize speech waveform using the generated excitation and spectrum parameters. The attraction of this approach is that voice characteristics of synthesized speech can easily be changed by transforming HMM parameters. Similarly to other data driven speech synthesis approaches, HTS has a compact language dependent module i.e. a list of contextual factors, which can be extracted through Festival and they are called “features” in Festival framework. Thus, HTS could easily be extended to other language [14].

9. Conclusion

Speech synthesis system overview and its performance are described in this chapter. The way in which the text is synthesized i.e. how the system converts the text input into speech is explained with the help of block diagram in this chapter. Each and every module with its input and output are described. Further, different speech synthesis methods are discussed in detail. The HMM based speech synthesis technique, its significance and

different phases involved in the development of system are explained.

REFERENCES

[1]. Nisako Baloyi, "A Text-to-Speech synthesis System for Xitsonga Using Hidden Markov Models", M.Tech Mini Dissertation at University of Limpopo, June 2012, pp.1-61.

[2] Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "A Marathi Hidden-Markov Model Based Speech Synthesis System" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 34-39e-ISSN: 2319 –4200, p-ISSN No. : 2319 – 4197

3) Sangramsing Kayte, Monica Mundada,Dr. Charansing Kayte" Speech Synthesis System for Marathi Accent using FESTVOX" International Journal of Computer Applications (0975 – 8887) Volume 130 – No.6, November2015

4)Sangramsing Kayte, Monica Mundada, Dr. Charansing Kayte "Marathi Text-To-Speech Synthesis using Natural Language Processing "IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 63-67e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197

5) Sangramsing Kayte, Monica Mundada,Dr. Charansing Kayte "Screen Readers for Linux and Windows – Concatenation Methods and Unit Selection based Marathi Text to Speech System" International Journal of Computer Applications (0975 – 8887) Volume 130 – No.14, November 2015

6) Sangramsing Kayte, Monica Mundada,Dr. Charansing Kayte " Performance Evaluation of Speech Synthesis Techniques for Marathi Language " International Journal of Computer Applications (0975 – 8887) Volume 130 – No.3, November 2015

7) Sangramsing Kayte, Monica Mundada, Jayesh Gujrathi, " Hidden Markov Model based Speech Synthesis: A Review" International Journal of Computer Applications (0975 – 8887) Volume 130 – No.3, November 2015

[8] Jonathan Allen," Linguistic Aspects of Speech Synthesis", Volume 92, pp.9946-9952, October 1995.

9) Sangramsing N. Kayte,Monica Mundada,Dr. Charansing N. Kayte, Dr.Bharti Gawali "Rule-based Prosody Calculation for Marathi Text-to-Speech Synthesis" Sangramsing N. Kayte et al. Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part - 5) November 2015, pp.33-36

10) M.B.Chandak, Dr.R.V.Dharaskar and Dr.V.M.Thakre,"Text to Speech with Prosody Feature: Implementation of Emotion in Speech Output using Forward Parsing", International Journal of Computer science and Security, Volume (4), Issue (3).

11) Newton, "Review of methods of Speech Synthesis", M.Tech Credit Seminar Report, Electronic Systems Group, November, 2011, pp. 1-15

12) 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

13) 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

14) Sangramsing Kayte, Monica Mundada, Jayesh Gujrathi, " Hidden Markov Model based Speech Synthesis: A Review" International Journal of Computer Applications (0975 – 8887) Volume 130 – No.3, November 2015

15) Black, A., Zen, H., Tokuda, K, "Statistical Parametric Synthesis", in Proc. ICASSP, Honululu, USA, 2007

16) 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.

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

18) 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

19) 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).