Speech Synthesis: A Review

Archana Balyan[1], S. S. Agrawal[2], Amita Dev[3]

[1] Department of Electronics and Communication Engineering, MSIT, New Delhi, India
[2] Advisor C DAC & Director KIIT, Gurgaon, India
[3] Bhai Parmanand Institute of Business Studies, Delhi, India

Abstract

Attempts to control the quality of voice of synthesized speech have existed for more than a decade now. Several prototypes and fully operating systems have been built based on different synthesis technique. This article reviews recent research advances in R&D of speech synthesis with focus on one of the key approaches i.e. statistical parametric approach to speech synthesis based on HMM, so as to provide a technological perspective. In this approach, spectrum, excitation, and duration of speech are simultaneously modeled by context –dependent HMMs, and speech waveforms are generated from the HMMs themselves. This paper aims to give an overview of what has been done in this field, summarize and compare the characteristics of various synthesis techniques used. It is expected that this study shall be a contribution in the field of speech synthesis and enable identification of research topic and applications which are at the forefront of this exciting and challenging field.

Key words: *Text-to- speech, concatenative synthesis, Database, Hidden markov model, feature extraction*

1. Introduction

Speech synthesis is a process of automatic generation of speech by machines/computers. The goal of speech synthesis is to develop a machine having an intelligible, natural sounding voice for conveying information to a user in a desired accent, language, and voice. Research in T-T-S is a multi-disciplinary field: from acoustic phonetics (speech production and perception) over morphology (pronunciation) and syntax (parts of speech, grammar), to speech signal processing (synthesis). There are several processing stages in T-T-S system: the text front –end analyses and normalizes the incoming text, creates possible pronunciations for each word in context, and generates prosody (emotions, melody, rhythm, intonation) of the sentence to be spoken. For evaluation of T-T-S systems three parameters need to be evaluated: accuracy, intelligibility and naturalness. The fig. 1 shows a block diagram of T-T-S synthesis (X.Huang, 2001) [1].

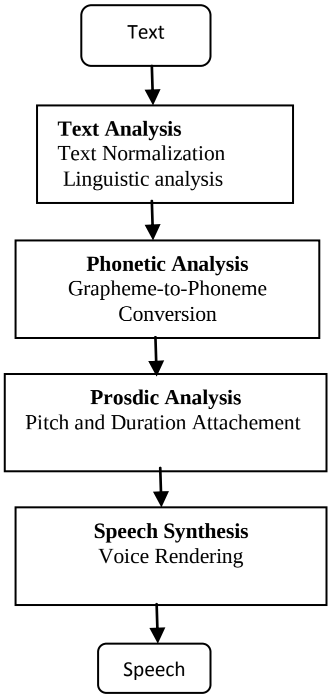

Fig. 1: Block diagram of TTS

Implementation of T-T-S

The process of transforming text into speech contains broadly two phases: 1) Text analysis and 2) generation of speech signal.

Text analysis consists of normalization of the text wherein the numbers and symbols become words and abbreviations are replaced by their whole words or phrases etc. The most challenging task in the text analysis block is the linguistic analysis which means syntactic and semantic analysis and aims at understanding the context of the text. The statistical methods are used to find the most probable meaning of the utterances. This is significant because the pronunciation of a word may depend on its meaning and on the context.

Phonetic Analysis converts the orthographical symbols into phonological ones using a phonetic alphabet. For e.g. the alphabet of the *International Phonetic Association* contains phoneme symbols, their diacritical marks and other symbols related to their pronunciation, other phonetic alphabets such as SAMPA (*Speech Assesment Methods-Phonetic Alphabet), Worldbet and Arpabet* are available.

Prosody is a concept that contains the rhythm of speech, stress patterns and intonation. At the perceptual level, naturalness in speech is attributed to certain properties of the speech signal related to audible changes in pitch, loudness and syllabic length, collectively called prosody. Acoustically, these changes correspond to the variations in the fundamental frequency (F0), amplitude and duration of speech units (T. Dutoit, 1997 & D. Jurafsky, 2000) [2, 3].

Speech Synthesis block finally generates the speech signal. This can be achieved either based on parametric representation, in which phoneme realizations are produced by machine, or by selecting speech units from a database. The resulting short units of speech are joined together to produce the final speech signal.

T-T-S systems have numerous potential applications. Few are listed below.

1. **In telecommunication service:** Most of the calls required very less connectivity, T-T-S systems are show huge presence in telecommunication services by making it possible to access textual information over the phone.
2. **In e-governance service**: T-T-S can be very helpful by providing government policy information over the phone, polling centre information, land records information, application tracking and monitoring etc.
3. **Aid to disabilities**: T-T-S can give invaluable support to voice handicapped individuals with the help of an especially design keyboards and fast sentence assembling program, also helpful for visually handicapped.
4. **Voice browsing**: T-T-S is the backbone of voice browsers, which can be controlled by voice instead of by mouse and keyboard, thus allowing hands-free and eyes free browsing.
5. **Vocal monitoring**: At times oral information is supposed to be more efficient than its written counterpart. Hence, the idea of incorporating speech synthesizers in the measurement or control systems, like cockpits to prevent pilots from being overwhelmed with visual information.
6. **Complex interactive voice response systems:** With the support of good quality speech recognizers, speech synthesis systems are able to make complex interactive voice response systems a reality.
7. **Multimedia, man-machine communication**: Multimedia is first but promising move in the direction and it includes talking books and toys, mail and document readers. However, as the applications spread, the issue of naturalness is of prime importance in the development of unlimited text to speech synthesizers.

Over the last decade, TTS technologies have shown a convergence towards statistical parametric approaches (H.Zen, K.Tokuda 1989) [4].The most extensively investigated generative model has been the hidden Markov model (HMM) that was first proposed for the use in ASR (L.R. Rabiner, 1989) [5] and in more recent years the HMM has also become the focus of increasing interest in TTS research (A.Falaschi, 1989) [6]. In this paper we restrict the scope of our study to the dominant paradigm in speech modeling for T-T-S- The hidden Markov model. In this paper, we will review some of the approaches used to generate synthetic speech and discuss some of the basic factors for choosing one method over another. This paper is organized as follows: Section 2 gives overview of various existing synthesis approaches and techniques with underlying assumptions. Section 3 presents an overview of HMM based speech synthesis .Section 4 description of implementation of statistical models for TTS is presented also discussing their advantages and disadvantages. Section 5 gives the details of the various major databases that are available for development of T-T-S and discusses speech and database development in Indian scenario. In section 6, we conclude the study and give suggestions for future work in this field of research.

2. Recent techniques of speech synthesis

The techniques which have been developed in the recent past could be divided into three categories: (i)Articulatory synthesis,(ii) formant synthesis and (iii) concatenative synthesis. These have been classified on the basis of how they parameterize the speech for storage and synthesize.

2.1 *Articulatory synthesis*

Articulatory synthesis is based on physical models of the human speech production system. It involves simulating the acoustic functions of the vocal tract and its dynamic motion. An *articulatory model*; reconstitutes the shape of the vocal tract as a function of the position of the phonatory organs (lips, jaw, tongue, velum). The signal is calculated by a mathematical simulation of the air flow through the vocal tract. The control parameters of such a synthesizer are: sub-glottal pressure, vocal cord tension, and the relative position of the different articulatory organs. An articulatory model is then reproduced which corresponds to the shape of the vocal tract. The problems faced in this technique are that of obtaining accurate three-dimensional vocal tract representations and of modeling the system with a limited set of parameters. S. Martincic- Ipsic, 1989 [7] cites lack of knowledge of the complex human articulation organs being the main reasons why articulatory synthesis has not lead to quality speech synthesis. In the publications by Fant (1960), Holmes, Mattingly, and Shearme (1964), Flanagan (1972), Klatt (1976), Allen, Hunnicutt, and Klatt (1987) the foundations for speech synthesis based on acoustical or articulatory modeling can be found.

2.2 *Formant speech synthesis*

Formant speech synthesis is based on rules which describe the resonant frequencies of the vocal tract. The formant method uses the source-filter model of speech production, 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. Rule-based formant synthesis can produce quality speech which sounds unnatural, since it is difficult to estimate the vocal tract model and source parameters. Typically the adjustable parameters include at least the fundamental frequency, the relative intensities of the voiced and unvoiced source signals, and the degree of voicing. The parameters controlling the frequency response of the vocal tract filter – and those controlling the source signal – are updated at each phoneme. The vocal tract model can be implemented by connecting the resonators either in cascade or parallel form.

An important step in synthesizing good quality speech was development of terminal analogue or formant synthesizers-both serial and parallel type. Several versions of formant synthesizers such as PAT, OVE-II, and INFOVOX were developed. The demonstration of parallel formant synthesizers by John Holms made a remarkable impact for English speech. Klatt has used combined version of serial and parallel formant synthesizer, which formed the basis of the MITalk and KLattalk models of the synthesizer. A set of source and tract parameters were used to control the synthesizer to dramatically vary the output waveform by changing them in accordance with the knowledge/data obtained from the analysis of original speech. Agrawal S.S., 2001[8] reports that KLSYN88 and KLSYN93 version has been used for synthesizing Hindi speech. At CEERI, PC version of the Klatt T-T-S model of cascade/parallel formant synthesizer was developed. The vowels and voiced sounds, semi-vowels and aspirated sounds were generated by using serial tract while the fricative sounds and the burst of the stop consonants were generated by parallel track. The synthesizer was controlled by a set of about 60 parameters (consonants and variables). A set of parameters which have been varied more frequently are shown in Table 1 and Table 2.

Table 1: Source Parameters varied

Parameter	Type	Min	Max	Def	Description
F0	V	0	1000	5000	Fundamental frequency, in tenths of Hz
AV	V	0	60	80	Amplitude of voicing, in dB
OQ	V	10	50	99	Open quotient(voice opening time /period),in %
SQ	V	100	200	500	Speed quotient(rise/fall time, LF model), in %
TL	V	0	0	41	Extra tilt of voicing spectrum, dB down @3KHz
AH	V	0	0	80	Amplitude in aspiration, in dB
AF	V	0	0	80	Amplitude of frication , in dB

Table 2: Vocal Tract Parameter Varied

Parameter	Type	Min	Max	Def.	Description
F1	V	180	500	1300	Frequency of 1st formant, in Hz
B1	V	30	60	1000	Bandwidth of 1st formant, in Hz
F2	V	550	1500	3000	Frequency of 1st formant, in Hz
B2	V	40	90	1000	Bandwidth of 1st formant, in Hz
F3	V	1200	2500	4800	Frequency of 1st formant, in Hz
B3	V	60	150	1000	Bandwidth of 3rd Formant in Hz
F4	V	2400	3250	4990	Frequency of 3rd Formant in Hz
B4	V	100	200	1000	Bandwidth of 4rth Formant in Hz
F5	V	3000	3700	1500	Frequency of 3rd Formant in Hz
B5	V	100	200	4990	Bandwidth of 3rd Formant in Hz
F5	V	100	4990	1500	Frequency of 1st formant, in Hz
B6	V	0	500	4990	Bandwidth of 1st formant, in Hz
A2F	V	0	0	4000	Amp of ric-excited parallel 2nd formant, in Hz
A3F	V	0	0	80	Amp of fric-excited parallel 2nd formant,in Hz
A4F	V	0	0	80	Amp of fric-excited parallel 2nd formant, in Hz
A5F	V	0	0	80	Amp of fric-excited parallel 2nd formant, in Hz
A6F	V	40	250	80	Amp of fric-excited parallel 2nd formant, in Hz
B2F	V	60	300	1000	BW of fric-excited parallel 2nd formant, in Hz
B3F	V	100	320	1000	BW of fric--excited parallel 2nd formant, in Hz
B4F	V	100	360	1000	BW of fric-excited parallel 2nd formant,in HZ
B6F	V	100	1500	1500	BW of fric-excited parallel 2nd formant,in Hz
A2F	V	0	0	4000	Amp of fric-excited parallel 2nd formant, in Hz
FNP	V	180	280	80	Frequency of nasal pole, in Hz.
BNP	V	40	90	500	Bandwidth of nasal pole, in Hz
FNZ	V	180	280	1000	Freq uency of nasal zero, in Hz
BNZ	V	40	90	800	Bandwidth of nasal zero, in Hz
FTP	V	300	2150	1000	Frequency of nasal pole, in Hz
BTP	V	40	180	3000	Bandwidth of tracheal pole, in Hz
FTZ	V	300	2150	3000	Frequency of tracheal zero, in Hz

Due to high degree of control that the formant synthesizers provide, it has been widely used. These include Janet Cahn's Affect Editor [9] [10],[11],[12],[13] and Iain Murray et al.'s HAMLET,1989 [11] [12].The common feature is that both have used DECtalk as a formant synthesis system, providing dedicated processing modules which adapt their input according to the acoustic properties of the number of emotions. In both cases, the acoustic profile for each emotion category was derived from the literature and manually adapted. However, the Affect editor requires the input to be manually annotated; HAMLET processes its input entirely by rule. Burkhardt,2000 [14] [15] has used systematic, perception- oriented approach to find good acoustic relates for German speech. In addition to the resonators that model the formants, the synthesizer can contain filters that model the shape of the glottal waveform and the lip radiation, and also an *anti-resonator* to better model the nasalized sounds.

2.3 *Concatenative Speech synthesis*

More natural speech can be produced using concatenation techniques. In these techniques, stored speech units (segments) that are tied together to form a complete speech chain of sub-word units (e.g. phonemes, diphones) and has become basic technology. However, differences between natural variations of speech and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches in the output. There are two main sub-types of concatenative synthesis: 1) Diphone concatenation synthesis and 2) corpus based speech synthesis.

2.3.1 Diphone concatenation synthesis

Attempts to build utterances from *phoneme* wave forms have been of limited success, due to coarticulation problems. The use of larger concatenative units, particularly diphones (i.e. excised wave forms from the middle of one phoneme to the middle of the next one) provides rather good possibilities to take account of coarticulation because a diphone contains the transition from one phoneme to another and latter half of the first phoneme and the former half of the first phoneme. Consequently, the concatenation points will be located at the center of each phoneme, and since this is usually the most steady part of the phoneme, the amount of distortion at the boundaries are expected to be the minimum and must be subjected to a minimum of smoothing. While the sufficient number of different phones in a database is typically around 40-50, the corresponding number of diphones is from 1500 to 2000 and a synthesizer with a database of this size is implementable (S.Lemmetty) [16]. However, while diphone concatenation can produce a reasonable quality speech, a single example of each diphone is not enough to produce good quality speech.

2.3.2. Use of Diphone synthesis for emotional speech synthesis:

Diphone recordings are usually carried out with a monotonous pitch. At synthesis time, the required F0 contour is generated through various signal processing techniques which introduces certain amount of distortion, but with a resulting speech quality much more natural than formant synthesis. Various studies have been conducted to study whether F0 and duration are sufficient to express emotions. While (] Heuft. B 1996, Vroomen 1993, Montreo J.M. 1999, Iriondo 2000, Edgington 1991, Iriondo 2000, Schröder, M., 1998) [17] [19] [21] [22] [24] [25] report that synthesized emotions can be recognized reasonably well, (Edgington, 1991, Rank, E.1998) [18] [20] report recognition rates close to chance level. One approach to emotional speech synthesis with diphones, used by Murray,I.R., 2000) [23] is copy synthesis. Mozziconacci,S.J.L., 1998 and Chung, S.-J., 1999 [26] [27] formulated prosody rules for emotions. However, among the problems with diphone synthesis remains the danger of major discontinuities occurring at the interface between two halves of a vowel, in cases where dissimilar formant targets are used on the two sides of the interface.

2.3.3. Diphone concatenation using Linear Prediction Coefficients

Synthesis systems based on coding have as long a history as the vocoder. Stevens, 1960 [28] proposed a conceptual model designed to improve speech recognition by using speech synthesis technique. Then the idea known as “analysis by synthesis” (AbS) was applied to various models using linear predictive coding (LPC) since LPC corresponds to the vocal tract filter. The underlying principle is that natural human speech is transformed into parameter sequences and stored in such a way that it can be assembled into new utterances. Synthesizers such as the systems from AT&T, (Olive, 1977, 1990, and Olive and Liberman, 1985), [29][30][31] ,NTT (Hakoda et. al., 1990 and Nakajima and Hamada, 1988) [32][33]and ATR (Sagisaka 1988, Sagisaka, Kaiki, Iwahashi, and Mimura, 1992) [34][35][36]are based on the source-filter technique where the filter is represented in terms of LPC or equivalent parameters. The development of the linear predictive coding (LPC) technique for speech analysis and re-synthesis has made it possible to store relatively large inventories of high quality speech wave forms in limited space. (Atal and Hanauer, 1971)[36].The system is an all-pole linear filter that simulates the source spectrum and the vocal tract transfer function. The technique has many advantages, such as the automatic analysis of the original signal, fairly easy algorithmic integration, and fidelity to the original sound. This filter is excited by a source model that must be able to handle all types of sounds: voiced, aspirative and fricative. It has, however, been found that the use of LPC is not successful in text-to-speech probably because of its limited ability to represent speech parameters [37].

Even though diphone synthesizers produce a reasonable quality speech waveform, in many cases the pitch and duration of the speech units from database need to be modified to the pitch and duration required for proper sounding synthetic speech. Considerable success has been achieved by systems that base sound generation on concatenation of natural speech units (Mouline et.al., 1990) [38]. The most important aspects of prosody can be imposed on synthetic speech without considerable loss of quality. The introduction of *PSOLA* (Pitch-Synchronous Overlap-Add) in 1985 considerably facilitated the research and development of concatenative synthesis systems.

The PSOLA (Carpentier and Moulines, 1989) [39] methods are based on a pitch-synchronous overlap-add approach for concatenating waveform pieces. The idea in PSOLA is to extract speech frames pitch-synchronously, i.e., the center of each frame is located at the pitch pulse position (the highest peak within a pitch period). At the synthesis stage these frames are partly overlapped and summed so that the desired time- and pitch-scale are realized. This way the prosodic features, especially with respect to duration and fundamental frequency, of speech can be adjusted independently from each other. PSOLA is best applicable to voiced speech in which the pitch period can be determined. Also, PSOLA is very sensitive to errors in the pitch estimate, which often causes problems in practice. The frequency domain approach, FD-PSOLA, is used to modify the spectral characteristics of the signal (Moulines et al. 1995) [40]; the time domain approach, TD-PSOLA, provides efficient solutions for real-time implementation of synthesis systems (Kortekaas et al. 1997) [41]. Earlier systems like SOLA (Roucos and Wilgus, 1985) [42], and systems for diver´s speech restoration also did direct processing of the waveform, (Liljencrants, 1974) [43].

2.3.4 Corpus- based speech Synthesis

Most state- of- the- art speech synthesis systems which are able to produce more natural speech are generalization of the concatenative synthesis(R.sproat,1992)[44] which is based on dynamic selection of units are based on large amounts of speech data. This method is also known as corpus synthesis. This method has become popular due to high quality synthetic voice that it provides due to utilization of natural speech as units of concatenation, improved naturalness and intelligibility it offers. The main characteristic of corpus-based T-T-S method is use of large database.

2.3.5 Preparation of database for corpus based T-T-S

The main problem with the corpus-based approaches is the need for an annotated database. These systems always require a significant amount of human effort in labeling the phonetic boundaries of the corresponding corpus [Van erp *et al.* 1988][45] [Wand et al. 1999][46].In [Ljolje *et al.*1993, 1994][47] [Demuynck *et al.* 2002][48] used HMM based recognizers. Several works have focused on automatic phonetic labeling, such as in [van Santen et al. 1990] [49] broad- band and narrow-band edge detection has been adopted. Bonafonte *et al.* [50] took Guassian probability density distribution as a similarity measure. In [Torre Toledano *et al. 1998*], [51]Toledano *et al.* tried to mimic human labeling using set of fuzzy rules using rule-based approach. In [Sethy *et al.* 2002][52], Sethy *et al.*used adapted CDHMMM (continuous density hidden Markov model) models using statistics based methods . The main focus of these studies had been English speech utterances and does not produce desirable results for another language. Several explicit segmentation approaches have been proposed in the literature. Malffere et *al.* [53] proposed an alignment of synthetic speech against natural speech, using the dynamic time warping (DTW) algorithm. Keshet *et al.* [54] introduced a phonetic alignment algorithm based on discriminative learning. In [55], Torkkola described a method for automatic alignment of speech waveforms using nueral networks followed by boundary refinement using heuristic speech-specific knowledge. In [56], Pellom and Hansen examine HMM-based segmentation performance in noisy conditions. In [57], Brugnara et al. present HMM architecture for speech segmentation. In [58], Adell et al. do a comparative study of automatic phone segmentation methods for text-to-speech. Finally, in [59] Mporas te al. introduced a hybrid HMM based method for speech segmentation, consisting of iterative isolated unit training of phone recognizers, initialized from embedded training. The hybrid HMM-based method has proved to significantly improve the speech segmentation performance in the case of TIMIT [60] multi-speaker database.

2.3.6. Unit selection synthesis

One of the major approaches in corpus-based speech synthesis is sample based one; Unit selection synthesis (A.J Hunt 1996) [61] can offer high quality synthesis without the expert work that would be required to build a formant synthesizer. Although unit selection can produce high quality synthesis the database must be appropriately designed to have the right coverage for the language or domain so that quality is reasonable.A.Black,2002[62] discusses the limitations and optimizations that can help in achieving high quality databases for unit selection. A. Black and K. Lenzo, 2001 [63]experimented with more elaborate selection technique, where they first model a particular speaker’s acoustic variation and select data based in their actual usage rather than general phonemes. The performance was good but it was more computationally expensive and required an existing model of the speaker, which may not be available when building a new language. J.Kominek and A.Black,2003[64]used a simpler technique in building the CMU ARCTIC voices, and have successfully used very similar techniques for a wide range of languages including as Croatian, Thai and Spanish. Chou, F.-C 1998 [65] noted that given a suitably balanced set of utterances we can more accurately label the data using acoustic modeling HMM tools in any

language. Out of this large database, units of variable size, e.g., HMM state, half-phone, phone, diphone, or syllable, a unit sequence corresponding to a given context-dependent sub-word sequence is selected by minimizing its total cost, consisting of target and concatenation cost (W. N. Campbell and A. Black 1998) [66] .These cost functions have been formed from a variety of heuristic or ad hoc quality measures based on features of the acoustics signals
and given texts. N. Mizutani 2002, C. Allauzen 2004, S. Sakai and H. Shu 2005, Z.-H. Ling R.2006 and Christian Weiss 2006 [67], [68], [69], [70] and [71] proposed and investigated target and concatenation cost functions based on statistical models. If perfect matching units are found in the database, the synthesis gives very good results else the results can be bad when no appropriate units are found.

The feature of unit selection synthesis to preserve the features of recorded speech very well has been exploited by Lida et al. [72] for the synthesis of emotional speech. For each of three emotions (anger, joy, and sadness), an entire unit selection database was recorded by the same speaker. In order to synthesize a given emotion, only units from the corresponding database are selected. The emotions in the resulting synthesized speech are well recognized (50-80%). Another, theoretically more demanding approach is to select the material appropriate for the targeted emotion from one database. The equivalent of prosody rules is then used as selection criteria. This has been attempted by Marumoto & Campbell [73], who used parameters related to voice quality and prosody as emotion-specific selection criteria. The results indicated a partial success: Anger and sadness were recognized with up to 60% accuracy, while joy was not recognized above chance level.

In an attempt to improve naturalness (X. Huang & A. Acero, 1997) [74], reports variety of techniques which expand the inventory of units used in the concatenation from the basic diphone schema. This could be done, both in changing the size the units, the classification of the units themselves, and the number of occurrences of each unit. According to Nagy[75], as the length of the elements used in the synthesized speech increases, the number of concatenation points decreases, resulting in higher perceived quality. In the work of Sagisaka *et al.* 1992[76], units are of variable length, giving rise to the term *non-uniform unit* synthesis. The selection algorithm use clustering based on acoustic distance but only using phonetic information. Donovan and Woodland,1995 [77] use clustering techniques based on acoustic distance, in which all the members from the cluster are used so that continuity costs may take part in the criteria for selection of the best unit. Campbell and Black,1997 [78] also use similar phonetic based clustering but further cluster the units based on prosodic features, but still resorts to weighted feature target distance for ultimate selection.   Alan Black and Paul Taylor, 1997 [79], in their work, resorts to creating a large inventory by automatically clustering units of the same phone class (uniform synthesis) based on their phonetic and prosodic context. In their algorithm, they use acoustic distance measure for clustering units, candidate units from clusters are selected by decision trees built by using CART(L. Breiman, 1996) [80] method and an optimal coupling (A. Conkie, 1997) [81] technique to measure the concatenation costs between two units. Although this method removes the need to generate the target feature weights generated in [61] [Hunt and Black, 1996] but parameters like acoustic cost and continuity cost need to be estimated.

3. Hidden Markov Model (HMM) based speech synthesis

3.1 Hidden Markov Models (HMMs)

In the early 1970s, Lenny Baum of Princeton University invented a mathematical approach to recognize speech called Hidden markov model (HMM). The Hidden markov model (HMM) (J. Ferguson 1980, L.R. Rabiner 1989, L.R. Rabiner & B.H. Juang, 1993) [83] [84] [85] is a doubly stochastic process which has an underlying stochastic process that is not observable , but can be observed through another stochastic process that produces a sequence of observations. Table 3 compares the Unit selection and HMM based speech synthesis system.

Table 3: Relation between unit selection and generation (HMM) approaches

Unit Selection	HMM
Clustering(possible use of HMM)	Clustering(use of HMM)
Multi template	Statistics
Single tree	Multiple tree(spectrum, F0, duration)
Advantage 1)High quality at waveform level Disadvantage 1) Discontinuity 2) Hit or miss	Disadvantage 1) Vocoded speech(buzzy) Advantage 1)Smooth 2)stable
Large run-time data	Small run-time data
Fixed voice	Various voices

3.1.1 Recent Development of HMM- based speech synthesis system (HTS)

HMM based speech synthesis continues to dominate other synthesis approaches due to existence of freely available open source software such as HTS (K. Tokuda & H. Zen) [86] named “HMM-based speech synthesis system” to provide a research and development platform for statistical parametric speech synthesis. The HMM-based speech synthesis system (HTS) has been developed by the HTS working group as an extension of the HMM toolkit (HTK) (S. Young, 2006) [87].The source code of HTS is released as a patch for HTK. The first version 1.0 HTS was first released in December 2002. After an interval of three years, HTS version 2.0 was released in December 2006 with major update and inclusion of number of new features, such as introduction of global mean and variance calculation tool, for large databases the previous version often suffered from numerical errors.HTS version 2.0.1 was a bug – fixed version and the latest version, HTS version 2.1, was released in July 2008.This version includes important features; Hidden semi- markov models (HSMMs)(H. Zen & K. Tokuda, 2007, J. Yamagishi, 2007)[88][89], the speech parameter generation algorithm considering global variance (GV)(T. Toda and K. Tokuda, 2007) [90], advanced adaptation techniques (J. Yamagishi, 2009) [91], and stable version of run time synthesis engine API. The HTS version 2.1, with the STRAIGHT analysis/synthesis techniques (H. Kawahara 1999) [92], provides the ability to construct the state-of-art HMM based speech synthesis systems developed for the past Blizzard Challenge events(H. Zen& T. Toda, 2007, H. Zen & T. Toda, 2006, J. Yamagishi, 2009) [93][94][95]. H. Zen, 2009 [96] describes the details of new features included in version 2.1.

3.2. Architecture of a Typical HMM based speech synthesis system

T.Yoshimura, 2000 [82] suggested a trainable approach in which speech waveform is synthesized from parameters directly generated from Hidden Markov Models (HMM) has gained popularity. One of the main advantages of the referred HMM –based synthesis techniques when compared with unit selection and concatenation method is the fact that the voice alteration can be performed without large databases, being at par with quality with unit selection and concatenation ones. Figure 2 shows the system overview[82] . 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 by using a speech parameter generation algorithm. Finally, the excitation generation module and synthesis filter module synthesize speech waveform using the generated excitation and spectrum parameters. The training part performs the maximum likelihood estimation by using the Expectation Maximization (EM) algorithm (Dempster et al., 1977) [97]. In this process, spectrum (e.g., mel-cepstral coefficients) (Fukada et al., 1992) [98] and their delta and delta-delta coefficients) and excitation (e.g., log F_0 and its dynamic features) parameters are extracted from a database of natural speech and modeled by a set of multi-stream (Young et al., 2006) [99] context-dependent HMMs (phonetic, linguistic, and prosodic contexts being taken into account).

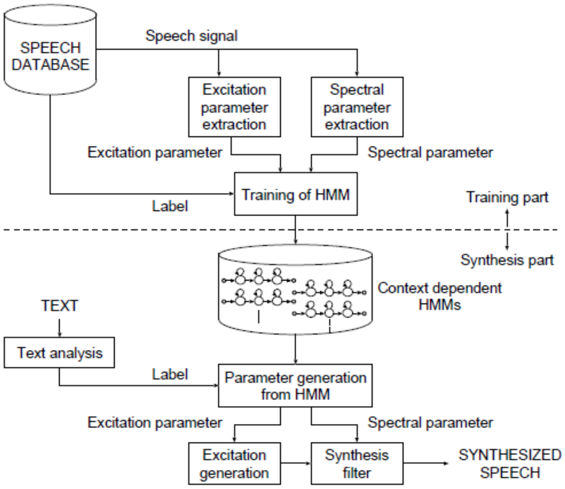

Fig 2: Typical Architecture of HMM- Based speech synthesis system

To model fixed-dimensional parameter sequences, such as mel cepstral coefficients, single multi-variate Gaussian distributions are typically used as their stream-output distributions. Several methods have been studied for modeling log F_0 sequences (Freij and Fallside, 1988[100]; Jensen et al., 1994[101]; Ross and Ostendorf, 1994, [102], the HMM-based speech synthesis system adopts multi-space probability distributions (Tokuda et al., 2002a) [103] as their stream-output distributions. To model the temporal structure of speech, each HMM has its state-duration distribution namely, the Gaussian distribution (Yoshimura et al., 1998) [104] and the Gamma distribution (Ishimatsu et al., 2001) [105]. They are estimated from statistical variables obtained at the last iteration of the forward-backward algorithm. As they have their own context dependency, each of spectrum, excitation, and duration is clustered individually by using phonetic decision trees (Odell, 1995) [106]. Hence, the system can model the spectrum, excitation, and duration in a unified framework. In the synthesis part, a given word sequence is converted into a context dependent label sequence, and then the utterance HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. Then, various kinds of speech parameter generation algorithm (Tokuda et al., 2000; Tachiwa and Furui, 1999), [107] [108] have been used to generate the spectrum and excitation parameters HMM. Finally, the excitation generation module and synthesis filter module filter, such as Mel log spectrum approximation (MLSA) filter (Imai et al., 1983) [109] synthesize speech waveform using the generated excitation and spectrum parameters.

3.2.2 Transforming voice characteristics, speaking styles, and emotions

The main advantage of statistical parametric synthesis is that it can synthesize speech with various voice characteristics such as speaker individualities, speaking styles, and emotions etc. The combination of unit-selection and voice-conversion (VC) techniques (Stylianou et al., 1998) [110] can alleviate this problem but high-quality voice conversion is still difficult. However, we can easily change voice characteristics, speaking styles, and emotions in statistical parametric synthesis by transforming its model parameters. There are three major techniques to achieve this, namely adaptation, interpolation, and eigenvoices.

(1) Speaker Adaptation (mimicking voices)
The use of adaptation to create new voices for speech synthesis makes HMM-based speech synthesis very attractive. The most popular speaker adaptation approaches in speech synthesis are based on maximum likelihood linear transforms (MLLT) (M.Gales, 1998) [111] and maximum a posteriori (MAP) adaptation (Gauvain, 1994)[112]. MAP estimation involves the use of prior knowledge about the distributions of model parameters. A major drawback of MAP estimation is that since every Gaussian distribution is individually updated, if the adaptation data are very few then many of the model parameters will not be updated and this results in the speaker characteristics of synthesized speech to often switch between general and target speakers within an utterance. Several attempts, such as vector field smoothing (VFS) (Takahashi and Sagayama, 1995) [113] and structured MAP estimation (Shinoda and Lee, 2001) [114] have been made to overcome this limitation. The two approaches may also be used in combination (V. Digalakis and L. Neumeyer, 1996) [115].These approaches provide means to adjust models using relatively few parameters, thus requiring only a small quantity of speaker-specific data. Several variations of linear transform-based speaker adaptation exists that may be applied to model parameters. These are 1) maximum likelihood linear regression (MLLR) (Leggetter, C, 1995) [116], 2) structural maximum a posteriori linear regression (SMAPLR) (Yamagishi 2009, [117], 3) features- spaced MLLR (constrained maximum likelihood linear regression (CMLLR) (M.Gales, 1998) [111] and 4) constrained structural maximum a posteriori linear regression (CSMAPLR) (Y. Nakano 2006, O. Siohan, 2002) [118] [119]. The baseline T-T-S uses CMLLR is used during training and of synthesis system.Anastasakos, 1996 [120] describes Speaker adaptive training (SAT) that uses speaker dependent transforms during training of speaker independent HMM acoustic model, such that the speaker acoustic model is comprised of both the canonical acoustic model(average voice model) (Yamagishi, J., Kobayashi, T., 2007) [121] and speaker dependent transforms (Yamagishi, 2006)[122].Adaptation may be performed in supervised mode-where the full correct context-dependent (supra segmental features) labels that are predicted from text using a T-T-S front-end manually or annotated automatically for the adaptation data is known and in unsupervised form-where where the true transcription of adaptation data is unknown. Till date, supervised adaptation has been mostly used: These rich ‘full context’ models make unsupervised adaptation difficult for synthesis. King et al., 2008 [123] proposed a solution to this problem by only using phonetic labels for adaptation and evaluated the performance of this approach. He reported that the use of unsupervised adaptation degraded its intelligibility but its similarity to the target speaker and naturalness of synthesized speech were less severely impacted.

(2*) Interpolation (mixing voices)*

The interpolation technique enables us to synthesize speech with untrained voice characteristics. The idea of using interpolation was first applied to voice conversion, where pre-stored spectral patterns were interpolated among multiple speakers (Iwahashi and Sagisaka, 1995) [124]. It was also applied to HMM- based speech synthesis, where HMM parameters were interpolated among some representative HMM sets (Yoshimura et al., 1997) [125]. The main difference between Iwahashi and Sagisaka's technique and Yoshimura et al.'s one was that as each speech unit was modeled by an HMM, mathematically-well-defined statistical measures could be used to interpolate the HMMs.

(3) Eigenvoice (producing voices)

The use of the interpolation technique enables us to obtain various new voices by changing the interpolation ratio between representatives HMM sets even if no adaptation data are available. However, as we increase the number of representative HMM sets to enhance the capabilities of representation, it becomes difficult to determine the interpolation ratio to obtain the required voice. To address this problem Shichiri et al., 2002 [127] applied the eigenvoice technique (Kuhn et al., 2000) [126] to HMM-based speech synthesis. The eigenvoice technique, which can reduce the number of parameters to be controlled, and this enables us to manually control the voice characteristics of synthesized speech by setting the weights. However, it introduces another problem in that it is difficult to control the voice characteristics intuitively because none of the eigen-vectors usually represents a specific physical meaning.

(4) Footprint

In statistical parametric synthesis, the footprint is usually small because we store statistics of acoustic models rather than the multi-templates of speech units as in the case of unit-selection synthesis. For example, the footprints of Nitech's Blizzard Challenge 2005 voices were less than 2 MBytes with no compression (Zen et al., 2007c)[128]. Additional reduction was also possible with small degradation in quality by utilizing vector quantization, using fixed-point numbers instead of floating-point numbers, pruning phonetic decision trees (Morioka et al., 2004) [129]and/or tying model parameters (Oura et al., 2008b)[130]. For example, (Morioka et al., 2004) [129] demonstrated that HMM-based speech synthesis systems whose footprints were about 100 KBytes could synthesize intelligible speech by properly tuning various parameters.

(5) Robustness

Statistical parametric speech synthesis is more "robust" than unit-selection synthesis. Factors such as 1) presence of noise or fluctuations due to the recording conditions 2) lack of phonetically balanced sentences resulting in lack some units would significantly degrade the quality of synthetic speech. Yamagishi et al.,2008 [131] reported that statistical parametric speech synthesis, especially AVSS, was much more robust to these kinds of factors .The reason cited is that adaptive training can be perceived as a general version of several feature-normalization techniques such as cepstral mean/variance normalization and stochastic matching.

(6) Development of Multilingual Text-to-speech synthesis

The statistical parametric speech synthesis can support multiple languages because only the contextual factors to be used depend on each language. Takamido *et al.*,2002 [132] showed that an intelligible HMM-based speech synthesis system could be built by using approximately 10 minutes from a single-speaker, phonetically balanced speech database. This property is of significant importance to support numerous languages because few speech and language resources are available in many languages. However, within statistical parametric synthesis, the adaptive training and adaptation framework allows multiple speakers and even languages to be combined into single models, thus enabling multilingual synthesizers to be built. Latorre et al., 2006 [133] and Black, A., and Schultz, T, 2006 [134] proposed building such multilingual synthesizers using combined data from multiple languages.

3.2.3 Disadvantages:

Although the operation and advantages of statistical parameter speech synthesis is impressive, a few disadvantages are associated with it. First, the parameters must be automatically derivable from databases of natural speech;

second the parameters must give rise to high quality synthesis; finally, the parameters must be predictable from text; the synthesis quality is intelligible but nowhere close to natural speech.

4. Implementation of HMM –based speech synthesis system

In this section, the several key system components namely such as lexicon and phone set, acoustic feature extraction, HMM topology and speaker adaptation which are very important for implementation of HMM based speech synthesis has been described. Table 4 shows typical configurations of HMM –based T-T-S systems followed by a brief description of the components (John Dines & Yamagishi, 2009) [136].

Table 4: Configurations of HMM-Based T-T-S Systems

Configuration	T-T-S
General	
Lexicon	Unisys
Phone set	GAM(56 phones)
Acoustics parameterization	
Spectral analysis	STRAIGHT (Fo adaptive window)
Feature extraction	Mel-generalized cepstrum(+Δ+Δ2) +logF0 +bndap(+Δ+Δ2)
Feature dimensionality	120+3+15
Frame shift **Acoustic modeling**	5ms
Number of states per model	5
Number of streams	5
Duration modeling	Explicit duration distribution(HSMM)
Parameter tying	Shared decision tree(MDL)
State emission distribution	Single Gaussian pdf
Context	Full(quinphone + prosody)
Training	Average voice(ML-SAT)
Speaker adaptation	CMLLR or CSMAPLR

4.1 Lexicon and phone set:

The lexicon describes the set of words known by the system and their pronunciation(s). We can generate pronunciations that lie outside the lexicon using letter –to-sound (LTS) methods. The Unisys lexicon [135] with general American accent (GAM) consists of 56 phones. A version of the Unisyn lexicon using an Arpabet-like set consists of 45 phonemes. The results of lexicon evaluations are shown in Table 5 [John Dines & Yamagishi, 2009] [136]. It is observed that the Unisyn lexicon gives slightly better objective measures Mel cepstral distance (MCD) and V/UV error. For an optimal lexicon for applications in T-T-S, the phone sequences produced by the lexicon should have good correlation with acoustic data.

Table 5: Comparison of Lexica for T-T-S

Lexicon	Phone set (size)	T-T-S		
		MCD	RMSE of log Fo	V/UV Error
CMU	CMU(39)	5.63	198	16.9
Unisys	GAM(56)	5.56	198	15.7
Unisys	Arpabet(45)	5.60	198	16.3

4.2 Acoustic Feature extraction:

Acoustic features should provide necessary information to reconstruct the speech signal, normally including pitch and excitation information. The characteristics of LSP-type parameters such as good quantization and interpolation are considered to be of importance in statistical parametric synthesis because statistical modeling is closely related to quantization and synthesis is closely related to interpolation. LSP-type parameters have been applied instead of cepstral parameters to HMM-based speech synthesis in [137][138][139][140] (Nakatani et al., 2006; Ling et al., 2006; Zen et al., 2006b; Qian et al., 2006). The Marume et al., 2006 [141] compared LSPs, log area ratios (LARs), and cepstral parameters in HMM based speech synthesis and reported that LSP-type parameters achieved the best subjective scores for these spectral parameters. Kim et al. [142] also reported that 18-th order LSPs achieved almost the same quality as 24-th order mel-cepstral coefficients.Several techniques of combining spectral analysis and model training have recently been proposed. These techniques, especially those of (Toda and Tokuda, 2008) and (Wu and Tokuda, 2009[143] [144] are based on a similar concept to analysis-by-synthesis in speech coding and the closed-loop training (Akamine and Kagoshima, 1998) [145] for concatenative speech synthesis. Such closed-loop training can eliminate the mismatch between spectral analysis, acoustic-model training, and speech-parameter generation, and thus improves the quality of synthesized speech.

Most current synthesis systems use Mel-frequency cepstral coefficients (MFCCs) (Dominik Niewiadomy) [146] as a feature vector although the standard MFCC does not provide a proper synthesis scheme. The T-T-S quality degrades as the feature analysis order decreases and T-T-S intelligibility is not significantly affected by order analysis. T-T-S features are normally based on variations of Mel-generalized cepstrum analysis (K. Koishida, 1994) [147] and may incorporate STRAIGHT F0-adaptive spectral analysis (H. Kawahara, 1999) [148].

4.3 Model Topology:

Model topology describes the manner in which the states in the HMM set are arranged. The two aspects namely, 1) number of emitting states in each model and 2) as state transition modeling (eg. Left-right, ergodic, explicit duration pdf) are considered as part of model topology. In T-T-S, 5 state left-right HSMM topology is normally used. K. Prahallad & A. W. Black, 2006 [149] experiments with two different HMM topologies (fully connected state model and forward connected state model) for sub-phonetic modeling to capture the deletion and insertion of sub-phonetic states during speech production process and shown that the experimented HMM topologies have higher log likelihood than the traditional 5-state sequential model. However, a 5 state left to right topology has been chosen to be the optimal configuration.

Parameter smoothing and parameter tying techniques, such as decision tree state tying can also be viewed as model topology research. Minimum description length (MDL) (Rissanen, 1980) [150] criterion-based phonetic decision-tree clustering (Shinoda and Watanabe, 2000) [151] has been used in the HMM-based speech synthesis system to balance model complexity and accuracy. As the amount of training data used in speech synthesis is usually less, MDL criterion that is based on asymptotic assumption, is theoretically invalid because the assumption fails. One possible solution to this problem is dynamically changing the complexity of models. Kataoka et al.,2004 [152] proposed a phonetic decision-tree backing-off technique for HMM-based speech synthesis that could dynamically vary the size of phonetic decision trees at run-time according to the text to be synthesized.

4.4 Improving Durational modeling accuracy using HMM:

The HMM only provides a coarse approximation of the underlying process for the generation of acoustic observations especially; the underlying Markov assumption constrains the state occupancy duration to be exponentially distributed. This is often inconsistent with the known duration distributions of the observation sequences being modeled. However, these assumptions hold for real speech. Because speech parameters are directly generated from acoustic models, their accuracy affects the quality of synthesized speech. Beginning from with the work of Ferguson, 1980 [153] and Levinson, 1986 [154], the most primary step taken to improve modeling of the HMM has been to include dynamic features (S. Furui, 1981) [155] in the feature vector and has significant impact on T-T-S. To improve the model structure accuracy, methods such as hidden semi-Markov models (HSMMS) (H.Zen, K.Tokuda, &A.W Black, 2009) [156] that provides explicit model of state duration through simple modification was introduced in the training section (M. Ostendorf, 1996) [157]. Zen *et al.*, 2004 [158] reported slight improvements in speaker-dependent systems. The use of HSMMs makes it possible to simultaneously re-estimate state output and duration models. The adaptation and adaptive training techniques for HSMMs were also derived (J. Yamagishi, 2009) [159]. However, Tachibana *et al.*, 2006 [160] reported that the use of HSMM was essential to adapt state-durations distributions. Y. Nakano, M. Tachibana, 2006 [161] exploited the explicit relationship between static and dynamic relationship has during inference of feature vectors. For consistency, this explicit relationship should also be taken into account during model parameter estimation, leading to the development of the trajectory HMM (K. Tokuda, 2003) [162]. Jian Yu &Meng Zhang, 2007[163] derived new training frameworks, e.g. minimum generation error (MGE) criterion which has been shown to benefit the T-T-S performance.

4.5 Over-smoothing

In the basic system, the speech parameter generation algorithm is used to generate spectral and excitation parameters from the HMMs that are often excessively smooth compared with those of natural speech. Poor modeling accuracy may cause over-smoothed parameters, and lead to quality degradation of synthesized speech. Over-smoothing is classified into two types: the over-smoothing in time domain and over- smoothing in frequency domain (Meng Zhang, 2008)[164].T. Drugman, 2009[165] shows that the over-smoothing in frequency domain is the main factor which influences the quality of synthesized speech and it is generally caused by training algorithm (ML-estimation) accuracy problem whereas over-smoothing in time domain which is caused due to limited model structure [5 state left to right with no skip] can nearly be ignored.

4.6 A new Articulatory paradigm for controlling synthetic speech quality

HMM based speech synthesizers present a certain unnaturalness degree due to the waveform generation part, which consists of a source-filter model wherein the excitation is assumed to be either a periodic pulse train or a white noise sequence. However, this model makes synthetic voice sound buzzy. Toda et al.,2007 [90] proposed a speech parameter generation algorithm considering global variance (GV) that reduces the buzziness in synthesized speech and improves the speech quality. This was one of the main components of Nitech’s Blizzard Challenge 2005 system. Raitio, 2008[166] uses inverse filtering technique in parametric speech synthesis which tries to better approximate the voiced excitation to the residual that represent more details of source than the noise but do not model relevant characteristics of the glottal source. The source-tract type of speech model has been successfully used in HMM based synthesis (J.Cabral, 2010) [167]; the system models the glottal source and vocal tract filter using LPC parameters. During synthesis, the excitation is obtained by transforming a real glottal pulse using F0 and the glottal parameters generated by the synthesizer. However, this approach does not allow control over glottal parameters related to voice quality and does not model the correlation between F0 and the glottal parameters. Joao et al., 2011[169] used an acoustic glottal source model, the Liljencrants-Fant (LF) model (G.Fant, J. liljencrants, 1985) [168] in the synthesis part. Here, a selected LF-model signal was passed through a post-filter to obtain a spectrally flat excitation (glottal post filtering).The synthesized speech was generated by shaping the excitation with the spectral envelope. The results based on perceptual tests showed that speech thus generated was more natural than that obtained using the impulse train. Further, Joao et al., 2011[169] incorporated the LF-model into a standard HMM- based speech synthesizer by using the Glottal Spectral Separation (GSS) method (D.Talkin)[170] for analysis –synthesis and adapting the acoustic modeling part to train the glottal parameters. This proposed HTS-LF system has a major advantage as it provides control over glottal parameters for voice quality transformations.

5. Speech Databases for speech synthesis

5.1 Characteristics of major databases:

Building high quality synthetic voices requires high degree of control, since the flavor of the voice invariably reflects the nature of the recordings. For a speech database to serve as the basis for constructing a synthetic voice, the recordings should be of studio quality and free of noise. Since perfect quality open-domain synthesis is not yet possible, the recorded utterances need to reflect the target domain – in particular, by being phonetically balanced. Finally, the prosody of speech needs to be controlled so that the synthetic voice's style of delivery is both consistent and appropriate satisfying these requirements makes a corpus designed for synthesis, as opposed to merely collected.

1) *FM Radio News Corpus*: The most common resource for speech synthesis research is Boston University's FM Radio News Corpus (M. Ostendorf, 1996) [171] was recorded in 1994. It consists of seven professional radio announcers reading either pre-edited or off-the-wire news stories. As such, the recordings are well suited for a study of prosody in speech – the primary intention of this corpus.

2) *TIMIT*: The TIMIT corpus was recorded in 1986 and collected to support the training and testing of automatic speech recognition systems. TIMIT was designed to study acoustic-phonetic knowledge and was commissioned by DARPA (W.Fisher, 1986) [172]. In 1997, a freely available, single-speaker version of the TIMIT prompt set was

released for synthesis research by the University of Edinburgh (CSTR USKED TIMIT, 2002) [173]. But because the phoneme sequences of this database are unusual, experience has shown that TIMIT based voices tend to be sub-par.

3) *ARCTIC*: An Arctic "database" is a reading of the Arctic prompt set (plus associated files) by a single speaker in a specified style of delivery. Each Arctic database consists of nearly 1150 utterances, most being between one and four seconds long. The prompt list is split into two sets (A and B), each of which is designed to be phonetically balanced American English and have diphone coverage representative of the source material. The wave files were recorded in a sound proof booth at 32,000 Hz with simultaneous EGG (laryngograph) measurements. In all cases the lexical and phonetic descriptions derive from the US English front-end module distributed with Festival. In this configuration Festival employs CMUDICT [174] as its dictionary component. Thus the two accented databases are described using a General American phoneme set and lexicon, despite any speaker-specific deviation.

5.2 Speech Synthesis and Development in Indian Scenario

Speech technologies can play a very important role in development of applications for common people in a multilingual society such as India which has about 1652 dialects/native languages. Till 1990s, Indian speech synthesizers were research synthesizers, generating small segments of speech in non-real time and the progress was very slow. Speech synthesizers were not developed for commercial purpose. In the 90s, Government of India had funded Indian language projects generously, through Technology Development for Indian Languages (TDIL) and other schemes.

5.2.1 Current Research projects in India:

Some of the institutions in India are engaged in speech synthesis. The IIT Madras has worked on a novel scheme where the 'unit 'is a character of written 'text'. The Tata Institute of Fundamental Research (TIFR), Mumbai has reported unlimited continuous speech synthesizer using formant synthesis technique. Whereas TIFR (Furtado X A & Sen A,1996) [175]and Central Electronics Engineering Research Institute (CEERI) (Agrawal S S.,1992)[176] worked with formant synthesis, ISI, Kolkata(Dan T K, Datta,1995) [177], Indian Institute of Information Technology (IIIT), Hyderabad (Kishore S.P.,2002)[178], center for Development of Advanced Computing (CDAC), Pune and Kolkata developed concatenation-based synthesizers. Between the concatenation and formant synthesizers, the quality obtained so far is comparable. Speech synthesizers based on Festival has been developed in languages including Hindi, Bangla, Kannada, Marathi and Tamil.

5.2.2 Speech Corpora Collected by the LDC-IL

Linguistic Data Consortium for Indian Languages (LDCIL) is the Consortium responsible to create the database and shall provide forum for the researchers all over the world to develop speech application using the collected data in various domains. The LDC-IL has collected Speech databases in various Indian languages, the details are described in (Agrawal S. S., 2010) [179]. The research that has been carried out is mostly for text to speech synthesis which uses phoneme/syllables concatenation on isolated words and is either based either on concatenative or formant synthesis techniques. The need of the hour is to work on the continuous speech and apply latest techniques such as Hidden Markov Models for development of T-T-S for general purpose or limited domain to achieve true application potentials of speech synthesis. Although Indian language speech synthesis has come up a long way, the amount of work for Indian languages in speech domain has not yet reached to a critical level to be used as real communication tool, as that in other languages of developed countries.

6. Discussions and Conclusions

Synthetic speech has been developed steadily especially during the last decades. We have presented an overview of speech synthesis-past progress and current trends, giving step by step progress in this field. The three basic methods for synthesis are the formant, concatenative, and articulatory synthesis. The formant synthesis is based on the modeling of the resonances in the vocal tract and is perhaps the most commonly used during last decades. However, the concatenative synthesis which is based on playing prerecorded samples from natural speech is more popular. In theory, the most accurate method is articulatory synthesis which models the human speech production system directly, but it is also the most difficult approach. Currently, the statistical parametric speech synthesis has been the most rigorously studied approach for speech synthesis. We can see that statistical parametric synthesis offers a wide range of techniques to improve spoken output. Its more complex models, when compared to unit-selection synthesis, allow for general solutions, without necessarily requiring recorded speech in any phonetic or prosodic contexts. The unit-selection synthesis requires very large databases to cover examples of all required prosodic, phonetic, and stylistic variations which are difficult to collect and store. In contrast, statistical parametric synthesis enables models to be combined and adapted and thus does not require instances of any possible combinations of contexts. Additionally, T-T-S systems are limited by several factors that present new challenges to researchers. They are 1) The available speech data are not perfectly clean 2) The recording conditions are not consistent & 3) Phonetic balance of material is not ideal. Means to rapidly adapt the system using as little data as a few sentences would appear to be an interesting research direction. It is seen that synthesis quality of statistical parametric speech synthesis is fully understandable but has “processed quality” to it. Control over voice quality (naturalness, intelligibility) is important for speech synthesis applications and is a challenge to the researchers. As described in this review, unit selection and statistical parametric synthesis approaches have their own advantages and drawbacks. However, by proper combination of the two approaches, a third approach could be generated which can retain the advantages of the HMM based and corpus based synthesis with an objective to generate synthetic speech very close to the natural speech. It is suggested that a more detailed evaluation and analysis, plus integration of HMM based segmentation and labeling for building database and HMM based search for selecting best suitable units shall aid in using the better features of the two methods.

References and Literature

[1] X.Huang, A.Acero, H.-W. Hon, “Spoken Language Processing”, *Prentice Hall PTR*, 2001
[2]T. Dutoit, “An Introduction to Text-to-Speech Synthesis”, *Kluwer Academic Publishers*, 1997
[3] D. Jurafsky and J.H. Martin, “Speech and Language Processing”, *Pearson Education*, 2000
[4]H.Zen, K.Tokuda , &A.W Black “ Statistical parametric speech synthesis”, *speech communication* , doi:10.1016/j.specom.2009.04.004 2009
[5]L.R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition”, In *proc. of the IEEE*, Vol. 71, no.2, pp.227-286, Feb 1989
[6]A.Falaschi, M.Guistianiani, M.Verola, “A hidden markov model approach to speech synthesis”, In *proc. of Eurospeech*, Paris, France, 1989, pp 187-190
[7]S. Martincic- Ipsic and I. Ipsic, “Croatian HMM Based Speech Synthesis,” *28th Int. Conf. Information Technology Interfaces ITI 2006*, pp.19-22, 2006, Cavtat, Croatia
[8] S.S. Agrawal, “ Speech Synthesis for Natural Sounding” *10th M.S. Narayana Memorial Lecture* (Keynote address) delivered during NSA-2001, held at VIT, Vellore(TamilNadu),2001
[9]Cahn, J. E., “Generating Expression in Synthesized Speech”, *Master’s Thesis*, MIT, 1989.http://www.media.mit.edu/~cahn/masters-thesis.html
[10] Cahn, J. E., The Generation of Affect in Synthesized Speech, *Journal of the American Voice I/O Society*, 8, July 1990, p. 1-19.
[11] Murray, I. R., “Simulating emotion in synthetic speech”, *PhD Thesis*, University of Dundee, UK, 1989.
[12] Murray, I. R., & Arnott, J. L., “Implementation and testing of a system for producing emotion-by-rule in synthetic speech”, *Speech Communication, 16,* p. 369-390.
[13] Montero, J. M., Gutiérrez-Arriola, J., Palazuelos, S.,Enríquez, E., Aguilera, S., & Pardo, J. M., “ Emotional Speech Synthesis: From Speech Database to T-T-S”, *ICSLP 98*, Vol. 3, p. 923-926.
[14] Burkhardt, F., “*Simulation emotionaler Sprechweise mitSprachsyntheseverfahren”* [Simulation of emotional manner of speech using speech synthesis techniques], PhD Thesis, TU Berlin, 2000. http://www.kgw.tuberlin. de/~felixbur/publications/diss.ps.gz
[15] Burkhardt, F., & Sendlmeier, W. F., “Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis”, *ISCA Workshop on Speech &Emotion, Northern Ireland 2000,* p. 151-156.
[16] S.Lemmetty, “Review of Speech Synthesis Technology”, *Master’s Thesis*, Helinski University of Technology
[17] Heuft, B., Portele, T., & Rauth, M. (1996), “Emotions in Time Domain Synthesis” *ICSLP 96*.
[18] Edgington, M., “Investigating the Limitations of Concatenative Synthesis”, *Eurospeech 97*.
[19] Vroomen, J., Collier, R., & Mozziconacci, S. J. L., “Duration and Intonation in Emotional Speech”, *Eurospeech 93*, Vol. 1, p. 577-580.
[20] Rank, E., & Pirker, H., “Generating Emotional Speech with a Concatenative Synthesizer”, *ICSLP 98, Vol. 3*, p.671-674.
[21] Montero, J. M., Gutiérrez-Arriola, J., Colás, J., Enríquez,E., & Pardo, J. M., “Analysis and Modeling of Emotional Speech in Spanish”, *ICPhS 99*, p. 957-960.
[22] Iriondo, I., Guaus, et al., “Validation of an Acoustical Modeling of Emotional Expression in Spanish using Speech Synthesis Techniques”, *ISCA Workshop on Speech & Emotion, Northern Ireland 2000*, p. 161-166.
[23]Murray, I. R., Edgington, M. D., Campion, D., & Lynn., “ Rule-based Emotion Synthesis Using Concatenated Speech”, *ISCA Workshop on Speech & Emotion, Northern Ireland 2000*, p. 173-177.
[24]Schröder, M., “Can emotions be synthesized without controlling voice quality?” Phonus 4, Research *Report of the Institute of Phonetics*, University of the Saarland, p.37-55. http://www.dfki.de/~schroed.
[25]Mozziconacci, S. J. L., “*Speech Variability and Emotion: Production and Perception”*, *PhD Thesis*, Technical University, Eindhoven, 1998.
[26]Mozziconacci, S. J. L., & Hermes, D. J.,“Role of intonation patterns in conveying emotion in speech”, *ICPhS 1999*, 2001-2004.
[27]Chung, S.-J., “Vocal Expression and Perception of Emotion in Korean”, *ICPhS 99*, p. 969-972.
[28]Stevens, K.,“Towards a model for speech recognition,” *J. Acoustic. Soc. Am.*, 32, pp.47-55, 1960
[29]Olive, J.P. (1977), “Rule synthesis of Speech from Dyadic Units”, *Proc. ICASSP-77*, pp568-570
[30]Olive, J. P. (1990), "A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds," *Proc. ESCA Workshop on Speech Synthesis*, Autrans, France.

[31]Olive, J.P. and Liberman, M.Y. (1985), “Text-to-speech- an overview” JASA Suppl 1, vol. 78 (Fall), S6

[32]Hakoda, K. S. Nakajima, T. Hirokawa and H. Mizuno (1990), "A new Japanese text-to speech synthesizer based on COC synthesis method," In *Proc. ICSLP90*, Kobe, Japan.

[33]Nakajima, S. and H. Hamada (1988), “Automatic generation of synthesis units based on context oriented clustering”, In *Proc. ICASSP-88*

[34]Sagisaka, Y. (1988), “Speech synthesis by rule using an optimal selection of non-uniform synthesis units”, In *Proc. ICASSP -88.*

[35]Sagisaka, Kaiki, Iwahashi, and Mimura, 1992) Sagisaka, Y., Kaiki, N., Iwahashi, N. and Mimura, K. (1992), “ATR v-TALK speech synthesis system”, In *Proc. ICSLP 92*, Banff, Canada

[36]Atal and Hanauer, “Speech Analysis and Synthesis by Linear Prediction of the Speech Wave” no.2 part 2, vol.51, *Acoustical society of America*, 1971

[37]T.Irino, Y.Minami, T. Nakatani, M. Tsuzaki, and H. Tagawa, “Evaluation of a speech recognition/Generation method based on HMM and STRAIGHT”, *ICSLP2002*, Denver, Colorado

[38]Moulines E., Emerard F., Larreur D., Le Saint Milon J., Le Faucheur L., Marty F.,Charpentier F., Sorin C., “ A Real-Time French Text-to-Speech System Generating High-Quality Synthetic Speech”, *Proceedings of ICASSP 1990* (1): 309-312.

[39]Charpentier F., Moulines E. (1989), “Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones” *Proceedings of Eurospeech 89* (2): 13-19.

[40]Moulines E., Laroche J., “Non-Parametric Techniques for Pitch-Scale Modification of Speech” Speech *Communication* 16 (1995): 175-205.

[41]Kortekaas R., Kohlrausch A, “Psychoacoustical Evaluation of the Pitch-Synchronous Overlap-and-Add Speech-Waveform Manipulation Technique Using Single-Formant Stimuli”, *Journal of the Acoustical Society of America, JASA*, vol.101 (4): 2202-2213.1997

[42]Roucos and Wilgus, 1985, and systems for diver´s speech restoration also did direct processing of the waveform,

[43]Liljencrants, 1974, Metoder for propotionell frekvenstransponering av en signal.” Swedish patent number 362975.

[44]R.sproat, J. Hirschberg, and D. Yarowsky, “A corpus-based synthesizer”, *Proc. ICSLP*, pp.563-566, 1992

[45]Van Erp. A and L. Boves.,“Manual segmentation and labeling of speech”, *Proc. of speech* 1988, pp. 1131-1138.

[46]Wang, H. C., R. L. Chiou, S. K. Chuang and Y. F. Huang, “A phonetic labeling method for MAT database processing”, *Journal of the Chinese Institute of Engineers*, 22(5), 1999,pp. 529-534.

[47]Ljolje, A. and M. D. Riley, “Automatic segmentation of speech for T-T-S”, In Proc. *of European Conference on Speech Communication and Technology”*, 1993, pp. 1445-1448.

[48]Demuynck, K. and T. Laureys, “A Comparison of Different Approaches to Automatic Speech Segmentation,” *Proceedings of International Conference on Text, Speech and Dialogue*, 2002, pp. 277--284.

[49] van Santen, J. P. H. and R. Sproat, “High-accuracy automatic segmentation,” *Proceedings of European Conference on Speech Communication and Technology*, 1990, pp.2809–2812.

[50]Bonafonte, A., A. Nogueiras and A. Rodriguez-Garrido,“Explicit segmentation of speech using Gaussian models,” *Proceedings of International Conference on Spoken Language Processing*, 1996, pp. 1269-1272.

[51]Torre Toledano, D., M. A. Rodrguez Crespo and J. G. Escalada Sardina, “Trying to Mimic Human segmentation of Speech Using HMM and Fuzzy Logic Post-correction Rules, “*Proceedings of Third ESCA/COCOSDA Workshop on speech synthesis*, 1998, pp.207-212.

[52] Sethy, A. and S. Narayanan, “Refined Speech Segmentation for Concatenative Speech Synthesis” *Proceedings of International Conference on Spoken Language Processing*, 2002, pp. 149-152.

[53]F.Malfere, o.Deroo, T. Dutiot, and C. Ris, “Phonetic alignment: speech synthesis vs. Viterbi-based”, *Speech communication* vol. 40, pp.503-515, 2003.

[54]J.Keshet, S.S Shwartz, Y.Signer, and D.Chazan, “Phoneme alignment based on discriminative learning”, *Proc. of Interspeech’05*, pp.2961-2964, 2005.

[55]K. Torkkola, “Automatic alignment of speech with phonetic transcription in real time”, *Proceedings of IEEE ICASSP’98*.pp. 611-614, 1998

[56]B.L. Pellom and J.H. Hansen.,“Automatic segmentation of speech recorded in unknown noisy channel characteristics”, *Speech Communication*, vol 25.pp. 97-116, 1998.

[57] F. Brugnara, D. Falavigna , and Omologo, “Automatic segmentation and labeling of speech based on hidden markov models”, Speech Communication, vol. 12,pp 97-116,1998.

[58]J. Adell, A.Bonafonte, J.A Gomez, and M.J. Castro, “Comparative study of automatic phone segmentation methods for T-T-S”, *Proc. of IEEE ICASSP’08*, pp. 4457-4460, 2008.

[59]I. Mporas , T. Ganchev and N. Fakotakis, “A hybrid architecture for automatic segmentation of speech waveforms,” Proceedings of IEEE ICASSP’08,PP. 4457-4460, 2008

[60]J Garofolo, “Getting started with the DARPA-TIMIT CD-ROM: an acoustic phonetic continuous speech database, “National institute of Standards and technology (NIST), Gaithersburg, MD, USA, 1988.

[61] A.J. Hunt and A.W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. 1, pp. 373–376, 1996.

[62]A. Black and A. Font Llitj´os, “Unit selection without a phoneme set,” In *IEEE Workshop on Speech Synthesis*, Santa Monica, CA. 2002.

[63]A. Black and K.Lenzo, “Optimal data selection for unit selection synthesis,” *4th ESCA Workshop on Speech Synthesis*, Scotland. 2001.

[64]J. Kominek and A.Black,2003 ., “The CMU ARCTIC speech databases for speech synthesis research,” Tech. Rep. CMU-LTI-03-177 http://festvox.org/cmu arctic/, Language Technologies Institute, Carnegie Mellon University,PiT-T-Sburgh, PA, 2003.

[65]Chou, F.-C., C.-Y. Tseng and L.-S. Lee, “Automatic Segmental and Prosodic Labeling of Mandarin Speech,” *Proceedings of International Conference on Spoken Language Processing*, 1998, pp. 1263-1266.

[66]W. N. Campbell and A. Black, “Prosody and the selection of source units for concatenative synthesis,” in *Progress in Speech Synthesis*, R. Van Santen, R.Sproat, J.Hirschberg, and J.Olive, Eds. 1996, pp. 279–292, Springer Verlag.

[67]N. Mizutani, K. Tokuda, and T. Kitamura, “Concatenative speech synthesis based on HMM” In *Proc. Autumn Meeting of ASJ*, pages 241–242, 2002 (In Japanese).

[68]C. Allauzen, M. Mohri, and M. Riley,“Statistical modeling for unit selection in speech synthesis” In *Proc. of the 42nd meeting of the ACL*, 2004.

[69]S. Sakai and H. Shu, “A probabilistic approach to unit selection for corpus-based speech synthesis” In *Proc. Interspeech (Eurospeech)*, pages 81–84, 2005.

[70] Z.-H. Ling and R.-H. Wang, “HMM-based unit selection using frame sized speech segments” In *Proc. Interspeech (ICSLP)*, pages 2034–2037, 2006.

[71]Christian Weiss and Wolfgang Hess, “Conditional random fields for hierarchical segment selection in text-to-speech synthesis”, In *Proc. Interspeech (ICSLP)*, pages 1090–1093, 2006.

[72]Iida, A., Campbell, N., Iga, S., Higuchi, F., & Yasumura, M., “A Speech Synthesis System for Assisting Communication”, *ISCA Workshop on Speech & Emotion,Northern Ireland 2000*, p. 167-172.

[73]Marumoto, T., & Campbell, N., “Control of speaking types for emotion in a speech re-sequencing system [in Japanese]”, In *Proc. of the Acoustic Society of Japan, Spring meeting 2000*, p. 213-214.

[74] X. Huang, A. Acero,. Acero, H. Hon, Y. Ju, J Liu,S. Meridth, and M. Plumpe, “ Recent Improvements on Microsoft’s trainable text –to-speech synthesizer: Whistler” *In ICASSP-97,Vol II, pages959-962, Munich, Germany,1997*

[75]A. Nagy,P.Pesti, G.Nemeth, T.Bohm, “Design Issues in Corpus based speech synthesizer (In Hungarian)” *Hungarian Journal of Communications,vol 2005/1,pp,18-24,Budapest, Hungary,2005.*

[76]Y.Sagisaka, N.Kaiki, N.Iwahashi, and K. Mimura, “ATR-v-TALK speech synthesis system “*In Proc. of ICSLP 92, volume 1, pages 483-486, 1992.*

[77]R.Donovan and P.Woodland, “Improvement in an HMM- based speech synthesizer*”, In Eurospeech95, volume 1, pages 573-576, Madrid, Spain, 1995*

[78]Campbell, N. and Black, A., “Prosody and the selection of source units for concatenative synthesis” *Progress in Speech Synthesis, ed. van Santen, J. Sproat, R., Olive, J., Hirsberg J., Springer, New York. pp. 663-666. 1997.*

[79]Alan W Black and Paul Taylor, “Automatically clustering similar units for unit selection in speech synthesis” *In Proc. of Eurospeech 97, vol. 601-604, Rhodes, Greece.*

[80]L. Breiman and A. Black., “Prosody and the selection of the source units for concatenative synthesis”, *In J. van Santen, R.Sproat, J.Olive, and J.Hirschberg,editors, Progress in Speech Synthesis, pages 279-282,Springer Verlag,1996*.

[81]A.Conkie and S. Israd, “ Optimal coupling of diphones”, Springer, *New York. pp. 663-666. 1997.*

[82]T.Yoshimura, K.Tokuda, T. Masuko, T. Kobayashi and T. Kitamura,“Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis*”In Proc. of ICASSP 2000, vol 3, pp.1315-1318, June 2000.*

[83]J. Ferguson, Ed., “Hidden Markov Models for speech” *IDA, Princeton, NJ, 1980*

[84]L.R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition” *Proc. IEEE, 77(2), pp.257-286, 1989*

[85]L.R.Rabiner and B.H. Juang, “Fundamentals of speech recognition”, *Prentice-Hall, Englewood Cliff,New Jersey,1993.*

[86]K. Tokuda , H. Zen, J. Yamagishi, T. Masuko, S. Sako, T. Toda, A.W. Black, T. Nose , and K. Oura, “The HMM based synthesis system(HTS)” http://hts.sp.nitech.ac.jp/.

[87]S.Young,G. Evermann, M. Gales,et al.,“ *The Hidden Markov Model Toolkit (HTK) version 3.4”*, 2006. http://htk.eng.cam.ac.uk/.

[88]H. Zen, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura,“A hidden semi-Markov model-based speech synthesis system.” *IEICE Trans. Inf.Syst.*, E90-D (5):825–834, 2007.

[89]J. Yamagishi and T. Kobayashi. Average-voice based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Trans. Inf. Syst.*, E90-D (2):533–543, 2007.

[90]T. Toda and K. Tokuda, “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis”, *IEICE Trans. Inf. Syst.*, E90-D (5):816–824, 2007.

[91]J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm”, *IEEE Trans. Audio Speech Lang. Process.*, 17(1), pp.66–83, 2009.

[92] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveign´e, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based *F*0 extraction: possible role of a repetitive structure in sounds”, *Speech Comm.*, 27:187–207, 1999.

[93]H. Zen, T. Toda, M. Nakamura, and K. Tokuda, “Details of Nitech HMM based speech synthesis system for the Blizzard Challenge 2005. *IEICE Trans. Inf. Syst.*, E90-D(1):325–333, Jan. 2007.

[94]H. Zen, T. Toda, and K. Tokuda, “The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006”, *In Blizzard Challenge Workshop*, 2006.

[95]J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals,“A robust speaker-adaptive HMM-based text-to-speech synthesis”, *IEEE Trans. Audio Speech Lang. Process.*, 2009. (accept for publication).

[96]H.Zen, K.Oura, T.Nose, J. Yamagishi, S.Sako, T.Toda, T.Masuko, A.W. Black, K.Tokuda, “Recent development of the HMM-Based Speech Synthesis System(HTS)”, *Proc. 2009 Asia-Pacific Signal and Information Processing Association (APSIPA)*, Sapporo, Japan, October 2009.

[97]Dempster, A., Laird, N., Rubin, D., 1977,“ Maximum likelihood from incomplete data via the EM algorithm”, *Journal of Royal Statistics Society 39, 1–38.*

[98]Fukada,T., Tokuda, K., Kobayashi, T., Imai, S., 1992, “An adaptive algorithm for mel-cepstral analysis of speech”, *In Proc. ICASSP*. pp. 137–140.

[99]Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X.-Y., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2006,“The Hidden Markov Model Toolkit (HTK) version 3.4. http://htk.eng.cam.ac.uk/.

[100]Freij, G., Fallside, F., 1988,“Lexical stress recognition using hidden Markov models”, *Proc. ICASSP*. pp. 135–138.

[101]Jensen, U., Moore, R., Dalsgaard, P., Lindberg, B., 1994, “Modeling intonation contours at the phrase level using continuous density hidden Markov models*”, Comput. Speech Lang.* 8 (3), 247–260.

[102]Ross, K., Ostendorf, M., 1994, “A dynamical system model for generating F0 for synthesis”, *In Proc. ESCA/IEEE Workshop on Speech Synthesis*. pp. 131–134.

[103]Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T., 2002a,“Multi-space probability distribution of HMM”, *IEICE Trans. Inf. Syst*. E85-D (3), 455–464.

[104]Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T. 1998, “Duration modeling for HMM-based speech synthesis”, In *Proc. ICSLP*. pp. 29–32.

[105]Ishimatsu, Y., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2001,“Investigation of state duration model based on gamma distribution for HMM based speech synthesis”, *In Tech. Rep. of IEICE.* vol. 101 of SP 2001-81. pp. 57–62, (In Japanese).

[106]Odell, J., 1995,“The use of context in large vocabulary speech recognition”, *Ph.D. thesis*, University of Cambridge.

[107]Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000,“Speech parameter generation algorithms for HMM-based speech synthesis”In *Proc. ICASSP*. pp. 1315–1318.

[108]Tachiwa, W., Furui, S., “A study of speech synthesis using HMMs” In: Proc. Spring Meeting of ASJ. pp. 239–240,(In Japanese), 1999.

[109] Imai, S., Sumita, K., Furuichi, C., “Mel log spectrum approximation (MLSA) filter for speech synthesis”, *Electronics and Communications in Japan* 66 (2), 10–18, 1983

[110] Stylianou, Y., Cap´pe,O., Moulines, E., 1998, “Continuous probabilistic transform for voice conversion”, *IEEE Trans. Speech Audio Process.* 6 (2), 131–142.

[111] M.Gales, “Maximum Likelihood linear transformations for HMM-based speech recognition,” *Computer speech and language*, vol. 12, no. 2, pp.75-98,1998.

[112] Gauvain, J., Lee, C.-H., 1994,“Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains”, *IEEE Trans. Speech Audio Processing*, 2 (2), 291–298, 1995.

[113]Takahashi, J., Sagayama, S., “Vector-field-smoothed Bayesian learning for incremental speaker adaptation”, pp. 696–699

[114]Takahashi, T., Tokuda, K., Kobayashi, T., Kitamura, T., Shinoda, K., Lee, C.-H., 2001, “A structural Bayes approach to speaker adaptation”, *IEEE Trans. Speech Audio Process*.vol 9, pp. 276–287, 2001

[115]V. Digalakis and L. Neumeyer, “Speaker adaptation using combined transformation and Bayesian methods,” *IEEE Trans. Speech Audio Process*, vol. 2, pp. 294-300, July 1996.

[116]Leggetter,C., Woodland, P., 1995, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech Lang. 9, 171–185.

[117]Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J., 2009, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm”, *IEEE Trans. Audio Speech Lang. Process*. 17 (1), 66–83.

[118]Y. Nakano, M. Tachibana, J.Yamagishi, and T.Kobayashi, “Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis,” in *Proc. ICSLP* 2006, Sep. 2006, pp.2286-2289

[119]O.Siohan, T. Myrvoll, and C.-H. Lee, “Structural maximum a posteriori linear regression for fast hmm adaptation,” *Computer, Speech and language*, vol. 16, no.1, pp.5-24, 2002.

[120]Anastasakos, T., McDonough, J., Schwartz, R., Makhoul, J., “A compact model for speaker adaptive training” In *Proc. ICSLP*. pp. 1137–1140. 1996

[121]Yamagishi, J., Kobayashi,T., “Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training”, *IEICE Trans. Inf. Syst*. E90-D (2), 533–543, 2007.

[122]Yamagishi, J., “Average-voice-based speech synthesis”, *Ph.D. thesis*, Tokyo Institute of Technology, 2006.

[123] King, S., Tokuda, K., Zen, H., Yamagishi, J., 2008, “Unsupervised adaptation for HMM-based speech synthesis”, *In Proc. Interspeech.* pp. 1869–1872.

[124] Iwahashi, N., Sagisaka, Y., “Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks” *Speech Communication*, 16 (2), 139–151, 1995

[125] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., “Speaker interpolation in HMM-based speech synthesis system” In *Proc .of Eurospeech*. pp. 2523–2526, 1997

[126]Kuhn, R., Janqua, J., Nguyen, P., Niedzielski, N., 2000, “Rapid speaker adaptation in eigenvoice space”, *IEEE Trans. Speech Audio Process.* 8 (6), 695–707.

[127]Shichiri, K., Sawabe, A., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., “ Eigenvoices for HMM-based speech synthesis”, In *Proc. ICSLP*. pp.1269–1272, 2002.

[128]Zen, H., Toda, T., Nakamura, M., Tokuda, T., 2007c,“Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005”,*IEICE Trans. Inf. Syst*. E90-D (1), 325–333.

[129]Morioka, Y., Kataoka, S., Zen, H., Nankaku, Y., Tokuda, K., Kitamura, T., 2004, “Miniaturization of HMM-based speech synthesis”, In *Proc. Autumn Meeting of ASJ*. pp. 325–326 (in Japanese)

[130]Oura, K., Zen, H., Nankaku, Y., Lee, A., Tokuda, K., 2008b, “Tying variance for HMM-based speech synthesis”, *In Proc. Autumn Meeting of ASJ*. pp. 421–422 (In Japanese)

[131]Yamagishi, J., Ling, Z.-H., King, S., 2008a, “Robustness of HMM-based speech synthesis”, In *Proc. Interspeech*. pp. 581–584.

[132]Y. Takamido, K. Tokuda, T. Kitamura, T. Masuko, and T. Kobayashi, “A study of relation between speech quality and amount of training data in HMM-based TTS system,” *ASJ Spring meeting*, 2-10-14, pp. 291–292, Mar. 2002 (in Japanese).

[133]Latorre, J., Iwano, K., Furui, S., 2006, “New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer”, *Speech Communication ICAT*. 48 (10), 1227–1242.

[134]Black, A., Schultz, T., 2006, “Speaker clustering for mulitilingual synthesis”, In *Proc. ISCA* itrw multiling. no. 024.

[135]S. Fitt and S. Isard, “Synthesis of regional English using a keyword lexicon,” In *Proc. Eurospeech*, vol. 2, Sep. 1999, pp. 823–826.

[136]John Dines, Junichi Yamagishi and S.King, “Measuring the gap between HMM- based ASR and TTS”, In *Proc. Interspeech* 2009, Brighton,U.K., Sept. 2009

[137]Nakatani, N., Yamamoto, K., Matsumoto, H., “Mel-LSP parameterization for HMM-based speech synthesis”, In *Proc. SPECOM*. pp.261–264, 2006.

[138]Ling, Z.-H., Wang, R.-H., “HMM-based unit selection using frame sized speech segments”, In *Proc. Interspeech*. pp. 2034–2037, 2006

[139]Zen, H.,Toda, T., Tokuda, K., “The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006” In *Proc. Blizzard Challenge Workshop,2006.*

[140]Qin, L., Wu, Y.-J., Ling, Z.-H., Wang, R.-H., 2006, “Improving the performance of HMM-based voice conversion using context clustering decision tree and appropriate regression matrix format”, In *Proc. Interspeech*, pp. 2250–2253.

[141]Marume, M., Zen, H., Nankaku, Y., Tokuda, K., Kitamura, T., “An investigation of spectral parameters for HMM-based speech synthesis”, *In Proc. of Autumn Meeting of ASJ*. pp. 185–186, (in Japanese) 2006

[142]Kim, S.-J., Kim, J.-J., Hahn, M.-S., 2006a.,“HMM-based Korean speech synthesis system for hand-held devices”, IEEE *Trans. Consumer Electronics* 52 (4), 1384–1390.

[143] Toda, T., Tokuda, K., 2008, “Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM”,In *Proc. ICASSP*. pp. 3925–3928.

[144]Wu, Y.-J., Tokuda, K., 2008, “An improved minimum generation error training with log spectral distortion for HMM-based speech synthesis”, In Proc. Interspeech, pp. 577–580.

[145] Akamine, M., Kagoshima, T., 1998, “ Analytic generation of synthesis units by closed loop training for totally speaker driven text to speech system (TOS drive T-T-S)” In *Proc. ICSLP*. pp. 139–142.

[146]Dominik Niewiadomy, Adam Pelikant, “Implementation of MFCC vector generation in classification context”, In *Journal of Applied Computer Science*

[147] K. Koishida, G. Hirabayashi, K. Tokuda, and T. Kobayashi, “Mel generalized cepstral analysis - a unified approach to speech spectral estimation,” in *Proc. ICSLP*, vol. 3, Yokohama, Japan, September 1994,pp. 1043–1046.

[148]H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds,” *Speech Communication*, vol. 27, pp. 187–207, 1999.

[149]K. Prahallad, A. W. Black, and R. Mosur, “Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis,” In *Proc. ICASSP*, Toulouse, France, 2006, pp. 853–856.

[150]Rissanen, J., 1980, “Stochastic complexity in stochastic inquiry”, *World Scientific Publishing Company*

[151]K. Shinoda and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition,” *J. Acoust. Soc. Japan (E)*, vol. 21, pp. 79–86, Mar. 2000.

[152]Kataoka, S., Mizutani, N., Tokuda, K., Kitamura,T., 2004, “Decision-tree backing-off in HMM-based speech synthesis” In *Proc. Interspeech*. pp. 1205–1208.

[153]J. D. Ferguson, “Variable duration models for speech,” In *Proc. of Symp.App. Hidden Markov Models Text Speech*, 1980

[154]S. E. Levinson, “Continuously variable duration hidden Markov models for speech analysis,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*1986, pp. 1241–1244.

[155]S. Furui, “ Cepstral analysis technique for automatic speaker verification”, *IEEE Trans. on Acoustics, Speech, & Signal Process*, vol. 29, pp.254–272, April 1981.

[156]H.Zen, K.Tokuda, &A.W Black, “ Statistical parametric speech synthesis*”, Speech Communication* , doi:10.1016/j.specom.2009.04.004 2009.

[157]M. Ostendorf,V.V. Digalakis, and O. A. Kimball, “From HMM’s to segment models: A unified view of stochastic modeling for speech recognition,” *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 360–378,Sep. 1996.

[158]K. Tokuda, H. Zen, and A. W. Black, “HMM-based approach to multilingual speech synthesis,” in *Text to speech synthesis: New paradigms and advances*, S. Narayanan and A. Alwan, Eds. Prentice Hall, 2004.

[159]J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King,and S. Renals, “A robust speaker-adaptive HMM-based text-to-speech synthesis,” *IEEE Trans. Speech, Audio & Language Process.*, vol. 17,no. 6, pp. 1208–1230, Aug. 2009.

[160]Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, “Constrained structural maximum a posteriori linear regression for average voice based speech synthesis,” In *Proc. ICSLP 2006*, Sep. 2006, pp. 2286–2289.

[161]T. Irino, Y. Minami, T. Nakatani, M. Tsuzaki, and H. Tagawa, “Evaluation of a speech recognition / generation method based on HMM and STRAIGHT,” In *Proc. ICSLP*, Denver, USA, 2002, pp. 2545–2548.

[162] K. Tokuda, H. Zen, and T. Kitamura, “Trajectory modeling based on HMMs with explicit relationship between static and dynamic features,” In *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 865–868.

[162]Y.-J. Wu and R.-H. Wang, “Minimum generation error training for HMM-based speech synthesis,” in *Proc. ICASSP*, Toulouse, France,

[163 Jian Yu,Meng Zhang, Jianhua, Xia Wang, “A novel hmm-based T-T-S system using both continuous HMMs and discrete”, In *Proc. ICASSP* 2007

[164] Meng Zhang, Jianhua Tao, Huibin,Xia Wang , “ Improving HMM based speech synthesis by reducing over-smoothing problems”, *IEEE* 2008

[165] T. Drugman, G. Wilfart, and T.Dutiot,“ A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis,” In *Proc. of Interspeech,* Brighton, September 2009.

[166] Raitio, T.,Suni, H.Pullakka ,M.Vainio, and P.Alku, “ HMM based Finnish text –to- speech synthesizer using post glottal filtering”, In *Proc. of Interspeech*, Brisbane , 2008.

[167]J.Cabral, S. Renals, K.Richmond , and J. Yamagishi, “Glottal spectral separation for parametric speech synthesis ,” In *Proc. of the 7th SSW*, Japan, September 2010.

[168]G.Fant, J. liljencrants, and Q.Lin, “A four-parameter model of glottal flow”, *STL-QPSR, KTH*, Stockholm, 1985

[169] Jo˜ao P. Cabral, Renals S., Richmond K., Yamagashi J., “An HMM-based speech synthesizer using Glottal Post-Filtering” *IEEE* 2011

[170]D. Talkin, “A robust algorithm for pitch tracking (RAPT),” in *Speech Coding and Synthesis.*

[171]M. Ostendorf, P. Price, S. Shattuck-Hufnagel, “Technical Report ECS-95-001”, *The Boston University Radio News Corpus*, 1996

[172]W. Fisher, D. Doddington, K. Goudie-Marshall, “The DARPA speech recognition research database: specifications and status”, 1986

[173]University of Edinburgh, Center for Speech Technology Research, CSTR USKED TIMIT, 2002, http://festvox.org/dbs/dbs_kdt.html

[174]Carnegie Mellon University,“ The CMU pronunciation dictionary”, 2000,http://www.speech.cs.cmu.edu.

175]Furtado X A & Sen A, “Synthesis of unlimited speech in Indian Languages using formant-based rules”*’ Sadhana,19*96,pp 345-362

[176]Agrawal S S & Stevens K, “Towards synthesis of Hindi consonants using KLSYN88”, *Proc ICSLP92*, Canada, 1992, pp.177-180

[177]Dan T K, Datta A K & Mukherjee, B, “Speech synthesis using signal concatenation”, *J ASI*, vol. XVIII (3&4), 1995, pp 141-145

[178] Kishore S. P., Kumar R & Sanghal R, “A data driven synthesis approach for Indian language using syllable as basic unit”, *Proc ICON* 2002, Mumbai, 2002

[179]Agrawal S. S. 2010, “Recent Developments in Speech Corpora in Indian Languages: Country Report of India”, O-COCOSDA, Nepal.