A Comparative Study of Different Text-to-Speech Synthesis Techniques

Helal Uddin Mullah[1]
Speech & Image Processing Laboratory, Department of ECE, NEHU, Shillong - 793022
Email: mullahz251@gmail.com

Abstract— Speech synthesis is the artificial production of human speech. Attempts to control the quality of voice of synthesized speech have existed for more than a decade now. Several prototypes and fully operating systems also have been built based on different synthesis technique. This article reviews recent advances in research and development 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 speech synthesis techniques used.

Keywords- Text-to-speech(TTS), concatenative synthesis, database, hidden Markov model(HMM), feature extraction.

I. 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 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 that contains phoneme symbols, their diacritical marks and other symbols related to their pronunciation. Other phonetic alphabets such as SAMPA (Speech Assessment Methods-Phonetic Alphabet), World bet and Arpabet are available. The fig. 1 shows a block diagram of TTS synthesis [1].

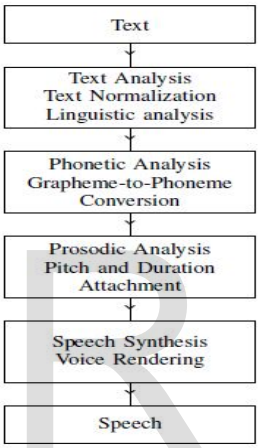

Fig. 1: *A simple functional block diagram of TTS system*

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 and 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.

TTS systems have numerous potential applications. Few are listed below.

1) In telecommunication service: Most of the calls required very less connectivity, TTS systems shows huge presence in

telecommunication services by making it possible to access textual information over the phone.

2) *In e-governance service*: TTS can be very helpful by providing government policy information over the phone, polling center information, land records information, application tracking and monitoring etc.

3) *Aid to disabilities*: TTS 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) *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.

5) *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.

6) *Multimedia, man-machine communication***:** Multimedia is first but promising move in the direction and it includes talking books and toys, mail, voice browsing 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 TTS, 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 II gives overview of various existing synthesis approaches and techniques with underlying assumptions. Section III presents an overview of HMM based speech synthesis and a description of implementation of statistical models for TTS is presented also discussing their advantages and disadvantages. Section IV includes a comparison table of unit selection and HMM based synthesis and section V describes about application of HMM based approach in multilingual TTS design. In section VI, we conclude the study and give suggestions for future work in this field of research.

II. 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.

A. 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[1] cites lack of knowledge of the complex human articulation organs being the main reasons why articulatory synthesis has not lead to quality speech synthesis.

B. 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 [7] 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. At CEERI, PC version of the Klatt TTS 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.

C. 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 subword 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.

1) Diphone concatenation synthesis :

Attempts to build utterances from phoneme wave forms have been of limited success, due to co-articulation 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 co-articulation 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) [8].

Linear predictive coding (LPC) method is used for diphone's concatenation. 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.

2) 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 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.

3) Preparation of database for corpus based TTS :

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. Several works have focused on automatic phonetic labeling, such as in van Santen et al. 1990 [9] broad- band and narrow-band edge detection has been adopted. Bonafonte et al. [10] took Gaussian probability density distribution as a similarity measure. In Torre Toledano et al. 1998[11], tried to mimic human labeling using set of fuzzy rules using rule-based approach. In Mporas et al. [12] 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 multi-speaker database.

4) Unit selection synthesis :

One of the major approaches in corpus-based speech synthesis is sample based one; can offer high quality synthesis without the expert work that would be required to build a formant synthesizer. Here, database must be appropriately designed to have the right coverage for the language or domain so that quality is reasonable. A database with unit 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. 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. S. Sakai and H. Shu 2005 [13], Z.-H. Ling R. 2006 [14] 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.

Today, with the increase in power and resources of computer technology and also the increase in speech and linguistics resources, larger speech databases can be collected and more appropriate speech units are selected that match both phonemes and other linguistic contexts such as lexical stress, pitch accent, and part-of-speech information in order to generate high quality natural sounding synthetic speech with appropriate prosody using the unit selection based technique.

III. STATISTICAL PARAMETER BASED SPEECH SYNTHESIS

We talk about a statistical parametric approach to speech synthesis particularly when we wish to learn speech models from data. The model is parametric because it describes the speech using parameters, rather than stored exemplars. It is statistical because it describes those parameters using statistics (e.g., means and variances of probability density functions) which capture the distribution of parameter values found in the training data [15].

The employment of hidden Markov models (HMM) in speech synthesis began after the success of the HMM for automatic speech recognition. HMM-based model is not a true representation of real speech but the availability of effective learning algorithms, automatic methods for model complexity control and computationally efficient search algorithms make the HMM a powerful model.

A. *Architecture of the system for HMM based synthesis*

The system used for training speech models and generating speech waveforms is HTS (HMM-based Speech Synthesis System). It consists of two main parts training and synthesis. During the training spectral coefficients (mel-cepstral coefficients [7] and their dynamic features) and excitation (logarithmic fundamental frequency (log F0) and its dynamic features) parameters are extracted from speech database and modeled by context-dependent HMM-s (phonetic, phonological and prosodic contexts are taken into account) [21]. Each HMM has state duration probability density functions (PDFs) to capture temporal structure of speech [16]. As a result, the system models spectrum, excitation and durations in a unified framework [17].

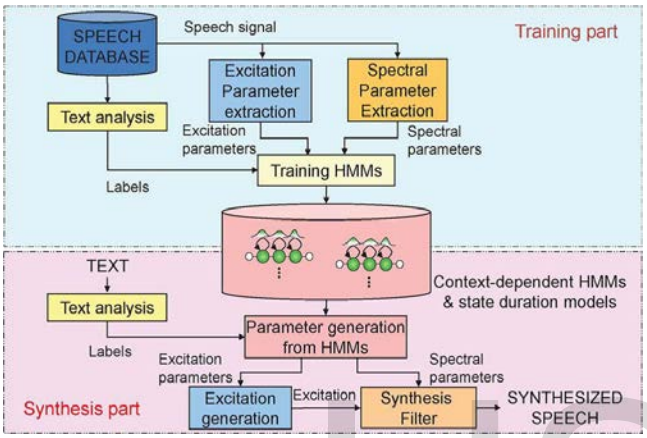

Fig. 2: *Architecture of HMM based synthesis system*

During the synthesis part, text to be synthesized is converted to a context-dependent label sequence and then an utterance HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. State durations of the utterance HMM are determined based on the state duration PDFs. Speech parameter generation algorithm generates the sequence of spectral and excitation parameters that maximize their output probabilities. Finally, a speech waveform is synthesized directly from the generated spectral and excitation parameters using the speech synthesis filter [18].

B. *Advantages of Statistical Parametric Speech Synthesis*

Most of the advantages of statistical parametric speech synthesis (against unit-selection synthesis) are related to its flexibility due to the statistical modeling process. Although the training process takes long to complete (up to tens of hours regarding the quantity of training data), it is of little relevancy because it happens only once. The footprint of speech model and synthesis engine is small and is suitable to use in devices with low computational performance.

Intelligible speech can be synthesized with models trained on small amount of data (as little as 100 sentences [20]) because HMM-based speech model is stable and can cover acoustic space despite sparseness of training data.

The main advantage of this approach is its flexibility in changing voice characteristics, speaking styles and emotions.

Since the system requires only language specific database and contextual factors to work, it is rather easily adapted to a language.

C. *Drawbacks of Statistical Parametric Speech Synthesis*

Compared to unit selection speech synthesis the biggest drawback of HMM-based speech synthesis is the lower quality of speech. There seem to be three factors that degrade quality, i.e., vocoder (analysis-synthesis system that reproduces speech [20]), acoustic modeling accuracy and over-smoothing [21].

The speech synthesized by the basic HMM-based speech synthesis system sounds buzzy since it uses a mel-cepstral vocoder with simple periodic pulse-train or white noise excitation. To alleviate the problem, high quality vocoders have been integrated that, for example, take into account the aperiodicity of fundamental frequency. Speaking of Estonian, the letters b, d and g (short stops) that appear between voiced letters tend to sound voiced rather than unvoiced so they cannot be synthesized correctly by a simple vocoder.

HMMs are useful for describing transitions between states but regarding one specific state the parameters are static. State-output probability depends on the current state and the probability factor for duration decreases exponentially. This does not hold for real speech. Extra models for describing duration have been employed to overcome this problem.

The statistical averaging in the modeling process improves robustness against data sparseness and the use of dynamic feature constraints in the synthesis process enables the system to generate smooth trajectories. Compared with natural speech the synthesized speech sounds muffled because the generated speech-parameter trajectories are over-smoothed, thus discarding the natural variability of speech.

IV. COMPARISION OF UNIT SELECTION AND HMM BASED SPEECH SYNTHESIS SYSTEM

Unit Selection Based	**HMM Based**
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

TABLE I: *Relation between unit selection and generation (HMM) approaches.*

V. 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 [20] 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. [23], 2006 and Black. A and Schultz et al. [24], 2006 proposed building such multilingual synthesizers using combined data from multiple languages.

VI. 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, TTS 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 and 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

[1] 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.

[2] T. Dutoit, "An introduction to text-to-speech synthesis", 2nd edition, Kluwer Academic Publishers, 1997.

[3] Daniel Jurafsky and James H. Martin, "Speech and Language Processing", 2nd edition, Prentice Hall, 2009.

[4] H.Zen, K.Tokuda and A.W Black, "Statistical parametric speech synthesis", speech communication, doi:10.1016/j.specom.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] D. Klatt, "Software for a cascade/parallel formant synthesizer", Journal of the Acoustical Society of America, vol. 67, pp. 971-995, 1980.

[8] S.Lemmetty, "Review of Speech Synthesis Technology", Masters Thesis, Helinski University of Technology.

[9] Van Santen, J. P. H. and R. Sproat, "High-accuracy automatic segmentation", Proceedings of European Conference on Speech Communication and Technology, 1990, pp.28092812.

[10] 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.

[11] 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.

[12] I. Mporas, T. Ganchev and N. Fakotakis, "A hybrid architecture for automatic segmentation of speech waveforms", Proceedings of IEEE ICASSP08, PP. 4457-4460, 2008.

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

[14] Z.-H. Ling and R.-H. Wang, "HMM-based unit selection using frame sized speech segments", In Proc. Interspeech (ICSLP), pages 20342037, 2006.

[15] S. King, "An introduction to statistical parametric speech synthesis", Sadhana, 36(5) (2011), 837-852.

[16] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Duration Modeling in HMM-based Speech Synthesis System", Proceedings of ICSLP 2 (1998), 29-32

[17] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis" Proceedings of EUROSPEECH 5 (1999), 2347-2350.

[18] S. Imai, "Cepstral analysis synthesis on the mel-frequency scale", Proceedings of ICASSP-83 (1983), 93-96.
[19] Y. Takamido, K. Tokuda, T. Kitamura, T. Masuko, T. Kobayashi, "A study of relation between speech quality and amount of training data in HMM based TTS system", ASJ Spring meeting 2002, 291-292 (in Japanese).
[20] Vocoder, http://en.wikipedia.org/wiki/Vocoder
[21] H. Zen, K. Tokuda, A.W. Black, "*Statistical parametric speech synthesis*", Speech Commun.51 (11) (2009), 1039-1064.
[22] Latorre, J., Iwano, K., Furui S., "*New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer*", 2006, Speech Communication ICAT. 48 (10), 12271242.
[23] Black, A., Schultz, T., 2006, "*Speaker clustering for multi-lingual Synthesis*", In Proc. ISCA itrw multiling. No. 024.