International Journal of Computer Applications Technology and Research
Volume 6–Issue 12, 511-515, 2017, ISSN:-2319–8656

A survey on speech signal synthesis system

Soumya Dath G
Assistant Professor, Dept. Of ISE
GSSSIETW, Mysuru, Karnataka, India

Abstract: The primary objective of this paper is to provide an overview of existing methods Text-To-Speech synthesis techniques. Text to speech synthesis can be broadly categorized into three categories, formant Based, Concatenative based and Articulatory. Formant based speech synthesis relies on different techniques such as cascade, parallel, klatt and PARCAS Model etc. Concatenative speech synthesis can be broadly categorized into three categories, Diphones Based, Corpus based and Hybrid whereas Articulatory synthesis involves Vocal Tract Models, Acoustic Models, Glottis Models , Noise Source Models . In this paper, all text to speech synthesis methods are explained with their pros and cones.

Keywords: Text to speech synthesis, Formant speech synthesis, Concatenative speech synthesis, Articulatory speech synthesis

1. INTRODUCTION

Text-to-speech (TTS) synthesis ultimate goal is to create natural sounding speech from arbitrary text. Moreover, the current trend in TTS research calls for systems that enable production of speech in different speaking styles with different speaker characteristics and even emotions. Speech synthesis generally refers to the artificial generation of human voice – either in the form of speech or in other forms such as a song. The computer system used for speech synthesis is known as a speech synthesizer. There are several types of speech synthesizers (both hardware based and software based) with different underlying technologies. For example, a TTS (Text to Speech) system converts normal language text into human speech, while there are other systems that can convert phonetic transcriptions into speech. The goal of a text-to-speech system is to automatically produce speech output from new, arbitrary sentences. The text-to-speech synthesis procedure consists of two main phrases. The first is text analysis, in which the input text is transcribed into a phonetic or some other appropriate representation, and the second is the actual generation of speech waveforms, in which the acoustic output is produced from the information obtained from the first phase [2]. A simplified version of the synthesis procedure is presented in figure 1.

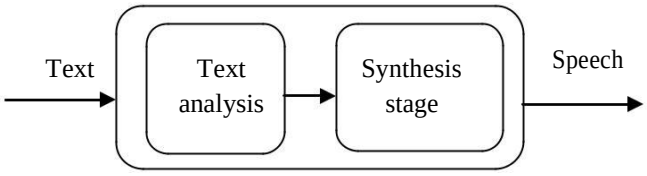

Figure 1: Phases of Text -to -speech system

The quality of a speech synthesizer is measured based on two primary factors – its similarity to normal human speech (naturalness) and its intelligibility (ease of understanding by the listener). Ideally, a speech synthesizer should be both natural and intelligible, and speech synthesis systems always attempt to maximize both characteristics [3].

A typical text to speech system has two parts – a front end and a back end. The front end is responsible for text normalization (the pre-processing part) and text to phoneme conversion. Text normalization or tokenization is the phase where numbers and abbreviations in the raw text are converted into written words.

Text to phoneme conversion or grapheme-to-phoneme conversion is the process of assigning phonetic transcriptions to each word and dividing them into prosodic units such as phrases, clauses, and sentences. The output of the front-end system is the symbolic linguistic representation of the text. It is composed of the phonetic transcriptions along with the prosody information. This output is then passed on to the back-end system or the synthesizer, which converts it into sound [1].

This paper is organized as follows. This section gives an introduction about text to speech synthesis. In section II, a review about various methods for text to speech synthesis is explained in detail. The conclusion is given in section III.

2. METHODS OF TEXT TO SPEECH SYNTHESIS

Various methods of text to speech synthesis are explained below.

2.1 Formant Synthesis

Formant synthesis is based on the source-filter-model of speech. There are two basic structures: parallel and cascade, but for better performance some kind of combination of these is usually used. Formant synthesis also provides infinite number of sounds which makes it more flexible than concatenation methods. In this approach, at least three formants are generally required to produce intelligible speech and to produce high quality speech up to five formants are used. Each formant is modelled with a two-pole resonator which enables both the formant frequency (pole-pair frequency) and its bandwidth to be specified. The input parameters may be the open quotient that means the ratio of the open-glottis time to the total period duration: Voicing fundamental frequency (F0), Voiced excitation open quotient (OQ) , Degree of voicing in excitation (VO), Formant frequencies and amplitudes (F1...F3 and A1...A3), Frequency of an additional low-frequency resonator (FN) and Intensity of low and high-frequency region (ALF, AHF)[1].

2.1.1 A Cascade Formant Synthesizer:

It consists of band-pass resonators connected in series and the output of each formant resonator is applied to the input of the next one. The cascade structure needs only formant frequencies as control information. The main advantage of the

cascade structure is that the relative formant amplitudes for vowels do not need individual controls.

The cascade structure is better for non-nasal voiced sounds because it needs less control information than parallel structure. Moreover, it is then simple to implement. However, with cascade model the generation of fricatives and plosive bursts is a problem [1].

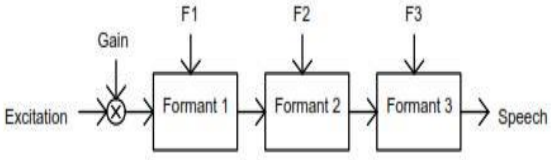

Fig. 2 Basic structure of cascade formant synthesizer [1]

2.1.2 A Parallel Formant Synthesizer:

It consists of resonators connected in parallel. Sometimes extra resonators for nasals are used. In this method, the excitation signal is applied to all formants simultaneously and their outputs are summed. Adjacent outputs of formant resonators must be summed in opposite phase to avoid unwanted zeros or anti resonances in the frequency response. The parallel structure enables controlling of bandwidth and gains for each formant individually and thus needs more control information [1].

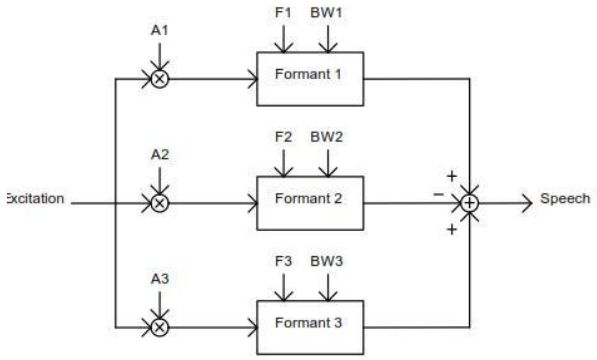

Fig. 3 Basic structure of a parallel formant synthesizer [1]

The parallel structure has been found to be better for nasals, fricatives, and stop consonants, but some vowels cannot be modelled with parallel formant synthesizer as well as with the cascade one, this is the main disadvantage of it.

2.1.3 Klatt Formant Synthesizer:

To see the good results with only either cascade or parallel formant synthesiser method is difficult. So to improve results, combination of these basic models is used. In 1980 Dennis Klatt proposed a more complex formant synthesizer which incorporated both the cascade and parallel synthesizers with additional resonances and anti-resonances for nasalized sounds, sixth formant for high frequency noise, a bypass path to give a flat transfer function, and a radiation characteristics. This system used quite complex excitation model which was controlled by 39 parameters updated every 5 ms. The quality of Klatt Formant Synthesizer was very promising and this model has been incorporated into several present TTS systems, such as MITalk, DECtalk, Prose-2000, and Klattalk[1].

2.1.4 PARCAS (Parallel-Cascade) model:

In the model, the transfer function of the uniform vocal tract is modelled with two partial transfer functions, each including every second formant of the transfer function. Coefficients k1, k2, and k3 are constant and chosen to balance the formant amplitudes in the neutral vowel to keep the gains of parallel branches constant for all sounds[1].

The PARCAS model uses a total of 16 control parameters: F0 and A0 - fundamental frequency and amplitude of voiced component, Fn and Qn - formant frequencies and Q-values (formant frequency / bandwidth), VL and VH - voiced component amplitude, low and high, FL and FH - unvoiced component amplitude, low and high, QN - Q-value of the nasal formant at 250 Hz[1].

The used excitation signal in formant synthesis consists of some kind of voiced source or white noise. The correct and carefully selected excitation is important especially when good controlling of speech characteristics is wanted. The formant filters represent only the resonances of the vocal tract, so additional provision is needed for the effects of the shape of the glottal waveform and the radiation characteristics of the mouth. Usually the glottal waveform is approximated simply with -12dB/octave filter and radiation characteristics with simple +6dB/octave filter[1].

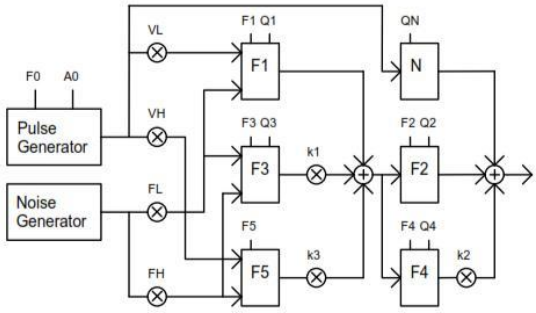

Fig. 4 PARCAS model [1]

2.2 Concatenative Speech Synthesis

Concatenative speech synthesis method involves production of artificial speech by concatenating pre recorded units of speech by phonemes, diphones, syllabus, words or sentences. This method involve the selection of appropriate units from the speech database and algorithms that join selected units and perform some signal processing to smoothen the concatenation the boundary. Here, speech is produced by selecting and concatenating appropriate speech unit from speech database where speech data base consists of speech units of different sizes such as phones, di-phones, syllables, words/sentences. Concatenative speech synthesis can be done by three different methods [2].

2.2.1 Di phone based speech sythesis:

Diphone is used as a basic speech unit for this synthesis method where diphone is two connected half phones starting in the middle of first phone and ending in the middle of second phone.

As only single instances of all speech units are available in the speech database, to obain good quality of synthesiesd speech with the procody, various signal processing methods are applied. PSOLA, TD – PSOLA, LP – PSOLA, ESNOLA, FD-

PSOLA are various signal processing methods are used for obtaining good quality synthesised speech.

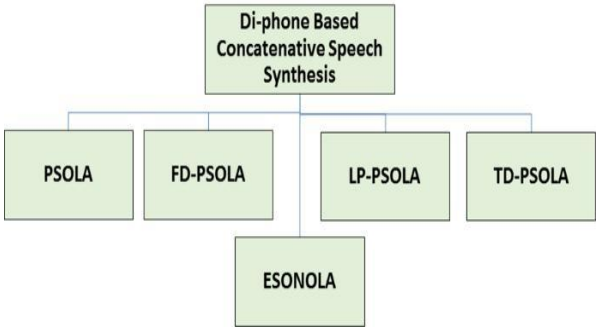

Fig 5. Various Signal Processing Techniques Used in Di-phone Based Speech Synthesis

2.2.1.1 Pitch synchronous Overlap Add(PSOLA):

This method is analysis synthesis method, that involves decomposition of speech signal into number of pitch short synchronised waveforms.To obtain multiple synthesis, short term signal the pitch synchronous short term waveform can be altered time or spectral domain and final synbthesised speech is produced by overlapped addition of this short term signal.It preserves the spectral envelope when pitch shifting is used. This is the main advantage of this method. In addition it does not lose any information of signal, since it work directly on the signal.

2.2.1.2 Frequency Domain Pitch Synchronous Overlap Add (FD-PSOLA):

This method is used for modifying the spectral envelope and for compute spectral envelope it uses linear prediction and pitch is modified by linear interpolation on the spectral envelope.

In this method, all operations are performed in frequency domain and main advantage of this system is easy implementation. FDPSOLA leads discontinuities at the concatenation boundary. Moreover, it is computationally intensive and has high memory requirements for storage which are the main disadvantages of systems.

2.2.1.3 Linear Prediction Pitch synchronous overlap Add (LP-PSOLA):

By manipulating the LP residual, modification of pitch and duration of speech is achieved. This method is suitable for pitch scale modification as it provides independent control over spectral envelop for synthesis.

Due to overlap and adding of the windowed residual segments, it producing phase mismatch and audible distortion, this is main disadvantage of this method.

2.2.1.4 Time Domain Pitch Synchronized Overlap Add (TD-PSOLA):

This method is used for the modification of the prosody of speech waveforms, as it facilitates the high quality pitch and time scale modification. This method is computationally efficient method but it has a drawback of large speech database requirement. Moreover, the quality of synthesized speech is affected by the detection of epochs in the speech signal which is very difficult to achieve in in real time applications.

2.2.1.5 Epoch Synchronous Non Overlap Add (ESNOLA):

In this method, the synthesized speech is generated by concatenating basic speech segments at the epoch positions of the voiced speech, where epochs represent quasi periodic sounds. Pitch and duration modifications of speech synthesized speech can be done through ESNOLA.

Main advantage of this method is that, it allows selection of smaller parts of phonemes called partnemes as concatenation units that reduces the size of the speech inventory. ESNOLA supports the introduction of jitter shimmer and complexity perturbations that leads to naturalness in phonetic quality of synthesized speech.

2.2.2 Corpus based speech synthesis

This method uses data driven approach. This method depends on the availability of good speech inventory, having good phonetic and prosodic features for the language under consideration.

Segmentation and labelling of the speech inventory is major issue related to corpus based approach that can be achieved using automatic segmentation algorithm.

2.2.2.1 Unit Selection Synthesis:

In this method multiple instance of speech units having different prosodic features are stored. Here, unit is selected from the database based on two costs: A target cost and a concatenation cost.

Target Cost:

It estimates how similar the features of database speech unit are to the features of the desired speech unit..It comprises of target sub-costs where each target sub cost is a cost of a single attribute of a speech unit such as energy, pitch etc.

The target cost can be calculated as: $Ct(ti,vi) = \sum_{pj=1} wtj\ Ctj(ti,vi)$

Where, ti is the target unit , Vi is the candidate unit and p is the number of sub-costs used. Ctj is the jth target sub-cost, Wtj it is the weight given to the jth target sub-cost.

Concatenation Cost:

It is a measure of how well two speech units join and match each other when they are concatenated . The concatenation cost also comprises of multiple subcosts where each of these sub-costs is related to a specific continuity metric such as spectral continuity etc.

The concatenation cost can be calculated as: $Cc(vi-1,vi) = \sum_{qj=1} wcj\ Ccj(vi-1,vi)$

Where vi-1 and vi are candidate speech units for the (i-1)th and ith target speech units, q is the total number of subcosts used and Wcj is the weight associated with the subcost Ccj .

An exhaustive search is performed to select optimum speech units from the speech database.

This method is sample based method and main advantage of this system is that it requires minimum or no signal processing and it gives Good quality in naturalness of synthesized speech. But Requires the main drawback of this system is that it requires a large speech database.

2.2.2.2 Statistical Parametric Synthesis:

Statistical parametric synthesis makes use of averaged acoustic inventories that are extracted from the speech corpus. The extracted parameters of speech are the spectral

parameters such are cepstral coefficients or line spectral pairs, and excitation parameters such as fundamental frequency.

Statistical Parametric synthesis has the advantages of requiring less memory to store the parameters of the model, rather than the data itself and it allows more Variation in the speech produced for example, an original voice can be converted into another voice. The most commonly used statistical parametric speech synthesis technique is the Hidden Markov Model (HMM) synthesis [4].

2.2.2.3 Hidden Markov Model (HMM) synthesis:

HMM synthesis has two phases: A training phase and a synthesis phase.

In the training phase speech parameters are extracted from utterances in the speech training database and they are modeled as HMMs and In the synthesis phase the words to be synthesized their corresponding HMMs are identified from the database and parameters are extracted from these HMMS. Finally speech is synthesized from these extracted parameters.

This method is based on parameters. Main advantage of HMM based parametric speech synthesis is the flexibility since speech is stored in the form of parameters and it is easy to modify these parameters. Apart from it, it requires small speech inventory but it has the disadvantage of poor quality in the naturalness of the synthesized speech due to over smoothing of the parameters in the statistical model and required more Signal processing [3].

2.2.3 Hybrid Text To Speech Synthesis:

The Hybrid TTS approach is a combination of the two main approaches of synthesis: Concatenative synthesis and Statistical Synthesis. The hybrid TTS combines the characteristics of smooth transitions between adjacent speech segments of a Statistical TTS with the naturalness of a Concatenative TTS. This is achieved by interweaving natural speech segments and statistically generated speech segments. The statistical segments are positioned so as to smooth discontinuities in the synthesized speech, while enabling as far as possible natural speech sequences as they appear in the training inventory disadvantages of this system are the Degradation in speech quality when CTTS speech inventory is small and more signal processing requirement [5].

2.3 Articulatory Synthesis

Articulatory synthesis refers to computational techniques for synthesizing speech based on models of the human vocal tract and the articulation processes occurring there. “Articulatory speech synthesis models the natural speech production process as accurately as possible. This is accomplished by creating a synthetic model of human physiology and making it speak. [6].Articulatory synthesis systems comprise (i) a module for the generation of vocal tract movements (control model), (ii) a module for converting this movement information into a continuous succession of vocal tract geometries (vocal tract model), and (iii) a module for the generation of acoustic signals on the basis of this articulatory information (acoustic model).[7]

2.3.1 Vocal Tract Models:

The task of vocal tract models is to generate the complete geometrical information concerning the vocal tract (shape and position of all vocal tract organs, i.e. lips, tongue, palate, velum, pharynx, larynx, nasal cavity) and its variation over time. Shape, position, and motion of movable vocal tract organs are generated on the basis of the time functions of all vocal tract parameters defined by the model [7]. A typical set of vocal tract parameters are: position of jaw, upper lips, lower lips, tongue tip, tongue body, velum, and larynx [6]. Vocal tract models can be subdivided into statistical, biomechanical, and geometrical models [7].

Statistical models are based on large corpora of vocal tract movements measured by different techniques (MRI, EMA, or X-Ray [7]. Biomechanical models aim to model the physiological basis of all vocal tract organs and their neuromuscular control [7]. For geometrical models the positioning and shape of the vocal tract organs is calculated by using a set of a priori defined vocal tract parameters.

2.3.2 Acoustic Models:

The task of the acoustic models is to calculate the time varying air flow and air pressure distribution within the vocal tract and to calculate the acoustic speech signal radiated from the facial region of the model[8]. The input information for acoustic models is lung pressure, subglottal air flow, and the geometric shape of the vocal tract tube (trachea, glottis, pharyngeal, oral, and nasal tract) for each time instant. A timevarying tube model is specified from the geometrical vocal tract model information, which represents the vocal tract cavities (trachea, pharynx, nasal, and oral cavity) [7]. Acoustic models can be subdivided into reflection type line analog models, transmission line circuit analog models, hybrid time-frequency domain models, and finite element wave propagation models

In the case of reflection type line analog models forward and backward traveling partial flow or pressure waves are calculated for each vocal tract tube section in the time domain on the basis of scattering equations which reflect the impedance discontinuity at tube junctions [12]. In the case of transmission line circuit analog models, pressure and flow within each vocal tract tube section is calculated by a digital simulation of electrical circuit elements, representing the acoustic and aerodynamic properties within each vocal tract tube section.

2.3.3 Glottis Models:

The task of glottis models is to generate the acoustic source signal for phonation and its insertion into the vocal tract tube model[7]. The source signal is propagated through the supraglottal cavities (pharyngeal, oral and nasal cavity) as well as through the subglottal cavities (trachea, lungs) by the acoustic model. Glottis models can be subdivided into self-oscillating models, parametric glottal area models, and parametric glottal flow models [11].

2.3.4 Noise Source Models:

The task of noise source models is to generate and to insert noise source signals into the acoustic transmission line model. Noise signals result from turbulent air flow, mainly occurring downstream in front of a vocal tract constriction in the case of a high value of volume flow. Noise source models can be subdivided into parametric and generic noise source models [9][10].

3. CONCLUSION

Speech synthesis has been developed steadily over the last decades and it has been incorporated into several new applications. Text to speech synthesis is a rapidly growing aspect of computer technology and is increasingly playing a more important role in the way we interact with the system and interfaces across a variety of platforms. We have identified the various operations and processes involved in text to speech synthesis.

For most applications, the intelligibility and comprehensibility of synthetic speech have reached the acceptable level. However, in prosodic, text preprocessing, and pronunciation fields there is still much work and improvements to be done to achieve more natural sounding speech. The three basic methods used in speech synthesis have been introduced in Chapter 2. The most commonly used techniques in present systems are based on formant and Concatenative synthesis. The latter one is becoming more and more popular since the methods to minimize the problems with the discontinuity effects in concatenation points are becoming more effective. The Concatenative method provides more natural and individual sounding speech, but the quality with some consonants may vary considerably and the controlling of pitch and duration may be in some cases difficult, especially with longer units. With concatenation methods the collecting and labeling of speech samples have usually been difficult and very time-consuming. With formant synthesis the quality of synthetic speech is more constant, but the speech sounds slightly more unnatural and individual sounding speech is more difficult to achieve. Formant synthesis is also more flexible and allows a good control of fundamental frequency [1]. The third basic method, the Articulatory synthesis, is perhaps the most feasible in theory especially for stop consonants because it models the human articulation system directly. On the one hand, the Articulatory based methods are usually rather complex and the computational load is high [7][8].

4. REFERENCES

[1] Sami Lemmetty. Review of Speech Synthesis Technology. Helsinki University of Technology Department of Electrical and Communications Engineering. March 30, 1999.

[2] Rubeena A. Khan , J. S. Chitode, Concatenative Speech Synthesis: A Review, *International Journal of Computer Applications (0975 – 8887). Volume 136 – No.3, February 2016.pg-1 to 4.*

[3] Raitio, Tuomo, et al. "HMM-based speech synthesis utilizing glottal inverse filtering." Audio, Speech, and Language Processing, IEEE Transactions on vol.19, no.1, 2011, pp. 153-165.

[4] Heiga Zen, Keiichi Tokuda, Alan W. Black ,"Statistical parametric speech synthesis", Speech Communication vol.51,no.11,2009,pp. 1039–1064.

[5] Pertti Palo. A Review of Articulatory Speech Synthesis. Espoo, June 5, 2006

[6] Birkholz P, Martin L, Willmes K, Kröger BJ, Neuschaefer-Rube C (2015) The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study. Journal of the Acoustical Society of America 137:1503-1512

[7] Louis Goldstein and Carol A. Fowler. Articulatory Phonology: A phonology for public language use

[8] Richard S, Mc gowan and Alice Faber. Introduction to papers on speech recognition and perception from an articulatory point of view.

[9] Shuangyu Chang. A Syllable, Articulatory-F eature, and Stress-Accent Model of Speech Recognition. September 2002

[10] Kelly and Lochbaum 1962, Liljencrants 1985, Meyer et al. 1989, Kröger 1998.(e.g. Flanagan 1975, Maeda 1982, Birkholz et al. 2007.