Survey on Text to Speech Synthesis Models and Methods

Pooja A.Gundle, R.K. Chavan

Abstract—TTS provide an overview of existing text to speech techniques by highlighting there recent works, models, challenges and advanced techniques. The quality of text to speech synthesis is like natural sounding. Hence today's interest is in high quality speech for applications. There are the many challenges are discussed in this paper. It has focus on text to speech synthesis models. Such as comman form model, signal to signal model, pipeline model and grapheme and phoneme form model, segmentation model.Also provide the overview of speech synthesis methods, is broadly categories into three types. Formant based, Concatenative based and Articulatory based, HMM based synthesis. Formant synthesis has three different techniques cascade, parallel, klatt formant synthesizer. Concatenative speech synthesis broadly categorize into three types diaphones Based, domain based and unit selection based synthesis, articulatory synthesis have Vocal Tract Models, Acoustic Models, Glottis Models , Noise Source Models.

Index Terms— TTS, TTS models, speech synthesis, TTS methods, HMM (Hidden Markov model).

1 INTRODUCTION

The text to speech synthesis is Synthesizing speech from the text (TTS) [1].TTS systems has based on the complex pipeline. TTS is the one of the major application of NLP. The conversation of text to speech involved three important stages text analysis, text processing and waveform generation i.e. formation of speech. TTS is an application that convert written text into speech, user enter a text and gets output as a sound. In this survey Speech Synthesis is becoming one of the most important steps towards improving the human interface to the computer. The objective of text to speech is convert arbitrary text into spoken waveform. The quality of speech synthesizer is based on two factors naturalness and intelligibility [2]. Intelligibility means it describes the clarity of audio, naturalness that describes the information which is not directly captured by intelligibility. Nowadays TTS system useful in many application there are many traditional speech synthesis approaches. This paper explains detail information about formant synthesis and its classification.

1.1 Text-To-Speech synthesis:

The main motto of TTS is to convert arbitrary text into waveform. Speech generation is generation of an acoustic wave form corresponding text and each of these units in the sequence. This involved text analysis, text normalization, text processing, grapheme-to phoneme conversion and speech synthesis [1]. Text analysis which analyse the input text such as dividing the text into words and sentences. Text normalization is the transformation of text into the pronounceable form, It is the front end of TTS that assign phonetic transcription to each and every word [3]-[4]. The process of assigning phonetic transcription to word is called grapheme to phoneme conversion. The phonetic analysis also known as word analysis focuses on phone within the word [5].finally symbolic linguistic representation produce sound. Fig1 shows the flow diagram of TTS.

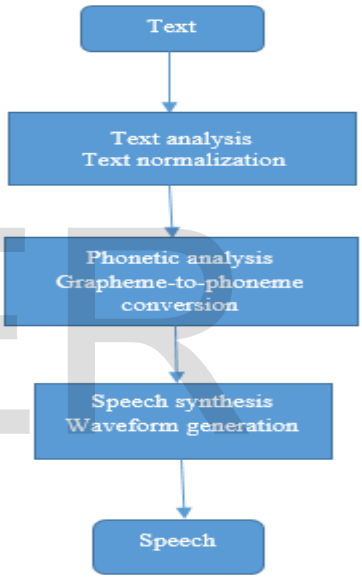

Fig 1.1 Block diagram of speech synthesis

1.2Goals and the challenges of text to speech synthesis

Speech synthesis has developed for over the various systems, it has been integrated into several applications. Developing speech synthesis is a complicated process. There are many goals in TTS, the main goal of TTS is providing the high quality speech to users.

1) The development of computer based multi voice TTS system.

2) Developing of TTS system requires the knowledge about the language and produce human like speech.

3) The goals in building a computer system capable of speaking are to first build a system that clearly gets across the message.

4) The system is able to take any written input, if we build an English text-to-speech system, it should be capable of reading any English sentence given to it.

5) TTS system has made a good testing ground for many

models and theories.

This paper is organised as follows: This section gives introduction about TTS, how text is converted into speech, what are the factors which important to the speech synthesis. Section II illustrates the work that researcher carried in speech synthesis Section III illustrates the model of TTS, and in Section IV gives methods of TTS.

2. Related work

There were lot of research done in text to speech synthesis. Previous work uses neural network as substitutes for TTS components. Paul Taylor explains the text to speech concepts how text is converted into speech [1].Different methods of speech synthesis used in development of speech synthesis [10]. The first device to be considered as speech synthesizer was vocoder.Vocoder is key component of modern speech synthesis applications. It provides a parameterization of the speech waveform that is willing to statistical modelling [6].After demonstrating vocoder scientific world is more interested in speech synthesis.

A¨aron van den Oord [7] Wavenet paper on deep neural network. It is audio generative model based on PixelCNN. In wavenet, the audio data operates directly at waveform level. It develop new architecture based on dilated casual convolutions.Wavenet provide many application. Measuring the wavenet's audio modelling performance it evaluate on three different tasks. TTS, multi speaker speech generation, music audio modelling.

Recently lot of the work done in speech synthesis, wavenet [7], Char2wav [9].Char2wav extends sample RNN with attention based model which generates waveform samples. It produce speech directly from text.

Statistical parametric speech synthesis [2] it extracts parametric representation of speech. One of the instance of these technique called HMM based speech synthesis. HMM based speech synthesis is open source tool which provides a development platform for statistical parametric speech synthesis [11].

3. TTS system models

For our understanding how the text to speech conversion by computer carried out, we define common form model and several other models.

The Common Form model: In the common form model, there are two components, a text analysis system which decodes the text signal and uncovers the form, and a speech synthesis system which encodes this form as speech. The first system is one of resolving ambiguity from a noisy signal so as to find a clean, unambiguous message; the second system is one where we take this message and encode it as a different, noisy, ambiguous and redundant signal. In the basic common form model, we always read the words as they are encoded in the text; every word is read in the same order we encounter it.

The key features of the model are

- The two fundamental processes text analysis and speech synthesis
- The task of analysis is find the form that is words from the text.
- The task of synthesis is to generate the signal from this form.

- Signal to signal model: In this model the process of converting written signal to spoken signal. Here we directly convert text to speech. In such models, the process is not seen as one of uncovering a linguistic message from a written signal, and then synthesising from this, but as a process where we try and directly convert the text into speech. In particular, the system is not divided into explicit analysis and synthesis stages [1].
- Pipelined models: signal to signal model implemented as a pipeline model, the process is like the passing representation from one module to another. These type of systems are highly modular. Such that each module's job is defined as reading one type of information and producing another. Each module perform specific task such as speech tagging or pause insertion and so on. Modules are not explicitly linked that's why different theories and techniques are coexist.
- Grapheme and phoneme form model: This process is similar to comman form model in that first a grapheme form of the text input is found, and it is converted to a phoneme form for synthesis [1]. In this model grapheme form of the text input is -converted into phoneme form of speech synthesis [4] that is exact pronunciation of each word of the input sentences. Words are not central to the representation as is the case in the common form model. This approach is particularly attractive in languages where the grapheme-phoneme correspondence is relatively direct; in such languages finding the graphemes often means the phonemes and hence pronunciation can accurately be found. For other languages such as English, this is more difficult, and for languages such as Chinese this approach is probably impossible [12]. If it can be performed, a significant advantage of the grapheme/phoneme form model is that, an exhaustive knowledge of the words in a language is not necessary; little or no use is needed for a lexicon.

4. Text To Speech synthesis methods

The most differences usually are how the speech signal is generated from text. Among all the methods the easiest for understanding is concatenative speech synthesis. The audio signal is formed by concatenating pre-recorded speech samples. Another method that more closely related to articulatory speech synthesis is formant synthesis. There are several methods for synthesizing the speech. In this survey we explain three main speech synthesis method. Formant synthesis, articulatory speech synthesis, concatenative speech synthesis, HMM based speech synthesis.

1] Formant Synthesis

Formant synthesis was the first genuine synthesis technique and it was the dominant technique until the early 1980s.formant synthesis also called by synthesis by rule it produce quality speech which sounds the unnatural. Formant synthesis adopt model based, acoustic phonetic approach to the synthesis problem. The filter in a formant synthesizer is typically implemented using cascade or parallel second-order filter sections, one per formant [13]. Virtually in all formant synthesisers, the oral and nasal cavities are modelled as per parallel systems [13]. Most modern rule-based text-to-speech systems descended from software based on this type of synthesis model [14]-[16].

A formant synthesizer is a source-filter model in which the source models the glottal pulse train and the filter models the formant resonances of the vocal tract. Formant synthesizer is not an accurate model for vocal tract. Formant synthesis it is the combination of physical and spectral modelling approaches. In the Physical model there is an explicit division between glottal-flow wave generation and the resonance filter .It is a spectral modelling method in that parameters are estimated by matching short time audio spectra of the desired sounds [12]. 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 [13].The basic cascade format synthesis

(i) A Cascade Formant Synthesizer
(ii) A Parallel Formant Synthesizer
(iii) Klatt Formant Synthesizer

A cascade formant synthesizer Fig4.1 consist of band pass resonaters connected in series and the output of each formant synthesizer is applied to the input of the next one [10].A cascade formant synthesizer is good for non-nasal voice sounds rather than parallel formant synthesizer because it needs less control information

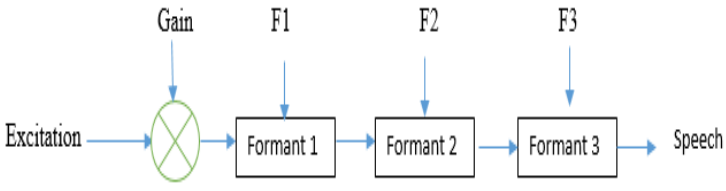

Fig 4.1 Basic structure of cascade formant synthesizer

In a parallel cascade synthesizer consist of resonaters that are connected in parallel sometimes extra resonaters are used for nasal. The parallel structure enables controlling of bandwidth and gain for each formant individually, and also need more control information. The excitation signal is applied to all formants, and output are summed. Adjacent output of formant synthesizer must be summed in opposite phase to avoid unwanted zeros.Parallel structure is better for nasal and stop consonants. But some vowels can't be model with parallel formant synthesizer it is well with the cascade formant synthesizer.

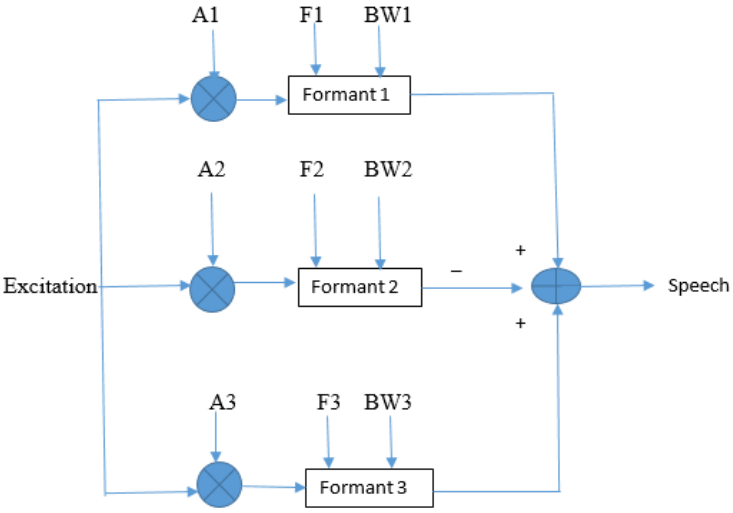

Fig. 4.2. Basic structure of a parallel formant synthesizer

Klatt formant synthesizer introduced by Dennis klatt. The qulity of klatt format synthesizer was very good for future views. The model has been incorporated into many TTS systems. Such as MLTALK, DECtalk, and Klattalk [24].Klatt formant synthesis is a synthesis technique, where set of parameters generated from text from which the wavform is builded from a cascade of modules to give the resultant signal.

2] Articulatory speech synthesis

The most obvious way to synthesise speech is to try direct simulation of human speech production. This approach is called as an articulatory speech synthesis. The machine was mechanical device with tubes. The device is of course mimicking vocal tract using sources and filters. Articulatory synthesis is the production of speech sounds using the model of vocal tract [20].Which directly simulates the movement of speech articulators [21]. Two difficulties that arise in articulatory synthesis is how to generate the control parameters form the specification and how to find the right balance between highly accurate model that closely follows human physiology[22].It produce complete synthetic output based on mathematical models of the structure.

i) Module for 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) [16].

(i) Vocal Tract Models:
(ii) Acoustic Models
(iii) Glottis Models
(iv) Noise Source Models

3] Concatenative Speech synthesis

Concatenative synthesis produces artificial speech by concatenating the pre-recorded units of speech such as phonemes, diphones, syllables, words or sentences [25][26].The size of speech unit stored affects the quality of the synthesized speech if large sentences are stored the speech synthesized will sound natural. Concatenative speech synthesis generate high quality synthesized speech [27]. Here we focus on two aspect of Concatenative speech synthesis, a search unit and speech database. In concatenative speech synthesis it takes small units of speech that is acoustically parameterised data and concatenate sequences of these small units together to produce waveform. Major factor influencing the quality of synthesized speech such as fundamental frequency, amplitude, speaking rate and the availability of speech units having appropriate prosody in the database [28].concatenative speech synthesis can done by three different methods.

(i) Diaphone based synthesis
(ii) Domain based synthesis and
(iii) Unit selection based synthesis

4] HMM-Based Synthesis

Hidden Markov models (HMMs) is a statistical machine learning speech synthesis to simulate real life stochastic processes [31]. The Hidden markov model (HMM) [31] [32] is a doubly stochastic process that produces a sequence of operations. In the HMM based speech synthesis the speech parameters of speech unit such as spectrum, fundamental frequency, and phoneme duration are statistically model and generated by using HMMs based on maximum likelihood criterion [33]. The major limitation of HMM is that they do not provide adequate representation of the temporal structure of speech. Hidden markov model is a collection of states connected by transition. In the synthesis part, a sentence HMM corresponding to an arbitrarily given text to be synthesized is constructed by concatenating context dependent HMMs, speech parameter vector sequences are generated from the HMMs. It can synthesize speech with various voice characteristics by transforming its model parameter such as speaker adaptation [34], [35], speaker interpolation [36], or eigenvoice technique [37], in the HMM-based speech synthesis system state durations are explicitly modelled by state duration models [38][39][40]. Then sequences of speech parameter vectors are generated from the given HMM using the speech parameter generation algorithm [41].

Conclusion

In this paper we discussed topics relevant to the development of the tts system. In this we have presented survey of several TTS techniques and challenges of TTS. Text to speech conversation is effective and efficient to users it produce high quality speech to the users. The desired speech is produced by using these methods, formant synthesis, articulatory and concatenative synthesis. The most commonly used techniques in present systems are concatenative synthesis and formant synthesis. The concatenative speech synthesis provide natural speech production. The overview of a concatenative speech synthesis system based on unit selection technique. The quality of the synthesized speech is affected by the unit length in the database

The naturalness of the synthesized speech increases with longer units. However, more memory is needed and the number of units stored in the database becomes very numerous. With formant synthesis it is more flexible and allows good control of fundamental frequency. .The synthesized speech is produced using an additive synthesis and an acoustic model. The technique produces highly intelligible synthesized speech. The third basic method is articulatory synthesis models the natural speech production process of human. Articulatory methods are usually complex.

References:

[1] Paul Taylor. Text-to-speech synthesis. Cambridge university press, 2009.
[2] Heiga Zen. Acoustic modeling in statistical parametric speech synthesis - from HMM to lstm-rnn. In Proc. MLSLP, 2015. Invited paper
[3] CLARK, J, and YALLOP, C. An introduction to phonetics and phonology. Basil Blackwell
[4] Mattingly I. G., Speech Synthesis for Phonetic and Phonological Models, T.A. Sebeok (Ed.) Current Trends in Linguistics, Vol. 12, (1974) p. 2451-2487.
[5]CLEMENTS, G. N. The geometry of phonological features. Phonology Yearbook 2 (1985),
225–252
[6] Yannis Agiomyrgiannakis. Vocaine the vocoder and applications in speech synthesis. In Acoustics,
Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pp. 4230–
4234. IEEE, 2015
[7] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves,Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. 09 2016.
[8] Mehri, Soroush, Kumar, Kundan, Gulrajani, Ishaan, Kumar, Rithesh, Jain, Shubham, Sotelo, Jose, Courville, Aaron, and Bengio, Yoshua. Samplernn: An unconditional end-to-end neural audio generation model
[9] Jose Sotelo, Soroush Mehri, Kundan Kumar, Jo˜ao Felipe Santos, Kyle Kastner, Aaron Courville, and
Yoshua Bengio. Char2Wav: End-to-end speech synthesis. In ICLR2017 workshop submission,
2017.
[10] Desai Siddhi, Jashin M. Verghese, Desai Bhavik, LIT, Sarigam "Survey on Various Methods of Text to Speech Syn-

thesis" International Journal of Computer Applications (0975 – 8887) Volume 165 – No.6, May 2017.
[11] 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).
[12] Rao, Kanishka, Peng, Fuchun, Sak, Has¸im, and Beaufays, Franc¸oise. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pp. 4225–4229.
[13] D. Klatt, ``Software for a cascade/parallel formant synthesizer,'' *Journal of the Acoustical Society of America*, vol. 67, pp. 13-33, 1980
[14] Archana Balyan, S.S. Agrwal and Amita Dev, Speech Synthesis: Review, IJERT, ISSN 2278-0181 Vol. 2 (2013) p. 57 – 75.
[15] Mark Tatham and Katherine Morton, Developments in Speech Synthesis (John Wiley & Sons, Ltd. ISBN: 0-470-85538-X, 2005).
[16] A. Indumati and Dr. E. Chandra, Speech processing –An Overview, Int. J. of Engg. Sci. and Tech., Vol. 4, (2012) p. 2853-2860.
[17] CHEN, S. F. Conditional and joint models for grapheme-to-phoneme conversion. In Proceedings
of Eurospeech 2003 (2003
[18]X. Rodet, ``Time-domain formant wave-function synthesis,'' *Computer Music J.*, pp. 6-14, 1984
[19] B. R. Glasberg and B. C. J. Moore, ``A model of loudness applicable to time-varying sounds,'' *Journal of the Audio Engineering Society*, vol. 50, pp. 331-342, May 2002.
[20] Pertti Palo. A Review of Articulatory Speech Synthesis. Espoo, June 5, 2006
[21] Fant, G. (1960). Acoustic Theory of Speech Production, Mouton and Co., Gravenhage, the Netherlands.
[22] Bernd J. Kröger, Peter Birkholz. Articulatory Synthesis of Speech and Singing: State of the Art and Suggestions for Future Research. Multimodal Signals: Cognitive and Algorithmic Issues. pp 306-319
[23] D. Klatt, ``Review of text-to-speech conversion for English,'' Journal of the Acoustical Society of America, vol. 82, pp. 737-793, Sept. 1987.
[24] D. H. Klatt and L. C. Klatt, ``Analysis, synthesis, and perception of voice quality variations among female and make talkers,'' Journal of the Acoustical Society of America, vol. 87, no. 2, pp. 820-857, 1990.
[25] Tabet Y, and Mohamed Boughazi. "Speech synthesis techniques. A survey." Systems, Signal Processing and their Applications (WOSSPA), 2011 7th International Workshop on. IEEE, 2011."
[26] A. Black and N. Campbell, "Optimizing selection of units from speech database for concatenative synthesis," Proc. of EUROSPEECH'95, vol. 1, pp. 581-584, Sept. 1995.
[27] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," Proc. of ICASSP, vol. 1, pp. 373-376, 1996.
[27] Samuel Thomas, "Natural sounding Text-to-speech synthesis based on syllable-like units," M S Thesis, IIT, Madras, 2007.
[28] Indumathi, A., and E. Chandra. "Survey on speech synthesis." Int J Signal Process 6 (2012): 140-5.
[29] Holmes, J., Holmes, W. (2001). Speech Synthesis and Recognition. Taylor and Francis, London, UK. 22.
[30] B.Kroger, "Minimal Rules for Articulatory Speech Synthesis", Proceedings of
EUSIPCO92, pp, 331-334, 1992
[31] K.Tucodo et al., "Hidden Semi-Marrov model based speech synthesis", Inter Speech
PP.1185-1180, 2004.
[32] X. D. Huang, Y. Ariki, and M. A. Jack, Hidden Markov models for speech recognition, Edinburgh University Press, 1990.
[33] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," IEICE Trans. D-II, vol.J83-D-II, no.11, pp.2099–2107, Nov. 2000 (in Japanese).
[28] J. Ferguson, Ed., "Hidden Markov Models for speech" IDA, Princeton, NJ, 1980
[29] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition" Proc. IEEE, 77(2), pp.257-286, 1989
[30]K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorigthms for HMM-based speech synthesis. In Proc. ICASSP 2000, pages 1315–1318, June 2000.
[31] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," Proc. of ICASSP, pp.1611–1614, 1997. [9] M. Tamura, T. Masuko, K. Tokud
[32] ACERO, A. Formant analysis and synthesis using hidden markov models. In Proceedings
Of Eurospeech 1999 (1999)
[33] AINSWORTH, W. A system for converting English text into speech. IEEE Transactions on
audio and electroacoustics 21 (1973).
[34] ALLEN, J., HUNNICUT, S., AND KLATT, D. From Text to Speech: the MITalk System.
[35] CAHN, J. A computational memory and processing model for prosody. In International
Conference on Speech and Language Processing (1998).Cambridge University Press, 1987.
[36] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," Proc. of ICASSP, pp.805–808, 2001.
[37] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," Proc. of Eurospeech, pp.2523–2526, 1997.
[38] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," Proc.of ICSLP, pp.1269–1272, 2002.
[39] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," Proc. of ICSLP, pp.29–32, 1998.
[40] Y. Ishimatsu, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura,"Investigation of state duration model based on gamma distribution.For HMM-based speech synthesis," Tech-

nical Report of IEICE, SP2001-81, pp.57–62, 2001. (In Japanese).
[41] J. Yamagishi, T. Masuko, and Kobayashi, "A study on state duration modeling using lognormal distribution for HMM-based speech synthesis," Proc. of ASJ, pp.225–226, March 2004. (In Japanese).
[42] https://www.greenbot.com/article/2105862/how-to-get-started-with-google-text-to-speech.html
[43] https://cloud.google.com/text-to-speech/