Survey On Speech Synthesis

A. Indumathi *induravi.cbe@gmail.com*
Asst.Professor,Department of Computer Applications(MCA)
Dr.SNS Rajalakshmi College Of Arts & Science
Coimbatore,Tamil Nadu,India

Dr. E. Chandra *crcspeech@gmail.com*
Director,Department of Computer Applications(MCA)
Dr.SNS Rajalakshmi College OF Arts & Science
Coimbatore, Tamil Nadu, India

Abstract

The primary goal of this paper is to provide an overview of existing Text-To-Speech (TTS) Techniques by highlighting its usage and advantage. First Generation Techniques includes Formant Synthesis and Articulatory Synthesis. Formant Synthesis works by using individually controllable formant filters, which can be set to produce accurate estimations of the vocal-track transfer function. Articulatory Synthesis produces speech by direct modeling of Human articulator behavior. Second Generation Techniques incorporates Concatenative synthesis and Sinusoidal synthesis. Concatenative synthesis generates speech output by concatenating the segments of recorded speech. Generally, Concatenative synthesis generates the natural sounding synthesized speech. Sinusoidal Synthesis use a harmonic model and decompose each frame into a set of harmonics of an estimated fundamental frequency. The model parameters are the amplitudes and periods of the harmonics. With these, the value of the fundamental can be changed while keeping the same basic spectral..In adding, Third Generation includes Hidden Markov Model (HMM) and Unit Selection Synthesis.HMM trains the parameter module and produce high quality Speech. Finally, Unit Selection operates by selecting the best sequence of units from a large speech database which matches the specification.

Keywords: TTS, HMM, Synthesis

1. INTRODUCTION

With advancement in global communication, Speech processing has been a main stream in the area of research. The key goal of speech research is to build a system that mimic human. It can be achieved by speech recognition and speech synthesis. The first one converts the speech into text information, whereas, the second converts the text message information into speech which is also called Text-To-Speech (TTS).

Speech synthesis is rapidly developing technology which consists of two major phases [1].

(i) Text Analysis where the input as text is transcribed into a phonetic or linguistic representation using pronunciation rules and
(ii) Generation of speech waveforms or speech synthesis, where the acoustic speech output is produced from phonetic and prosodic information

Speech Synthesis can be performed by using any one of the approach in Speech Generation Techniques [3]. They are

- First Generation Techniques
- Second Generation Techniques
- Third Generation Techniques

First Generation Techniques require a quite detailed, low-level description of what is to be spoken. Second Generation Techniques uses Data driven approach to increase the quality of speech by reducing modeling pitch and timing. On the other hand, Third Generation techniques use Statistical, Machine-learning techniques to infer the specification - to - parameter mapping from data. Main reason for the arise of third generation is due to less memory occupation for storing factors of the model than to memorize the data. In addition to that, it allows to modify the model in various ways, like converting the original voice into a different voice[2].

2. FORMANT SYNTHESIS

Formant Synthesis was the first synthesis technique to be developed and was the dominant technique until the early 1980's.Formant Synthesis is often called synthesis by rule. The basic assumption of Formant synthesis is to model vocal tract transfer function by simulating formant frequencies and formant amplitudes. The vocal track has certain major resonant frequencies[4] .The frequencies change as the configuration of the vocal tract changes like resonant peaks in the vocal track transfer function (frequency response) are known as "formants".

The synthesis is a sort of source-filter-method that is based on mathematical models of the human speech organ. The formant synthesizer makes use of the acoustic-tube model, where the sound is generated from a source, which is periodic for voiced sounds and white noise for obstruent sounds. This basic source signal is then fed into the vocal-tract model. This signal passes into oral cavity and nasal cavity and finally it passes through a radiation component, which simulates the load propagation characteristics to produce speech pressure waveform.

Formant synthesis technology generates artificial and robotic-sounding speech. Formant-synthesized speech is reliably intelligible, even at very high speeds. Formant synthesis is not a very computationally intensive process especially for today's computing systems [5]. The strength of formant synthesis is its relative simplicity and the small memory footprint needed for the engine and its voice data. This acts as main advantage for embedded and mobile computing applications. DecTalk, Apollo, Orpheus and Eloquence are well known TTS engines that use formant synthesis.

3. ARTICULATORY SPEECH SYNTHESIS

Articulatory speech synthesis uses mechanical and acoustic models of speech production to synthesize speech. Articulatory speech synthesis transforms a vector of anatomic or physiologic parameters into a speech signal with predefined acoustic properties [1]. It produces a complete synthetic output, based on mathematical models of the structure (Lips, Teeth,Tongue,Glottis & Velum) processes(transit of airflow along the supraglottal cavities) of speech. This technique is computation-intensive so a memory necessity is almost nothing.

Acoustic models contain number of smaller uniform tubes which generate natural speech. These tubes are controlled by themselves. Natural movements in tubes can give rise to the complex patterns of speech, thus bypassing the problems of modeling complex formant trajectories explicitly. Articulatory synthesis models have an interim stage, in which the motion of the tubes is controlled by some simple process (mechanical damping or filtering), intended to model the fact that the articulators move with a certain inherent speed. This motor-control space is then used to provide the parameters for the specification-to-parameter component.

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 and a more pragmatic representation that is easy to design and control[6].

4. CONCATENATIVE SPEECH SYNTHESIS

Concatenative synthesis depends on speech signal processing of natural speech databases. The segmental database is built to reflect the major phonological features of a language [12]. Concatenation techniques take small units of speech, either waveform data or acoustically parameterized data, and concatenate sequences of these small units together to produce either time varying acoustic parameters or, alternatively, waveforms. The time-varying acoustic parameters then need to be converted into a waveform by passing them through a speech synthesizer.

A concatenation system are concerned with the selection of appropriate units and the algorithms that join those units together and performs some signal processing to smooth unit transitions and to match predefined prosodic schemes. They are three vital subtypes of Concatenative synthesis [9].

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

(i) Diphone based synthesis

Diphone synthesis is most popular method used for creating a synthetic voice from recordings or samples of a particular person. It uses a nominal speech database[11] .The quantity of diphones in database depends on the phonotactics of the language. In diphone synthesis, the strength of speech depends on sentence or expression and the model used for prosody. The speech may sound a bit monotonic because of improper modeling. A diphone synthesis doesn't work well in languages where there is a lot of inconsequence in the pronunciation rules and in special cases where letters is pronounced differently than in general. The diphone works better for languages that have large consistencies in the pronunciation

The major problem with diphone synthesis is, discontinuities happening at the interface between two bisects of a vowel. In some cases, they create a bi-vocalic sound quality and an perceptible discontinuity at the diphone boundary

(ii) Domain based synthesis

Domain based synthesis concatenates prerecorded vocabulary and axioms to produce entire utterances [1]. It is used in places where output speech is narrowed to a specific domain, like announcement of transit schedule, weather condition reports etc .As these systems are restricted by vocabulary and axioms in their databases, they can only create the combinations of vocabulary and axioms with preprogrammed content.

5. SINUSOIDAL SYNTHESIS

Sinusoidal synthesis uses a harmonic model and decomposes each frame into a set of harmonics of an estimated fundamental frequency [3]. The fundamental parameters like amplitudes, frequencies and phases are changed by keeping the same spectral envelope. The basic idea is to model every significant spectral component as a sinusoid

This model composes harmonic component and a noise component. The first stage in this technique is to classify the frames into voiced and unvoiced portion and dictate the parameters of the harmonic and noise components and the relative strength of the contribution of each component to the frame [7]. Estimated pitch and fundamental frequency is fitted to each frame and from this error between the speech generated and the real waveform is found. Harmonic frames with low error are considered as voice part and frames with high error are considered as noise part.

6. HIDDEN MARKOV MODEL SYNTHESIS

Hidden Markov models (HMMs) is a statistical machine learning speech synthesis to simulate real life stochastic processes [8]. In HMM based synthesis, the speech parameters like frequency spectrum, fundamental frequency and duration are statistically modeled and speech is generated by using HMM based on maximum likelihood criterion. A hidden Markov model is a collection of states connected by transitions. Each transition carries two sets of probabilities: a transition probability, which provides the probability for taking the transition, and an output probability density function, which defines the conditional probability of emitting each output symbol from a finite alphabet, given that the transition is taken.

HMM synthesis provides a means by which to train the specification to parameter module automatically, thus bypassing the problems associated with hand –written rules[12]. The trained models can produce high quality synthesis, and have the advantages of being compact and amenable to modification for voice transformation and other purposes.

This technique consumes large CPU resources but very little memory. This approach seems to give a better prosody, without glitches, and still producing very natural sounding, human-like speech [15].

7. UNIT SELECTION SYNTHESIS

Unit selection synthesis is the dominant synthesis technique in text to speech .This technique is the extension of second generation concatenative technique and deals with the issues of how to manage large numbers of units, how to extend prosody and timing control, and how to alleviate the distortions caused by signal processing [11].

Unit selection technique uses a rich variety of speech, with the aim of capturing more natural variation and relying less on signal processing [5]. The specification and the units are completely described by a feature structure, which can be any mixture of linguistic and acoustic features. During synthesis, an algorithm selects one unit from the possible choices, in an attempt to find the best overall sequence of units which matches the specification and also it enables us to use the carrier speech and lessen the problems arising from designing and recording a database that creates a unit for every feature value.

The greatest difference between a Unit selection and a diphone voice is the length of the used speech segments. There are entire words and phrases stored in the unit database. This implies that the database for the Unit selection voices is many times bigger than for diphone voices [13]. Thus, the memory consumption is huge while the CPU consumption is low

8. CONCLUSION

Speech synthesis is a developing technology which has been incorporated in many real time applications. Existing speech synthesis technique produces quite intelligible and acceptable level of speech output [11] .There is still a long way to reach goal .A number of advances in the area of NLP or DSP have recently boosted up the quality and naturalness of available voices. However researcher have to concentrate on certain areas like prosodic, text preprocessing and pronunciation in order to produce natural and pleasant speech, improve voice quality and linguistic analysis.

In Formant synthesis, rules are needed to specify timing of the source and the dynamic values of all filter parameters which is difficult for simple words. Each synthesis technique has its own limitations and it can be selected depending on the applications. Articulatory synthesis produces intelligible speech, but its output is far from natural sounding .Collecting articulatory data is a costly and fairly invasive process. Articulatory synthesis is appealing for scientific purposes and may one day provide completely “tunable” high quality speech. In Unit selection synthesis, determination of high–level linguistic features are easy, but lack natural distances and can lead to

an explosion in feature combinations.HMM synthesis produce speech that is smooth but of poor voice quality however, it has several advantages over unit selection synthesis like flexibility, small footprint, can combine with new techniques to generate new voices with very small training sets.

9. FUTURE DIRECTIONS

Speech synthesis may be used in all kind of human machine interactions. For example, in warning and alarm systems synthesized speech may be used instead of warning lights or buzzers. Speech synthesis takes the major role in mobile phones. The future of text-to-speech can be evolved by improving the following properties.

Text analysis *:* TTS is entirely data-driven and the recent advances in statistical NLP can be used for search engine and document translation .I think many of these techniques are directly applicable to TTS and can be adopted.

Synthesis Algorithms: Its harder to predict concern speech synthesis algorithms. A few years ago, formant synthesis ,concatenative synthesis seemed to be the dominant technique, but recently HMM synthesis and Unit synthesis become dominant .I believe that each speech synthesis techniques has its own weight age according to their applications**.**

Quality Improvements*:* In terms of Overall quality and performance, the problems of text analysis can be fully solved with today's technology. The researchers have to concentrate on good quality databases.

Relationship with linguistics: Overall performance of speech synthesis can be increased by improving the relationship with linguistics.

10. REFERENCES

[1]. L.R.Rainer,"Applications of Voice processing to Telecommunications", *Proc.IEE,Vol.82,PP.199-228,1994.*

[2]. J.Allen,S.Hunncutt and D.H. Klatt,"From Text – to – Speech:The MItalk Systems", *Cambridge university Press, Cambridge, 1987.*

[3]. Hon,H.Acero,A.Huang,X.,Liu,J.,and plumpe,M."Automatic Generation of Synthesis units for Trainable Text-to-speech Systems" . *In proceedings of the IEEE international conference on Acousitics, Speech and Singnal Processing 1998.*

[4]. T.Styger ,E Keller, "Format Synthesis", *Fundamental of Speech Synthesis and Speech Recognition; Basi concept,State of the Art and future challenges (PP.109-128).*

[5]. D.H.Klatt "Review of Text – to- speech Conversion for English", *Journal of the Accoustical Society of America, Vol. 82(3),1987.*

[6]. B.Kroger, "Minimal Rules for Articulatory Speech Synthesis", *Proceedings of EUSIPCO92,pp,331-334,1992.*

[7]. Y.Stylianao, "Modeling Speech Based on the Harmoni Plus Noise Models,"*Springes 2005.*

[8]. K.Tucodo et al., "Hidden Semi-Marrov model based speech synthesis", *Inter Speech PP.1185-1180,2004.*

[9]. Moulines,E., and Charpertier, F."Pitch Synchronous waveform processing techniques for Text- to –Speech Synthesis using diphones." *Speech Communications 9,pp 453-67.1990.*

[10]. Garcia,G.J. Pampin,1999," Data Compression of Sinusoidal modeling parameters based on psychoaccousitc marking", *Proc.ICMC. International computer Music Conference,Beijing,China.*

[11]. A.Hunt and A.Black, " Unit selection in Concatenative Speech Synthesis system using large speech database," *Proc IEEE Int, Conf-Accoust., Speech,Signal Processing,pp 373-376,1996.*

[12]. A.Black and Po Taylor,"The Festival Speech Synthesis System:system documentation,Technical Report" *HCHC/TR – 8s, 1997.*

[13]. Wouter. J., and Macon, M. W. "Unit fusion for concatenative speech synthesis". *In proceedings of the International Conference on Spoken Language Proceedings 2000(2000).*

[14]. Yamagishi, J., Onishi,K., Masuko, T., and Kobayashi, T. "Modeling of Various Speaking styles and emotions for HMM-Based speech synthesis". *In Proceedings of Eurospeech 2003(2003).*

[15]. Zen.H. Tokuda, K.Masuko, T., Kobayashi, T., and Kitamura, T.Hidden "Semi-Makov model based synthesis." *In the Proceedings of 8th International Conference on Spoken Language Processing, Interspeech 2004(2004).*

[16]. *Othman.O.Khalifa,et al "SMA Talk: Standard malay Text to Speech Talk System"-Signal Processing :An International Journal (SPIJ),Vol:2,Issue:5,pp:1-26,2008.*