International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-2, Issue-1, March 2012

TEXT TO SPEECH: A SIMPLE TUTORIAL

D.Sasirekha, E.Chandra

Abstract:*Research on Text to Speech (TTS) conversion is a large enterprise that shows an impressive improvement in the last couple of decades. This article has two main goals. The first goal is to summarize the published literatures on Text to Speech (TTS), with discussing about the efforts taken in each paper. The second goal is to describe specific tasks concentrated during Text to Speech (TTS) conversion namely, Preprocessing & text detection, Linearization, Text normalization, prosodic phrasing, OCR, Acoustic processing and Intonation. We illustrate these topics by describing the TTS synthesis. This system will be highly useful for an illiterate and vision impaired people to hear and understand the content, where they face many problems in their daily life due to the differences in their script system. This paper starts with the introduction to some basic concepts on TTS synthesis, which will be useful for the readers who are less familiar in this area of research.*

Index Terms—TTS.

I. INTRODUCTION

A text to speech (TTS) synthesizer is a computer based system that can read text aloud automatically, regardless of whether the text is introduced by a computer input stream or a scanned input submitted to an Optical character recognition (OCR) engine. A speech synthesizer can be implemented by both hardware and software. It has been made a very fast improvement in this field over the couple of decades and lot of high quality TTS systems are now available for commercial use.

Speech is often based on concatenation of natural speech i.e units,that are taken from natural speech put together to form a word or sentence. Concatenative speech synthesis [1] has become very popular in recent years due to its improved sensitivity to unit context over simpler predecessors.

Rhythm [2] is an important factor that makes the synthesized speech of a TTS system more natural and understandable, The prosodic structure provides important information for the prosody generation model to produce effects in synthesized speech.

Many TTS systems are developed based on the principle, corpus-based speech synthesis [3] [10]. It is very popular for its high quality and natural speech output.

According to [4] [5], the next generation TTS systems are asked to deal with emotions in speaking styles. And there has been growing interest in developing commercial systems based on Limited Domain (LD-TTS) [6], which restricts the scope of the input text so as to obtain high quality speech synthesis.

As there are number of research prototypes of TTS systems has been developed and none was compared with the commercial grade TTS systems for quality. The main reason is that it needs improvisation in collaboration between linguistics and technologists.

Text to speech should be made audibly communicate information to the user, when digital audio recordings are inadequate, for developing a user friendly speech synthesizer. Thus this system widely helps

in developing a Computer-Human interaction like- voice annotations to files , Speech enabled applications, talking computer systems (GPS, Phone-based) etc.

Section II describes the evolution of the system and

Section III describes the steps involved in developing a effective text to speech (TTS) system..

II. EVOLUTION OF TTS

Let us start , with the understanding of the progression of text to speech (TTS) system. In 1779, the Danish scientist Christian Kratzenstein, working at the Russian Academy of Sciences, built models of the human vocal tract that could produce the five long vowel sounds they are [a], [e], [i], [o] and [u].In 1791, an Austrian scientist developed a system based on the previous one included tongue, lips and “mouth” made of rubber and a “nose” with two nostrils which was able to pronounce consonants. In 1837, Joseph Faber developed a system which implemented Pharyngeal Cavity, used for singing. It was controlled by keyboard.

Bell Labs Developed VOCODER, a clearly intelligible. keyboard-operated electronic speech analyzer and synthesizer. In 1939, Homer Dudely developed VODER which was a`n improvement over VOCODER.

The Pattern Playback was built by Dr. Franklin S. Cooper and his colleagues at Haskins Laboratories.First Electronic based TTS system was designed in 1968.

Concatenation Technique was developed by 1970’s. Many computer operating systems have included speech synthesizers since the early 1980s. From 1990’s , there was a progress in Unit Selection and Diphone Synthesis.

D.Sasirekha, Computer Science ,Research Scholar, Karpagam University, Coimabtore, India, dsasirekha@gmail.com

Dr.E.Chandra, Dean, School Of Computer Studies,
Dr. S.N.S. Rajalakshmi College of Arts and Science, Coimbatore, India, crcspeech@gmail.com

III. ARCHITECTURE OF TTS

The TTS system comprises of these 5 fundamental components:

A. Text Analysis and Detection
B. Text Normalization and Linearization
C. Phonetic Analysis
D. Prosodic Modeling and Intonation
E. Acoustic Processing

The input text is passed through these phases to obtain the speech.

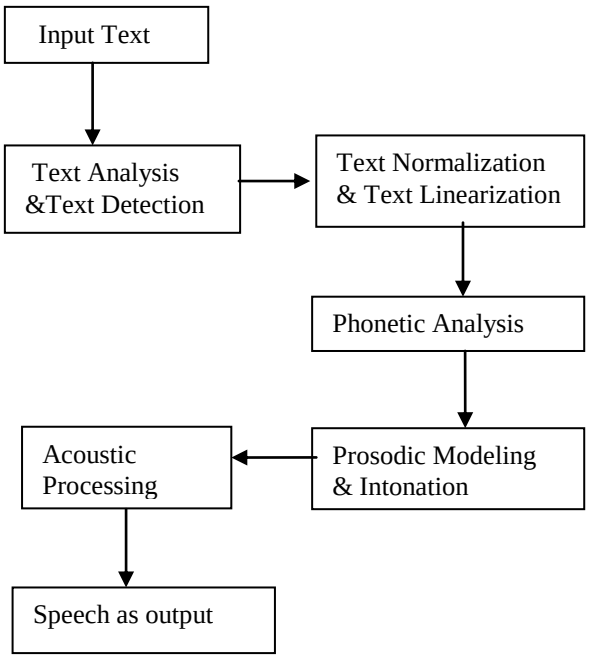

Fig 1: System Overview of TTS

A. Text Analysis and Detection

The Text Analysis part is preprocessing part which analyse the input text and organize into manageable list of words. It consists of numbers, abbreviations, acronyms and idiomatics and transforms them into full text when needed. An important problem is encountered as soon as the character level : that of punctuation ambiguity (sentence end detection). It can be solved, to some extent, with elementary regular grammars

Text detection is localize [8] the text areas from any kind of printed documents. Most of the previous researches were concentrated on extracting text from video. We aim at developing a technique that work for all kind of documents like newspapers, books etc

B. Text Normalization and Linearization

Text Normalization is the transformation of text to pronounceable form. Text normalization is often performed before text is processed in some way, such as generating synthesized speech or automated language translation. The main objective of this process is to identify punctuation marks and pauses between words. Usually the text normalization process is done for converting all letters of lowercase or upper case, to remove punctuations, accent marks , stopwords or "too common words "and other diacritics from letters .

Text normalization is useful for example for comparing two sequences of characters which represented differently but mean the same. "Don't" vs "Do not", "I'm" vs "I am", "Can't" vs "cannot" are some of the examples.

The main 4 phases of Text Normalization are

(i). Number converter: Number is pronounced differently in different situations. Like

1772 (date): seventeen seventy two.
1772(phone number): one seven seven two
1772 (quantifier): one thousand seven hundred and seventy two .

Fractional and decimal numbers are handled.
0.302 (number): point three knot two

(ii). Abbrevation converter:
Abbreviations area changed to full textual format.

Mrs. - Misses
St. Joseph St. - Saint Joseph Street

(iii). Acronym converter: Acronyms are replaced by single letter components.

S. I. - S I

(iv). Word segmentation:
Sentences are a group of word segments. Special delimiter to separate segments.
(i.e. '||').Segments can be an acronym, a single word or a numeral.

Examples of acronyms:

"NATO" - "nayto"
"HIV" - "aitch eye ve"
"Henry IV" - "Henry the fourth"
"Chapter IV"- "Chapter four"
Punctuation marks are also identified.

Linearization is the process of giving a hyper text link to give the user a quick overview of the page. Then the TTS system will help to read out the linearized data.This feature helps in selecting the text and reading and also to list the links in the hyper text.

C. Phonetic Analysis

Phonetic Analysis converts the orthographical symbols into phonological ones using a phonetic alphabet. Basically known as "grapheme-to-phoneme" conversion.

Phone is a sound that has definite shape as a sound wave. Phone is the smallest sound unit. A collection of phones that constitute minimal distinctive phonetic units are called Phoneme. Number of phonemes is relatively smaller than the graphemes, only 44.

Phoneme Set (English)

- Vowels (19) : /a/, /ae/, /air/, /ar/, /e/, /ee/, /i/, /ie/, /o/, /oe/, /oi/, /oo/, /ow/, /or/, /u/, /ur/, /ue/, /uh/, /w/.

- Consonants (25) : /b/, /ks/gz/, /c/k/, /ch/, /d/, /f/, /g/, /h/, /j/, /l/, /m/, /n/, /ng/, /p/, /kw/, /r/, /s/, /sh/, /t/, /th/, /th/, /v/, /y/, /z/, /zh/.

Examples:

- *o /air/ : square, bear.*
- *o /ow/ : down, house.*
- *o /ks/gz/ : box, exist*

Pronunciation of word based on its spelling has two approaches to do speech synthesis namely

(a)Dictionary based approach
(b) Rule based approach.

A dictionary is kept were It stores all kinds of words with their correct pronunciation, it's a matter of looking in to dictionary for each word for spelling out with correct pronounciation. This approach is very quick and accurate and the pronounciation quality will be better but the major drawback is that it needs a large database to store all words and the system will stop if a word is not found in the dictionary.

The letter sounds for a word are blended together to form a pronunciation based on some rule. Here main advantage is that it requires no database and it works on any type of input. same way the complexity grows for irregular inputs

D. *Prosodic Modelling and Intonation*

The concept of prosody is the combination of stress pattern , rhythm and intonation in a speech. The prosodic modeling describes the speakers emotion. Recent investigations suggest the identification of the vocal features which signal emotional content may help to create a very natural [9] synthesized speech.

Intonation is simply a variation of speech while speaking. All languages use pitch, as intonation to convey an instance, to express happiness, to raise a question etc. Modelling of an intonation is an important task that affects intelligibility and naturalness of the speech. To receive high quality text to speech conversion, good model of intonation is needed.

Generally intonations are distinguished as

(i) Rising Intonation
(when the pitch of the voice increases)
(ii) Falling Intonation
(when pitch of the voice decreases)
(iii) Dipping Intonation
(when the pitch of the voice falls and then rises)
(iv) Peaking Intonation
(when the pitch of the voice raises and then falls)

E. *Acoustic Processing*

The speech will be spoken according to the voice characteristics of a person, There are three type of Acoustic synthesing available

(i).Concatenative Synthesis
(ii).Formant Synthesis
(iii).Articulatory Synthesis

The concatenation of prerecorded human voice is called Concatenative synthesis, in this process a database is needed having all the prerecorded words .The natural sounding speech is the main advantage and the main drawback is the using and developing of large database.

Formant-synthesized speech can be constantly intelligible .It does not have any database of speech samples. So the speech is artificial and robotic.

Speech organs are called Articulators. In this articulatory synthesis techniques for synthesizing speech based on models of the human vocal tract are to be developed. It produces a complete synthetic output, typically based on mathematical models

IV. CONCLUSION

This paper made a clear and simple overview of working of text to speech system (TTS) in step by step process. There are many text to speech systems (TTS) available in the market and also much improvisation is going on in the research area to make the speech more effective, natural with stress and emotions. We expect the synthesizers to continue to improve research in prosodic phrasing, improving quality of speech, voice, emotions and expressiveness in speech and to simplify the conversion process so as to avoid complexity in the program.

References

[1] Frances Alias, Xavier Servillano, Joan Claudi socoro and Xavier Gonzalvo "Towards High-Quality Next Generation Text-to-Speech Synthesis:A multi domain Approach by Automatic Domain Classification",IEEE Transactions on AUDIO,SPEECH AND LANGUAG PROCESSING, VOL16,NO,7 september 2008.

[2] Qing Guo, Jie Zhang, Nobuyuki Katae, Hao Yu , "High –Quality Prosody Generation in Mandrain Text-to-Speech system", FujiTSu Sci.Tech,J., vol.46, No.1,pp.40-46 ,2010.

[3] Gopalakrishna anumanchipalli,Rahul Chitturi, Sachin Joshi, Rohit Kumar, Satinder Pal Singh,R.n.v Sitaram,D.P.Kishore, "Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition System",

[4] A.Black, H.Zen and K.Tokuda "Statistical parametric speech synthesis", in proc.ICASSP, Honolulu, HI 2007, vol IV, PP 1229-1232.

[5] G.Bailly, N.Campbell and b.Mobius, "ISCA special session: Hot topics in speech synthesis", in proc.Eurospeech,Genea, Switzerland, 2003, pp 37-40.

[6] M.Ostendorf and I.Bulyko, "The impact of speech recognition on speech synthesis", in proc, IEEE Workshop Speech Synthesis, Santa Monica,2002,pp. 99-106.

[7] Text To Speech Synthesis - a knol by Jaibatrik Dutta .

[8] Silvio Ferreia,Celina Thillou, Bernaud Gosselin, "From Picture to Speech: an Innovative Application for Embedded Environment",
[9] M.Nageshwara Rao, Samuel Thomas, T.Nagarajan and Hema A.Muthy, "Text-to-Speech Syntheis using syllable line units"
[10] Jindrich Matousek, Josef Psutks, Jiri Krita, "Design of speech Corpus for Text-to-Speech Synthesis"

D.Sasirekha , completed her BSc (CS)-2003 in Avinashlingam University for Women, coimbatore and M.Sc (CS)-2005 in Annamalai University, Currently doing Ph.D (PT) (CS) in Karpagam University, Coimbatore and working as a staff in Avinashilingam University for Women, Coimbatore,India.

Dr.E.Chandra received her B.Sc., from Bharathiar University, Coimbatore in 1992 and received M.Sc., from Avinashilingam University ,Coimbatore in 1994. She obtained her M.Phil., in the area of Neural Networks from Bharathiar University, in 1999. She obtained her PhD degree in the area of Speech recognition system from Alagappa University Karikudi in 2007. She has totally 16 yrs of experience in teaching including 6 months in the industry. At present she is working as Director, School of Computer Studies in Dr.SNS Rajalakshmi College of Arts & Science, Coimbatore. She has published more than 30 research papers in National, International journals and conferences in India and abroad. She has guided more than 20 M.Phil., Research Scholars. At present 3 M.Phil Scholars and 8 Ph.D Scholars are working under her guidance. She has delivered lectures to various Colleges in Tamil Nadu & Kerala. She is a Board of studies member at various colleges. Her research interest lies in the area of Neural networks, speech recognition systems, fuzzy logic and Machine Learning Techniques. She is a Life member of CSI, Society of Statistics and Computer Applications. Currently Management Committee member of CSI Coimbatore Chapter.