An Intelligent Text to Speech System for Windows based Systems and Mobile Devices

Abhishek Srivastava
PEC University of Technology
Electronics and Communication Engineering Department,
Chandigarh, India

Akshay Sharma
PEC University of Technology
Electronics and Communication Engineering Department,
Chandigarh, India

Neelu Jain
PEC University of Technology
Electronics and Communication Engineering Department,
Chandigarh, India

ABSTRACT
TTS (Text-to-speech) systems are used invariably as part of our daily lives and have come a long way. In this paper TTS system using Concatenative synthesis based on the SDK (Software Development Kit) platform has been presented. This system is compatible with both computer and mobile devices. It has a user friendly GUI (graphical user interface) to control various speech parameters. Speech signal produced can be saved and listened to whenever required. Signal analysis of the output speech can also be done using TTS System. The results of these signal analysis along with the stored speech signal can be used for further applications depending upon the requirements. It is an intelligent system and is able to overcome various normalization problems.

Key Words
TTS, SDK, Concatenative synthesis, GUI

1. INTRODUCTION
TTS synthesis is a technique for generating intelligible, natural-sounding artificial speech for a given text [1].It has been used widely in various applications such as e-book-readers, automated telecom services, as a part of a network voice server for e-mail, voice-over functions for the visually impaired, communication aids for the speech impaired, communicative robots and speech-to speech translation systems. It can also be used by people with dyslexia to read or to check self-written text by listening [2]

The methodology used in TTS is to exploit acoustic representations of speech for synthesis, together with linguistic analysis of text to extract correct pronunciations (‘‘content’’, what is being said) and prosody in context (‘‘melody’’ of a sentence; how it is being said).Synthesis systems are commonly evaluated [3]in terms of three characteristics: accuracy of rendering the input text (as: acronyms, names, URLs, email addresses, as a knowledgeable human would), intelligibility of the resulting voice message (measured as a percentage of a test set that is understood), and perceived naturalness of the resulting speech.

2. TTS SYSTEM OVERVIEW
A speech synthesis system can be divided into two parts (see Figure 1). Front End also called Natural Language Processing Module (NLP) [4] analyzes text, and Back End, also called Signal-Processing Module, generates the speech waveform based on information from the front end. Front end contains: text processor (normalization and letter-to-sound), prosody control, unit selection [5]. So it is basically concerned with the conversion of grapheme- to-phoneme. This process is also called ‘‘letter-to-sound’’ conversion. Back End is concerned with method used for synthesis. In the literature [6] we find two basic categories of methods: format synthesis [7,8,9] and concatenative synthesis[10,11,12].

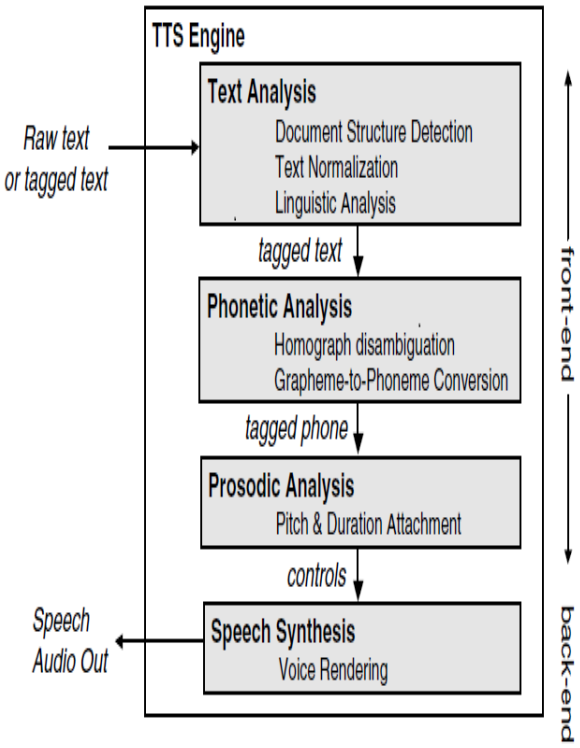

Fig 1: Block diagram of general Text to Speech System

Format synthesis depends on acoustical models in order to produce parametric driven speech, while concatenative synthesis, concatenates segments of recorded speech. Well known examples of format synthesis based systems are ‘‘MITalk’’ and ‘‘DECTalk’’[9].

Concatenative synthesis uses actual short segments of recorded speech that were cut from recordings and stored in voice database. There are three variants of concatenative synthesis, based on the types of speech units stored in the database of a Concatenative TTS system: domain specific synthesis, diphone synthesis and unit selection synthesis [13].Domain specific synthesis normally concatenates words or phrases of speech and can be used when the output of the synthesis system is limited to a small domain of utterances.

Diphone synthesis speech databases consist of only one unit of each diphone [14] occurring in the language. During synthesis,

pitch and duration modification are used to obtain a desired prosody. Unit selection synthesis is the most popular variant of concatenative synthesis and was first proposed by Nakajama and Hamada in 1988[15]. Since then various systems including commercial systems were developed resulting in a higher level of reading-style synthetic speech [16,17,18] and it is today considered as the state of art in text-to speech synthesis.

2.1 Various issues involved in text-to-speech analysis

The text analysis and normalization module in the front-end determines to a large extend the ‘‘what’’ and ‘‘how’’ of the resulting synthetic speech. Text normalization is difficult because it is context sensitive.

Abbreviations and acronyms fall in either of two categories. The first category contains a finite set of known ‘‘mappings’’ such as ‘‘Dr.’’ in the sentence ‘‘Dr. Marine drove to Marine Dr.’’. Note that a mapping may be ambiguous (Dr. can be ‘‘doctor’’ or ‘‘drive’’) .But more difficult to handle, is the open category of abbreviations that people invent on the fly eg. COMM could mean ‘‘communications’’, ‘‘committee’’ etc. Even the simple reading of numbers can be difficult, such as ‘‘452,’’ where the 452 can be part of a phone number (452–1111), read as ‘‘four five two . . .’’) or part of a name (e.g., INTEL452, read as ‘‘INTEL-four-fifty two’’. Other issue is to decide which synthesis method (Format synthesis or concatenative synthesis) to use for designing of a TTS system. Model based synthesis can be highly intelligible, but due to the difficult and complex task of obtaining good enough speech models, the synthesized speech has so far a degraded speech quality to some extent. Where as Concatenative synthesis can be very natural in the sense of having a speech quality close to human speech, but it may suffer from audible discontinuities at concatenation points

3. TTS SYSTEM AND SPEECH PROCESSING APPLICATION ARCHITECHURE

This TTS System has been designed using MATLAB GUI and is based on SDK platform of Windows based systems. The Speech API can be viewed as an interface used between applications and speech engine (recognition and synthesis). Speech API is provided either as a part of Microsoft Speech SDK or as part of the Windows OS itself and uses concatenative synthesis whose architecture is shown. (see Fig 2)

MATLAB GUI has been developed to communicate [19] with API by sending events using standard callback mechanisms (Window Message, callback proc or Win32 Event) (see Figure 3) such that they are accessible from a variety of programming languages by using a standard set of interfaces. In addition, it is possible for a 3rd-party to produce their own Speech Recognition and Text-To-Speech engines or adapt existing engines to work with API.

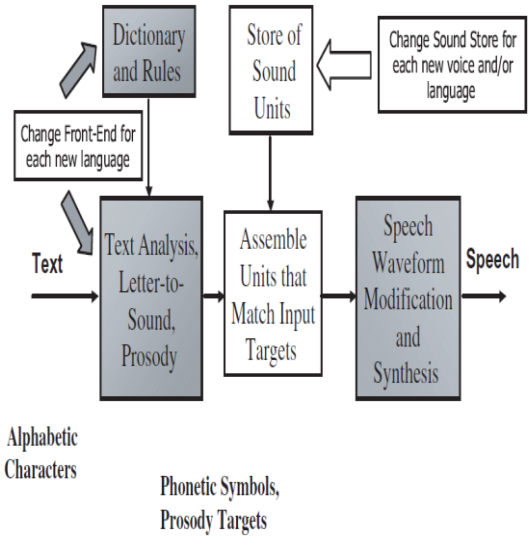

Fig 2: Architecture of a Concatenative text-to-speech system

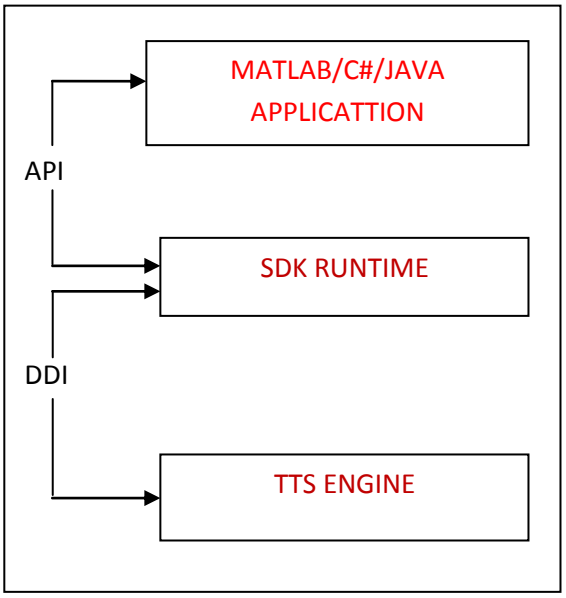

Fig 3: Architecture of TTS System

4. TTS Mobile Application Development

This TTS system can be converted into Mobile app (application) for windows based mobile phone. Windows Mobile OS (WP7/8) has an integrated development environment, which provides tools to allow a developer to write test and deploy applications into the target platform environment (see Figure 4).

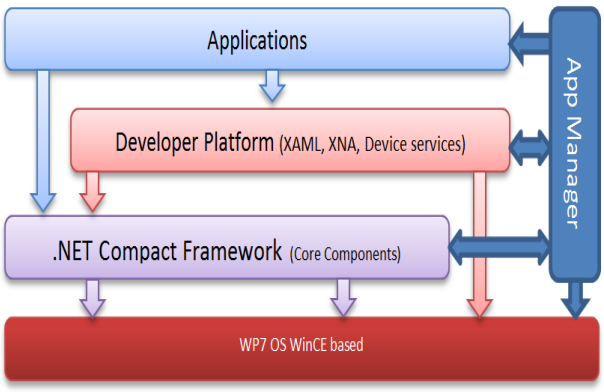

Fig 4: Architecture for development of windows mobile app

In Windows Phone, for apps to be designed, and tested following development tools are needed [19, 20].
Compilers: C#, Visual Basic, C, C++

Integrated development environment: Visual Studio 2010/12
Installer package: OTA deployment, XAP files

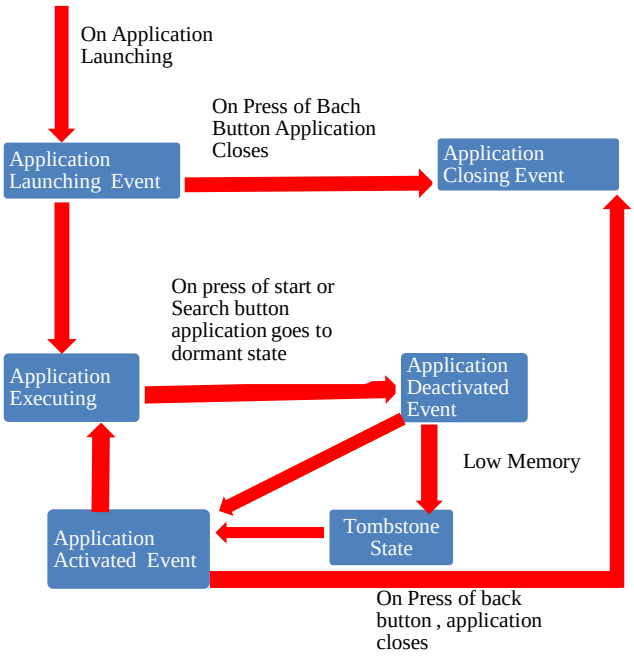

Fig 5: Windows phone application life cycle

The events that take place during the life cycle of a windows phone application are shown (see Figure 5).Application developer needs to take appropriate actions in response to each of these events while designing a mobile application.

5. MAIN FEATURES OF TTS SYSTEM

The developed TTS system has following features on its GUI (see Figure 6)

5.1 Browsing Feature

It is used to read text files present anywhere in the computer using this browsing feature.

5.2 Speaking Rate Control

Controls the speed at which the text is spoken. This tag can be empty or nested.

Attributes: Only one attribute may be applied within a tag. absspeed=Sets the absolute speed for the speech in a range between -10 and 10 with 0 being normal speech. eg. <rate absspeed="5"/>Speak all following text at rate 5.

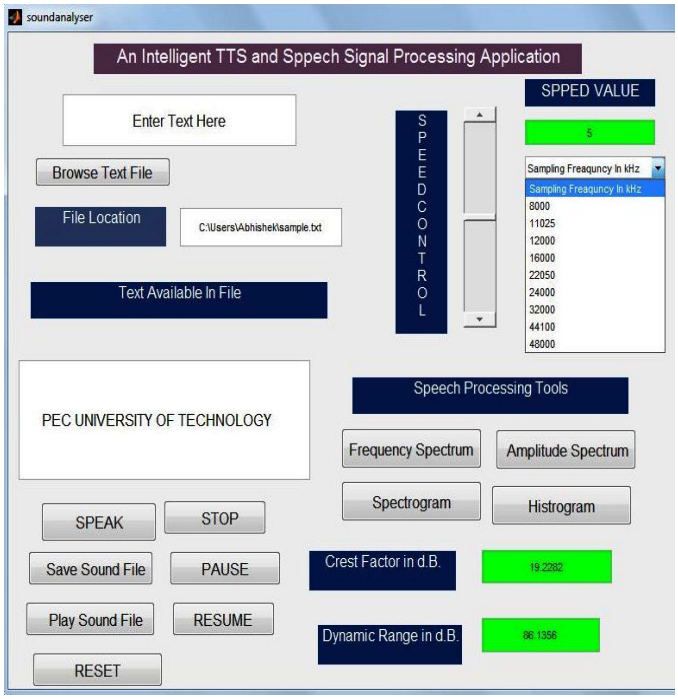

Fig 6: GUI of TTS system

5.3 Volume control

It controls the volume of speech. This tag can be nested or empty.

Attributes: level= Supports values between 0 and 100, being percentages of the system's set volume.eg. <volume level="50"/>Speak allow following text at level 50.

5.4Pitch Control

Controls the pitch at which the text is spoken. This tag can be empty or nested.

Attributes:absmiddle= Sets the absolute pitch for the speech in a range between -10 and 10 with 0 being normal speech.eg. <rate absmiddle="5"/>Speak all following text at pitch 5.

5.4 Word Emphasis

This feature is used to apply emphasis to a word or section of text. This tag must be nested.

Attributes: This tag has no attributes. eg. The telephone number is <spell>555 3468</spell>.

5.6 Forced Pronunciation

It forces the pronunciation of a word according to it usage if not correctly determined by the TTS speech engine or to override the engine. This tag must be nested.

Attributes: Part= Takes a value from noun, verb, modifier, function or interjection.eg. To <partofsp part="verb">record</partofsp>that <partofsp part="noun">record</partofsp> press 1.

5.7 Save Option

It is very powerful feature of our system. Speech signal produced can be saved.

5.8 Speech Signal Analysis

Speech signal analysis of produced speech in various domains and various other plots can be done.

5.9 Dynamic Range and Crest Factor Determination

It determines quality of speech signal in terms of crest factor and dynamic range. Crest factor is the ratio between peak (crest) level and RMS level of a wave form. Dynamic range is the ratio of the loudest sound to that of the quietest sound in a piece of equipment or a complete system, expressed in decibels (dB).

5.10. Stop and Pause feature

This feature allows the user to stop or pause the speech according to his needs.

5.11 Domain-Specific Text Filter

It will alter the raw text before it is passed on to the TTS system.eg. ‘‘Dr. Marine drove to Marine Dr.’’ is intelligibility read as ”Doctor Marine drove to Marine Drive”.

6. RESULTS

Quality of speech can be determined using crest factor and dynamic range. Ideally 120 dB is typical dynamic range [21] within which normal-hearing listener process acoustic intensity information. So in order to make the softest speech sounds audible and the loudest still comfortable, it is important to know the dynamic range for speech sounds. Crest factor is an important parameter when a voice is to be recorded or reproduced in an electro acoustic system.

Many text documents were converted into speech. It was found that crest factor is between 18 to 20 db and dynamic range is between 80 to 90 db which lies in permissible limit for audible speech [20]. Speech signal in time domain and frequency domain is represented and, its histogram and spectrograph is also plotted (see Fig.7 (a) to 7(d)).

For average long term speech spectrum (talking over one minute) maximum energy is in the 250Hz to 500Hz band and speech analysis in frequency domain (see Fig. 7(b)) is in conformation with this. These lower-frequency bands correspond to vowel sounds, the higher-frequency bands in the 2k Hz to 4k Hz region correspond to consonant sounds. Vowels carry the power of the voice and consonants provide intelligibility.

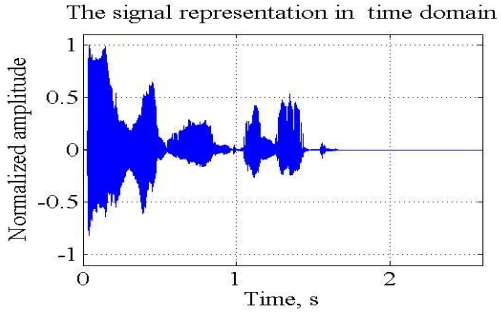

Fig 7(a): Speech Signal in time domain

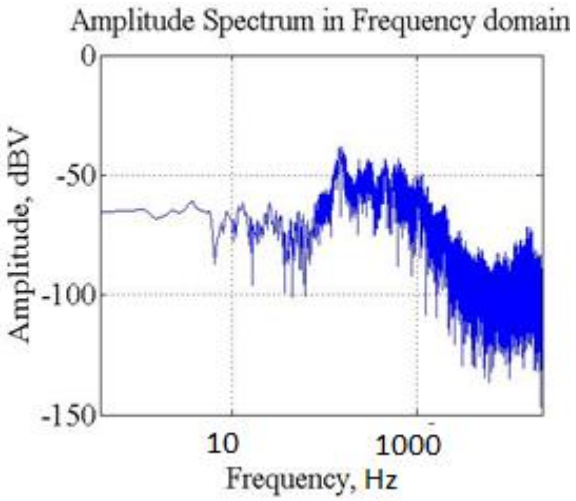

Fig 7(b): Speech Signal in frequency domain

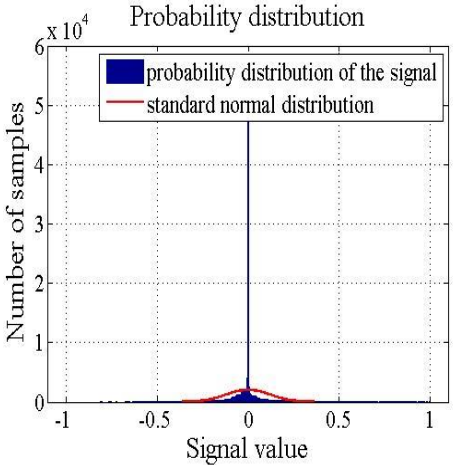

Fig 7(c): Histogram plot of the speech signal

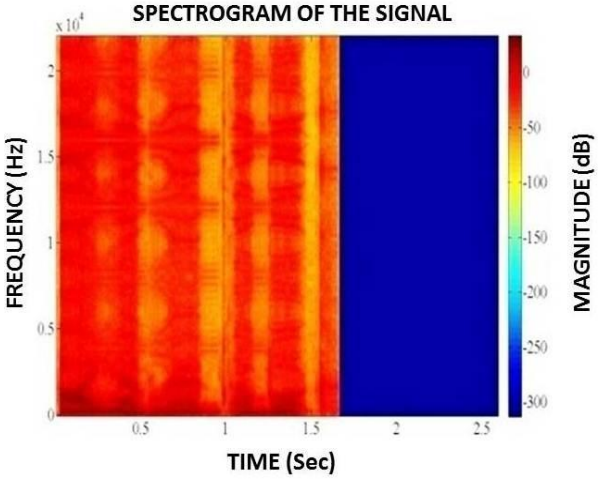

Fig 7(d):Spectrograph of the speech signal

Fig 7(a),(b),(c),(d): Various plots for speech signal

7. CONCLUSION AND FUTURE SCOPE

TTS System works efficiently for various input texts and was perfectly audible as confirmed by the values of Crest Factor and Dynamic range. This can be extended to include more voices. Even users voice can be used for speech output by modifying the program to include the database of the specified voice. This system can be extended to read pdf files. It can also be used to make applications for blind persons also, to read text from images.

8. ACKNOWLEDGMENT

We would like to thank Department of Electronics & Electrical Communication Engineering of PEC University of Technology, Chandigarh for giving us platform to do this work and providing all the support.

9. REFERENCES

[1] Tokuda et al," Speech Synthesis Based on Hidden Markov Models",Proceedings of the IEEE | Vol. 101, No. 5,pp.1234-1252 May 2013

[2] J. Hamzabegovic*, D.Kalpić "A Proposal for Development of Software to SupportSpecific Learning Difficulties",12th International Conference on Telecommunications - ConTEL 2013,pp.207-214,ISBN: 978-953-184-180-1, Zagreb, Croatia

[3] JuergenSchroeter AT&T Laboratories

[4] A. G. Ramakrishnan, Lakshmish N Kaushik, LaxmiNarayana. M, "Natural Language Processing for Tamil TTS", Proc. 3rd Language and Technology Conference, Poznan, Poland, October 5-7, 2007

[5] Chen, G.L., Yue, D.J., Zu, Y.Q., Yu, Z.L., "An embedded English synthesis approach based on speech concatenation and smoothing", ISCSLP2004, pp.157-160, Hong Kong, Dec. 2004

[6] T. Dutoit, "An Introduction to Text-to-Speech Synthesis"Dordrecht/Boston/London: Kluwer Academic Publishers, 1997.

[7] T. Styger and E. Keller, Fundamentals ofSpeech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges Formant synthesis, In Keller E. (ed.), 109-128, Chichester: John Wiley, 1994., 4,5

[8] 13. D.H. Klatt, ''Software for a cascade/parallel formant synthesizer,'' J. Acoust. Soc. Am., vol. 67, no. 3,971–995, 1980.

[9] J. Allen, M.S. Hunnicutt, and D. Klatt, From Text to Speech, The MITalk System, Cambridge: CambridgeUniversity Press, 1987

[10] Moulines, E., Charpentier, F. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Communication, Vol.9, pp.453-468, 1990

[11] Sproat, R., Hirschberg, J., Yarowsky, D., "A corpus-based synthesizer", ICSLP1992, pp.563-566, Alberta, Canada, Oct. 1992

[12] Van Santen, J., Sproat, R., Olive, J., Hirshberg, J., editors, Progress in Speech Synthesis, Springer Verlag, New York, 1995

[13] IngmundBjørkan,Speech Generation and Modification in Concatenative Speech Synthesis Ph D Thesis,Norwegian University of Science and Technology .Faculty of Information Technology, Mathematics and Electrical Engineering, Department of Electronics and Telecommunications 2010

[14] Sproat, R. and Oliver, J. "An Approach to Text-to-Speech Synthesis". Chapter 17 in book "Speech Coding and Synthesis", Elsevier, 1995

[15] S. Nakajima and H. Hamada, "Automatic generation of Synthesis Units based on context oriented clustering", Proc. ICASSP 1988, pp. 659-662, (New York, USA), 1988].

[16] R. E. Donovan and E. M. Eide, ''The IBM trainable speech synthesis system,'' in Proc. Int. Conf. Spoken Lang. Process., 1998, pp. 1703–1706.

[17] B. Beutnagel, A. Conkie, J.Schroeter, Y. Stylianou, and A. Syrdal, ''The AT&T Next-Gen TTS system,'' in Proc. Joint ASA/EAA/DAEA Meeting, 1999,pp. 15–19.

[18] G. Coorman, J. Fackrell, P. Rutten, and B. Coile, ''Segment selection in the L&H realspeak laboratory TTS system,'' in Proc. Int. Conf. Spoken Lang. Process., 2000,pp. 395–398.]

[19] http://msdn.microsoft.com/en-us/library/ms720151(v=vs.85).aspx

[20] http://msdn.microsoft.com/library/windowsphone/develop/ff402529(v=vs.105).aspx

[21] Zeng et a," Speech dynamic range for cochlear implants" . J. Acoust. Soc. Am., Vol. 111, No. 1, Pt. 1, Jan. 2002.