A Text to Speech Conversion Engine

Anubha Parashar, Vishwanath Bijalwan
{anubhaparashar1025, vishwanath.bijalwan}@gmail.com

ABSTRACT: A Text to Speech (TTS) Synthesizer is a computer application that is capable of reading out typed text. This generally involves two steps, text processing and speech generation. Speech synthesizers can be characterized by the size of the speech units they concatenate, as well as by the method used to code, store and synthesize the speech. Synthetic speech may be used in several applications like telecommunications services, language Education, aid to handicapped persons, fundamental and applied research etc. TTS has to face many challenges during the process of conversion of text to speech. The most important qualities expected from speech synthesis system are naturalness and intelligibility. In India different languages are spoken, each language being the mother tongue of tens of millions of people. While the languages and scripts are distinct from each other, the grammar and the alphabet are similar to a large extent. One common feature is that all the Indian languages are phonetic in nature.

INTRODUCTION

Speech and spoken words have always played a big role in the individual and collective lives of the people. On the contrary, the dependence of human computer interaction on written texts and images, makes the use of computers impossible for visually and physically impaired and illiterate masses. Automatic speech generation from natural language sentences can overcome these obstacles.

1.1 Introduction to Text to Speech Synthesis System

The function of text-to-speech (TTS) system is to convert an arbitrary text to a spoken waveform. This generally involves two steps, i.e., text processing and speech generation. Text processing is used to convert the given text to a sequence of synthesis units while speech generation is generation of an acoustic wave form corresponding to each of these units in the sequence [1] [2].

1.2 Applications of Text-to-Speech System

The application field of TTS is expanding fast whilst the quality of TTS systems is also increasing steadily. Speech synthesis systems are also becoming more affordable for common customers, which makes these systems more suitable for everyday use. Some uses of TTS are described below.

Aid to Vocally Handicapped

A hand-held, battery-powered synthetic speech aid can be used by vocally handicapped person to express their words. The device will have especially designed keyboard, which accepts the input, and converts into the required speech within blink of eyes.

Source of Learning for Visually Impaired

Listening is an important skill for people who are blind. Blind individuals rely on their ability to hear or listen to gain information quickly and efficiently. Students use their sense of hearing to gain information from books on tape or CD, but also to assess what is happening around them.

Games and Education

Synthesized speech can also be used in many educational institutions in field of study as well as sports. A teacher can be tired at a point of time but a computer with speech synthesizer can teach whole day with same efficiency and accuracy.

Telecommunication and Multimedia

TTS systems make it possible to access textual information over the telephone. Texts can be large databases which can hardly be read and stored as digitized speech. Queries to such information retrieval systems could be put through the user's voice (with the help of a speech recognizer), or through the telephone keyboard. Synthesized speech may also be used to speak out short text messages in mobile phones.

Man-Machine Communication

Speech synthesis may be used in several kinds of human-machine interactions. For example, in warning, alarm systems, clocks and washing machines synthesized speech may be used to give more accurate information of the current situation. Speech signals are far better than that of warning lights or buzzers as it enables to react to the signal more fast if the person is unable to get light due some obstacles.

Voice Enabled E-mail

Voice-enabled e-mail uses voice recognition and speech synthesis technologies to enable users to access their e-mail from any telephone. The subscriber dials a phone number to access a voice portal, then, to collect their e-mail messages, they press a couple of keys and, perhaps, say a phrase like "Get my e-mail." Speech synthesis software converts e-mail text to a voice message, which is played back over the phone. Voice-enabled e-mail is especially useful for mobile workers, because it makes it possible for them to access their messages easily from virtually anywhere (as long as they can get to a phone), without having to invest in expensive equipment such as laptop computers or personal digital assistants (PDAs) [3].

There are many TTS systems on Indian languages like Dhvani, Shruti, HP Lab System, Vani etc. We have used Dhvani because of its user friendly interface and Vani because it produces better results in prosody modification as parameters to control speech like volume, duration etc. is given by user. One of the key components of Text to Speech Synthesizer is prosody generator. There are basically two types of Text to Speech Synthesizer,

(i) single tone synthesizer and
(ii) multi tone synthesizer.

The basic difference between two approaches is the prosody feature. If the output of the synthesizer is required in normal form just like human conversation, then it should be added with prosody feature. The prosody feature allows the synthesizer to vary the pitch of the voice so as to generate the output in the same form as if it is actually spoken or generated by people in conversation.

Prosodic features of speech are quantity (duration), stress (intensity) and intonation(pitch). Prosody is one of the key components of Speech Synthesizers, which allows implementing complex weave of physical, phonetic effects that is being employed to express attitude, assumptions, and attention as a parallel channel in our daily speech communication. In general any communication is collection of two phases: Denotation, which represents written content or spoken content and Connotation, which represent emotional and attentional effects intended by the speaker or inferred by a listener. Prosody plays important role in guiding listener for speaker attitude towards the message, towards the listener and towards the complete communication event. From listener point of view, prosody consists of systematic perception and recovery of speaker intentions based on:[4][5]

a) Pauses: To indicate phrases and separate the two words

b) Pitch: Rate of vocal fold cycle as function of time

c) Rate: Phoneme duration and time

d) Loudness: Relative amplitude or volume.

PROBLEM AND SCOPE

Until now, in all the TTS, prosody modelling for hindi poems has not been done. We aim to perform phonetic and speech synthesis to produce enriched speech i.e. in the form of poems. The user will give text input and the output will be poetry recitation in two or three different emotions and styles , with varying duration, pitch and loudness. All the concepts relating metres, duration , intonation and emotion – styling of Hindi language are tried to be dealt with to generate the output as natural as possible.

Incorporating prosody modelling in Hindi TTS synthesis has a very wide scope, extending from high “naturalness” in poetry to domain independent text synthesizers like same TTS can be used in weather forecasting, automated railway announcement system etc.

REFERENCES

[1] Gupta, Jay Prakash, et al. "Analysis of Gait Pattern to Recognize the Human Activities." arXiv preprint arXiv:1407.4867 (2014).

[2] Bijalwan, Vishwanath, et al. "Machine learning approach for text and document mining." *arXiv preprint arXiv:1406.1580* (2014).

[3] Sati, Meenakshi, et al. "A Fault-Tolerant Mobile Computing Model Based On Scalable Replica." *IJIMAI* 2.6 (2014): 58-68.

[4] Kumar, K. Susheel, et al. "Sports video summarization using priority curve algorithm." *International Journal on Computer Science & Engineering* 2.9 (2010): 2996-3002.

[5] Bhaskar-Semwal, V., et al. "Accurate location estimation of moving object In Wireless Sensor network." *International Journal of Interactive Multimedia and Artificial Intelligence* 1.4 (2011).

[6] Bijalwan, Vishwanath, et al. "KNN based Machine Learning Approach for Text and Document Mining." *International Journal of Database Theory and Application* 7.1 (2014): 61-70.

[7] Gupta, Jay Prakash, et al. "Human activity recognition using gait pattern."*International Journal of Computer Vision and Image Processing (IJCVIP)* 3.3 (2013): 31-53.

[8] Kumari, Pinki, and Abhishek Vaish. "Brainwave's energy feature extraction using wavelet transform." *Electrical, Electronics and Computer Science (SCEECS), 2014 IEEE Students' Conference on*. IEEE, 2014.

[9] Kumari, Pinki, and Abhishek Vaish. "Brainwave based user identification system: A pilot study in robotics environment." *Robotics and Autonomous Systems* 65 (2015): 15-23.

[10] Kumari, Pinki, Santosh Kumar, and Abhishek Vaish. "Feature extraction using emprical mode decomposition for biometric system." *Signal Propagation and Computer Technology (ICSPCT), 2014 International Conference on*. IEEE, 2014.

[11] Vaish, Abhishek, and Pinki Kumari. "A Comparative Study on Machine Learning Algorithms in Emotion State Recognition Using ECG." *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012*. Springer India, 2014.