Procedia Computer Science

Volume 41, 2014, Pages 83–88

BICA 2014. 5th Annual International Conference on Biologically Inspired Cognitive Architectures

Natural Speech Synthesizer for Blind Persons Using Hybrid Approach

Mukta Gahlawat[a,b*], Amita Malik[a], Poonam Bansal[b]

[a] DeenBandhu ChotuRam University of Science & Technology Murthal, India.

[b] Maharaja Surajmal Institute of Technology, Jankpuri, New Delhi, India

Abstract

The major challenges faced by the researchers in speech synthesis are intelligibility and naturalness. Intelligibility means easily understandable and naturalness means the quality of speech being very near to human speech. Due to dynamic nature of human speech it is very difficult to mimic it, as the same content of speech in different situations is having different prosodic parameters. This paper discusses an approach to develop a natural sounding speech synthesizer. The developed Text To Speech system was tested on blind persons using subjective listening test. Test was performed using mean average score (MOS) and it was done on ten blind persons of age group varies from 14 years to 42 years. Five parameters naturalness, intelligibility, usability, localization awareness, expressions were considered for analysis of the speech synthesizer. As a result, good MOS was received for naturalness and usability, fair MOS for intelligibility and localization.

Keywords: Speech, Text to Speech, Expressive Speech, Unit selection, Concatenative Speech Synthesis

1 Introduction

Speech is the most natural way to communication between two or more persons. For effective communication expressions, clarity of speech and pronunciation play an important role to deliver the message correctly. When the speech synthesizer is developed, the researcher always tries to synthesize the speech as close as possible to human speech. Different peoples have different characteristics like pitch, prosody, accent, pronunciation etc. so it is very difficult to follow the standard speech characteristics all over the world. Even the individual's speech is full of variations depending upon his mood, biological fitness, and different state of mind. These are some reasons that justify why the

Selection and peer-review under responsibility of the Scientific Programme Committee of BICA 2014

doi:10.1016/j.procs.2014.11.088

natural sounding speech is still a state of art after having a long history of research. Speech synthesis means conversion of written text into spoken words by concatenating speech waveforms. There are number of ways of speech synthesis as discussed by (Lemmetty, 1999) in his review. First way, is the articulatory synthesis where the human vocal organs and articulation processes are modeled. Speech is created by digitally simulating the flow of air through the representation of the vocal tract. It produces high-quality synthetic speech but this technique is very hard to implement. Second technique is the Formant speech synthesis that involves an acoustic model for generating synthesized speech output. It does not use human speech samples instead there are a number of parameters which needs to be considered like fundamental frequency, voicing, and noise levels etc. This technique lacks naturalness of speech. Third method is the conatenative synthesis of speech which is considered as best for natural sounding speech synthesis because it is based on the concatenation of pre recorded segments of speech. Waveform is generated by selecting and concatenating the appropriate units from a database consisting of different types of speech units (like phones, diphones, syllables, words, phrases). Other methods like HMM based and linear prediction methods also exist in literature.

The aim of this work is to generate natural sounding speech; hence concatenative speech synthesis is implemented using unit selection algorithm (A. Hunt, 1996) (Black, 2003). For developing natural speech, a hybrid approach where the expressions and spatial parameters are unified, is used to make synthesized speech more natural. The normal vision persons can easily understand the expression of the speaker just by seeing his facial gestures but for visually impaired person it is not possible to indentify the mood or expressions of speaker. Moreover, majority of Text To Speech Synthesizer (or TTS) software's that are used by blind persons lack naturalness and expressions. Additionally, during testing one interesting input from listeners was received that this TTS system has the personalized database recorded by non-native speaker of English so they were able to understand the word more easily as compare to the software they were using in their labs. They mentioned the reason that the accent and pronunciation of words are same as they speak. The approach of adding expressions with spatial speech is purposed .This paper is organized in 6 sections, section 2 describe the related work, section 3 gives details of proposed approach , section 4 includes testing details followed by results obtained and last section include conclusion and future scope.

2 Related Work

The Speech synthesis is not new branch of research, it has a long history. Generating natural sounding speech is a big challenge of this field. When we talk about emotional speech, there are many authors who have done emotional speech synthesis using various techniques and in various emotions. (Akemi Iida, 2003) Synthesize the emotional speech by a corpus-based concatenative speech synthesis system using large emotional speech corpora. They have considered three kinds of emotions anger, joy, and sadness. They have created the corpora for Japanese language. (Daniel Erro, 2010) Designed the system which perform emotion conversion by manipulating prosody. Intonation, duration and intensity were taken as three prosody parameters. (Aimilios Chalamandaris, August 2010) implemented the unit selection technology into screen reading environments. They carried out subjective test using MOS to evaluate the resulting system. (Haojie Zhang, 2012) Synthesize the emotional speech by adjusting fundamental frequency and formant transition. (Roberto Barra-Chicote, 2010) have generated emotional speech by integrating unit selection and HMM based synthesis and found that unit selection require improvement in prosodic modeling and HMM require improvement in spectral modeling. Also there were some emotions which were not reproduced by either method. (Tonnesen & Steinmetz, 1993) Had work on synthesis of 3D speech. They described various ways to generate 3D sound, challenges for spatial sound and its applications. (Jaka Sodnikn, 2011) Designed multiple spatial sounds in hierarchical menu navigation for visually impaired computer users. They describe various benefits and drawbacks of simultaneous spatial sounds in auditory interfaces for visually impaired and blind computer users. They took two different auditory interfaces in spatial and non-spatial condition to represent the hierarchical menu structure of a simple word processing