Biologically Inspired Cognitive Architectures (2016) 15, 51–60

Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bica

RESEARCH ARTICLE

Integrating human emotions with spatial speech using optimized selection of acoustic phonetic units

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

[a] Computer Science & Engineering Department, DCRUST, Murthal, India
[b] Computer Science & Engineering Department, MSIT, C-4 Janakpuri, Delhi, India

Received 8 September 2015; received in revised form 18 September 2015; accepted 24 September 2015

KEYWORDS
Speech technology;
TTS;
Concatenative speech synthesis;
Unit selection synthesis

Abstract Synthesis of natural sounding speech is state of the art in the field of speech technology. Imitation of the dynamic human voice is required to generate this. The aim of this work is to develop and deploy a natural speech synthesizer for visually impaired persons. The synthesizer has been developed via an integrated approach of adding localization in expressive speech using a personalized speech corpus. A genetic algorithm has been implemented for optimal selection of acoustic phonetic units of speech. This concept has many applications. We tested one of those applications here in different aspects. Its performance is compared on various categories of listeners using a subjective listening test. Encouraging results on various parameters are received from visually impaired listeners.

Introduction

Speech is a major way for people to communicate. For effective communication, clear, high-quality speech is required. To attain speech similar to human beings, research has been conducted for several decades in the field of speech synthesis. Various characteristics of human voice have been used to mimic it. But this requires great effort as it is very hard to incorporate all parameters of human voice in synthesized speech. This challenge is the source of motivation for this work. Lots of researchers have tried in a number of ways to attain natural and intelligible speech. Various proposals have been used to imitate vocalizations. The best known is MBROLA (Dutoit, Pagel, Pierret, Bataille, and van der Vrecken (1996); ("MBROLA,"). It has

[*] Corresponding author. Mobile: +1 6178992636.
E-mail addresses: Mukta.gahlawat@gmail.com (M. Gahlawat), amitamalik.cse@dcrustm.org (A. Malik), pbansal89@yahoo.co.in (P. Bansal).

http://dx.doi.org/10.1016/j.bica.2015.09.008
2212-683X/© 2015 Elsevier B.V. All rights reserved.

been used internationally to develop a multilingual Text To Speech Synthesis System. Professionals of different countries and languages worked in collaboration to generate high quality diphone based speech using a time domain algorithm. A time domain algorithm requires less work for computation, so this has been used for modifying speech prosody. The Pitch Synchronization Overlap and Add algorithm (PSOLA) (Moulines & Charpentier, 1990) modifies the pitch and duration of speech signals. Small overlapping segments are formed from the waveforms of speech. These segments are moved closer to increase pitch and moved apart to decrease the pitch. In order to increase or decrease the duration of signal the segments are replicated or removed respectively. It is also used in voice conversion and synthesis of expressive speech using neutral speech by modifying prosodic characteristics (Valbret, Moulines, & Tubach, 1992). But PSOLA has no control over formant features. Natural speech synthesis can be done by modifying some combination of prosodic features such as line formants and pitch (Erro, Navas, Hernáez, & Saratxaga, 2010).

One important parameter that influences generation of natural speech is expression. Prominent methods used for embedding expression involve changing the prosody features and corpus base. Generation of emotional speech can be accomplished by manipulating prosodic features of neutral speech (Govind, Smily, Biju, & Binilkumar, 2014) or conversion of emotion by modifying various prosody features (Erro et al., 2010). Emotional speech has been also synthesized by doing adjustment in F0 (fundamental frequency) and transitions in formant (Zhang & Yang, 2012). Another method involves concatenation of pre-recorded audio files by selecting the most appropriate unit from a speech corpus (Iida, Campbell, Higuchi, & Yasumura, 2003). This uses a unit selection algorithm (Black, 2003) to select the best annotated speech unit from the database. This method has been used to develop screen readers (Chalamandaris, Karabetsos, Tsiakoulis, & Raptis, 2010). Speech synthesis can be based on statistical parameters for generating intelligible speech (Zen, Tokuda, & Black, 2009), for instance, Hidden Markov Model (HMM) based synthesis (Zen et al., 2007). A hybrid approach can be used for further improvement (Barra-Chicote, Jamagishi, King, Montero, & Macias-Guarasa, 2010). The benefits of HMM and unit selection can be put together to generate better speech quality (Taylor, 2006). Emotions are not only limited to text to speech but emotions have been considered for analyzing affect based on text (Ptaszynski et al., 2013) also.

There are some parameters that are imperative for natural speech generation but are rarely found in the literature. This includes addition of spatial positioning to increase the naturalness of speech. Research and development for spatial sounds are not so commonly found in speech synthesis. Some of the milestones of this field are cited here. Tonnesen & Steinmetz worked for three dimensional speech synthesis and suggested ways to produce spatial sound, its application and challenges associated with it. A U.S. Patent was filled by Moore and Farrett (1996) for 3D Speech Synthesis. Sodnik and Tomažič (2010) developed 3D Text To Speech System in Java. Sodnik, Jakus, and Tomazic (2010) developed an application for blind persons using spatial sounds coming from multiple directions. But they got degraded performance from multiple spatial signals as compared to non-spatial sounds. 3D spatial sounds have much more potential that require profound exploration in different circumstances. This paper discusses one such approach wherein the single spatial sound is integrated with expressive speech along with some additional parameters.

Most of the work towards natural vocalization is mainly mono parametric which aims to develop the speech either by modifying a parameter such as pitch, fundamental frequency, accent or other prosodic features or by generating emotional speech or by developing spatial sounds. But the present research is an attempt to apply multiple human characteristics to make synthesized speech more natural. Expressions, dynamic positioning of speakers in three dimensional spaces, accent and a personalized speech corpus have been collectively used to produce natural, intelligible and vibrant speech. The concept has different applications in areas like animated movies, theater shows for blind persons and many more. An expressive audio visual story is deployed and a responsiveness analysis has been done for different aspects of it.

System design

This developed natural speech synthesizer converts plain text to emotionally rich and spatial audio in local narrator intonation. Four parameters, i.e. expression, dynamic positioning of speaker in three dimensional space, local accent and personalized speech corpus are implemented. Fig. 1 illustrates a block diagram of the proposed system.

Recording of textual content

The first step is to select the relevant textual content for the specific domain. The domain chosen here is classroom learning for visually impaired students. All the words and sentences that can be used in the class room environment are selected. The selected material has been recorded using Audacity ("Audacity,") and files are recorded as .wav files. The units for recording taken are words and sentences. Various options of speakers are available including speakers of different gender, different dialectal backgrounds and different ages. A non-professional non-native female speaker was chosen for the recording of content in the English language under studio recording conditions.

Annotations of speech

The process of annotation involves the segmentation, labeling of speech and phonetic transcription. Both manual and automatic segmentation have been performed on the speech corpus. The wavesurfer tool ("WaveSurfer,") was used for manual segmentation. It is very time consuming and demands lots of manual effort. Due to the fact that manual segmentation is most precise this has been used