Review of Modern Speech Synthesis

Cheng Xian-Yi and Pan Yan

College of Computer Science, Nantong University, Nantong Jiangsu 226019, China

Abstract. Speech synthesis and recognition technology has become a research focus in the field of intelligent computer. For nearly 50 years of study, speech synthesis has had huge development as an interdisciplinary, this paper reviewed the modern speech synthesis technology and research, analyzes the problems existing in the research in this field.

Keywords: Speech synthesis, Based on the semantic speech synthesis, Chinese speech synthesis.

1 Introduction

Speech synthesis is an important part of man-machine speech communication, which is called text to speech (TTS). It involves acoustical, linguistics, digital signal processing, computer science, etc. It is a frontier technology of Chinese information processing field [1]. Speech synthesis technologies give machine the function of "artificial mouth", working out how to let the machine talk like a man [2]. As early as 200 years ago, people began to study speech synthesis, with the development of the modern computer technology and digital signal processing technology speech synthesis technology has developed.

In 1939, Dudley Homer exhibited his speech synthesizer in New York world expo, called "Parallel Bandpass Vocoder "[3]. In 1960, the Swedish linguists and words engineers G.Fant introduced the speech production theory systematically in "Acoustic Theory of Speech Production", which promoted the development of the speech synthetic technology. Since 1970s, linear prediction technology started to use in speech coding and recognition [4].

In 1973, Holmes made parallel formant synthesizer. In 1980 Klatt designed strings/parallel hybrid formant synthesizer [5]. These two synthesizers can synthesize natural language by adjusting the parameters. DECTALK of DEC of the United States was most representative in1987 [6]. This system uses Klatt's string/parallel formant synthesizer. It can provide all kinds of speech information services through standard interface, computer networking and separately receiving telephone network. Its pronunciation was vivid and it can produce 7 different tone of voice.

In recent years, a new speech synthesis method based on database is aroused people's attention. In this approach, the phonetic unit of synthetic statements are selected from an advance recorded of huge speech database, as long as speech database is enough big, including various possible speech units, it can stitch any statements.

W. Hu (Ed.): Electronics and Signal Processing, LNEE 97, pp. 517–524.
springerlink.com

Synthetic speech elements are from natural original pronunciation, the intelligibility and naturalness of speech statements will be very high [7].

2 Speech Synthesis Technology

2.1 Traditional Speech Synthesis Technology

Formant synthesis. Traditionally, the pole of track transmission frequency response is called formants, and the distribution characteristics of the resonant frequency (poles frequency) determine the timbre. It has the following three practical models [8].

①Cascade type formants model. In this model, the track is considered a group of series of second-order resonator. This model is mainly used in the synthesis of vowels.

②Parallel type formants model. Many researchers believe that the nasal vowels and other vowels and most consonants non-general, cascade model can not be described and simulated. Therefore, the parallel type formants model is produced.

③Mixed formants model., Formant filter first connected with tail in a cascade type formant model. Before amplitude adjustment the input signal through added to every resonant filter and superposed the output. Cascade type fulfills the acoustic theory speech production for the speech of synthetic source located at the end of track. Parallel model is more appropriate for the speech of synthetic source located at the middle of track, but its amplitude adjustment is very complex. So people combined both of them, mixed type formants model is proposed. It is an accurate simulation based on the track. It can be synthesized pronunciation of higher naturalness.

LPC rules synthesis LPC. Rules synthesis belongs to poles digital filter model of the whole linear source- channels speech production model. LPC rules synthesis technology is essentially encoding technology of time waveform. The synthesis process is essentially a simple decoding and stitching process. The advantages of LPC rules synthesis technology is simple, intuitive. Because speech of natural language and isolated streams has a great distinction, if simply put each isolated speech splicing together stiffly, the quality of the whole language will be very good. Therefore, the effect of whole continuous language of linear forecasting parameters synthesis is not good. It must combine with other technologies that can improve speech synthesis quality significantly.

PS0LA splicing synthesis. PSOLA is a kind of algorithm of modifying the rhythm of the synthesized speech, which was used for waveform edit synthesized speech technology. Main time-domain parameters of deciding speech waveforms rhythm included duration, intensity of a sound and pitch. According to the different methods of improving parameters, it mainly divided into LP-PSOLA, TD-PSOLA and FD-PSOLA. Its main characteristics are: Before stitching the requirements of speech waveforms PSOLA algorithm adjusted the prosodic feature of stitching unit according to the context requirements. The synthetic waveform maintains the main sound segment of the pronunciation and prosodic feature of stitching unit accord with context