International Journal of Advanced Engineering and Management Research

Vol. 6, No. 06; 2021

ISSN: 2456-3676

SPEECH SYNTHESIS BASED ON WAVEFORMS

Chih-Hsu Hsu
Department of Applied Healthcare Informatics, Ching Kuo Institute of Management and Health, No.336, Fu Hsin Rd., Keelung 203, Taiwan, R.O.C.,

Abstract

In this paper, we discuss a type of synthesis by compilation of speech waveforms. In order to synthesize any articles, we need control speech waveforms with pitch patterns. We proposed a novel method, which were features of the fourth tone with wide pitch frequency, to obtain the various pitch periods of natural speech waveforms, instead of interpolation or decimation of natural speech waveforms. Thus, we took a pitch period of natural speech waveforms as a basic unit, combining with synthesis rules of rhythm of natural speech, to synthesize sentences. In order to synthesize a complete sentence, we divided it into four steps. The first step of synthesis is a word, the second is a phrase, the third is an inspiration, finally synthesizes a complete sentence. Synthesis of words includes (i) vowels, (ii) diphthongs, (iii) consonant combined with vowels. The processing of vowels, we utilized the features of the fourth tone with wide range of pitch frequency, and retained all of the periodical speech waveforms for basically syntactic units. The processing of diphthongs, we preserved the parts of acutely changed waveforms in the middle, then processed others as vowels. The processing of consonant combined with vowels, we need retain consonants and a little parts of adjacent waveform of vowels, then imitate others as the above two classes. The results of listening tests showed that it could really synthesize sentences with fluency.

Keywords: speech synthesis, waveforms, pitch patterns, tones.

1. Introduction

Speech synthesis technology allows humans to communicate with computers using language. After decades of continuous efforts, this technology has matured and is generally divided into waveform compilation and parameter synthesis. We proposed a novel method, which were features of the fourth tone with wide pitch frequency, to obtain the various pitch periods of natural speech waveforms, instead of interpolation or decimation of natural speech waveforms.

Generally speaking, the way to synthesize speech artificially without going through the body's articulatory organs can be called speech synthesis. However, if the continuous speech produced by people is recorded and then played out intact, it cannot be called a speech synthesis system. Fast Fourier transform algorithms and cepstrum were applied to deal with speech problems, making digital signal processing technology a big step forward.

The compilation of waveforms is to individually record the words needed to form the text, and then synthesize them by voice according to the message to be transmitted after compilation. The compilation method is based on the voice waveform, and the characters must be matched properly to make the synthesized voice sound fluency, which is quite difficult to achieve.

Vol. 6, No. 06; 2021

ISSN: 2456-3676

Recently, most of the speech synthesis systems that have entered the practical stage use the compilation of speech waveform, such as telephone reservation systems and voice query service systems. However, with this kind of synthesis method, the sentence must be simple or the number of words is small, the synthesis effect is better, and it is less prone to problems. If you want to synthesize arbitrary sentences, consider the basic unit of synthesis, which can be divided into three situations: sentences, words, and periodic waves. When synthesizing arbitrary sentences by sentence as a unit, you only have to record all the sentences in advance, so the synthesized speech quality will be very good, but the memory occupied by the database will also be huge.

For Chinese speech, the length of a single-syllable speech waveform has different lengths of periodic waves depending on the characteristics of the four tones. Therefore, this paper proposed to retain various periodic waves of different speech as the basic synthesis unit, using regular synthesis. Baseband mode, we can synthesize natural and fluent speech. The Chinese speech synthesis rules are the fundamental frequency mode of four tones. Because the fundamental frequency modes of four tones have different sound length, pitch and volume according to the four tones, we only need to effectively use the fundamental frequency modes to control the speech waveforms. Thus, we could synthesize more natural speech.

2. Method

2.1 The Structure of Chinese Syllables

The structure of Chinese syllables are generally divided into three parts: consonants, vowels, and lexical tones. Table 1 shows the structure of the Chinese syllables. The finals can be divided into semivowels, main vowels and rhyming tails. The brackets are optional, but indispensable are the main vowels. Lexical tones are the characteristics of Chinese speech, and different tones have different distinguishing functions. Lexical tones refers to the tones of the rising and falling of the speech. Words have word tones, and each has its own tones. Tones refer to the relative pitch of a person's pronunciation. Pitch is an important factor in explaining the value of the tone; and because the tone is also related to the relatively long-term problem of a person's pronunciation, the length of the tone is also important to explain the value of the tone.

Table 1. The structure of the Chinese syllables

Consonant	Vowel		
(Consonant)	(Semivowel)	Main Vowel	(Rhyming Tail)

For the pronunciation of vowels, when the airflow passes through the vocal cords, the glottis will be closed tightly, making the airflow vibrate the vocal cords and produce sound, because their pronunciation, from the beginning to the end, regardless of the length of the sound, the pronunciation conditions remain unchanged. The pronunciation of diphthongs is formed by the close combination of the phonemes of two single vowels. Their loudness is from large to small. The previous phoneme has a larger volume and a longer time. Therefore, the front phoneme is called the main vowel, and the latter phoneme is called the rhyme end. Consonants are sounds

Vol. 6, No. 06; 2021

ISSN: 2456-3676

caused by the obstruction of two parts of the airflow from the lungs in the sound channel. The vocal cords do not vibrate during pronunciation, and the consonants emitted are noise, called unvoiced. The vocal cords vibrate during pronunciation, and the consonants emitted are a mixture of music and noise, called voiced sounds.

2.2 Synthesis Rule of Words

The characteristics of Chinese speech are one character and one sound. The structure of the Chinese syllables had been briefly described above. This section will introduce the three parts of tones, vowels and consonants, and then consider the parts of tone length and volume. According to the results of the experiment, we could control the sound of the output word at will, but we cannot change the timbre.

From the perspective of a complete four-tone fundamental frequency pattern, the third tones are the longest and the fourth is the shortest, but when the third tones appear in a sentence, they often only have the first half of the sentence. Use this fundamental frequency mode to synthesize four-tone speech.

Refer to the fundamental frequency mode of the four tones, the fourth tones cover almost all frequencies of the original sound, so the speech waveforms of various periods in the fourth tones are retained as the basic periodic waves, and then the fundamental frequency mode is used to control these speech periodic waves. Edit and synthesize a four-tone voice. We considered the change in the fundamental frequency pattern in the sentence, we need more frequency speech waveforms, just record more fourth tones, and then extract the required speech periodic waves from it, and we could synthesize the required speech.

The calculation of the period adopts the autocorrelation function, which is obtained separately with the center clipping function and the three-level center clipping function.

The autocorrelation function is defined as

$$\emptyset(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k) \tag{1}$$

3. Results

In terms of the synthesis of words, according to the speech waveform of the original sound, it can be divided into two categories: periodic waves and non-periodic waves. The unvoiced sounds in the initials are non-periodic speech waveforms, while the voiced sounds and vowels have periodic waves, which belong to the speech waveforms of periodic waves. The various speech periodic waves are retained, and continuously changing voice waveforms will have several waveforms with the same period but with large changes. If only a few periodic waves are used to control this voice waveform, there will be a sense of discontinuity and must be adjusted. The processing of consonant combined with vowels, we need retain consonants and a little parts of adjacent waveform of vowels, then imitate others as the above two classes. Use the

Vol. 6, No. 06; 2021

ISSN: 2456-3676

fundamental frequency mode to control the pitch, length, and volume of the voice to synthesize a natural and fluent speech.

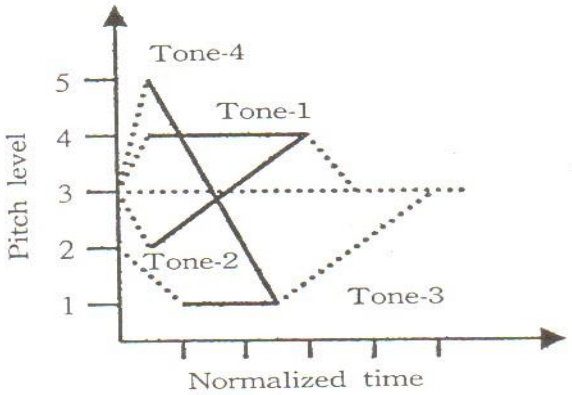

Figure 1. The pitch contours of four tones

Figure 1 shows the "pitch" and "pitch length" informations of the pitch contours of four tones.

From the four-tone fundamental frequency modes, the pitch length of the four tones can be roughly estimated, of which the third tones are the longest, the first and second tones are the second, and the fourth tones are the shortest. If it is consonant combined with vowel, the sound length is longer. In the structure of the Chinese syllabes, apart from the problems of pitch and length of tones, there are also energy problems. Among the four tones, the fourth tones have the highest energy, the first and second tones are the second, and the third tones have the lowest energy. Each word itself also has an energy problem, the volume at the beginning and the end is low, and the volume in the middle is higher. We adjust the volume of the reserved speech waveform, adjust the entire waveform at the same time, and use each periodic wave as the adjustment unit for the periodic wave. Since the periodic waves are extracted from the fourth sound, when synthesizing the word, the fourth sound only adjusts the periodic waves at the beginning and the end, and decreases or increases with a simple linear function. The volume of the first sound is reduced by about 10%, and the volume of the third sound is reduced by about 20%. In addition to the 10% reduction of the second sound, the volume of the second half decreases exponentially, because the periodic wave in this part happens to have the highest volume of the fourth sound.

Because the frequency range required to synthesize a sentence is wider, and more periodic waves of speech are retained, a sound with a wider frequency range must be emitted. The original method of extracting various periodic waves from the fourth sound may not be enough. Since there is no change in the fundamental frequency of the first sound, the second, third, and fourth tones from low to high are recorded. In order to effectively grasp the period of the periodic waves, try not to be affected by the period of the subsequent voice signal to avoid the calculation error of the autocorrelation function. The pronunciation should be slow and long, so that the periodic changes are slower, and more accurate periodic waves can be captured.

Vol. 6, No. 06; 2021

ISSN: 2456-3676

4. Discussion

Regarding the compilation method of waveforms, this paper proposed a novel method, which were features of the fourth tone with wide pitch frequency, to obtain the various pitch periods of natural speech waveforms, instead of interpolation or decimation of natural speech waveforms. It could retain various periodic waves of different speech as the basic synthesis unit to synthesize natural and fluent speech. It is necessary to use the fundamental frequency mode to control the speech waveform, and cooperate with the synthesis rules of natural speech. The so-called fundamental frequency mode of natural speech is also the fundamental frequency mode of four tones. Because the fundamental frequency mode of four tones has different pitch, pitch and volume according to the four tones, we only need to effectively use the fundamental frequency mode of four tones to control the speech waveform.

We took a pitch period of natural speech waveforms as a basic unit, combining with synthesis rules of rhythm of natural speech, to synthesize sentences. In order to synthesize a complete sentence, we divided it into four steps. The first step of synthesis is a word, the second is a phrase, the third is an inspiration, finally synthesizes a complete sentence. The processing of vowels, we utilized the features of the fourth tone with wide range of pitch frequency, and retained all of the periodical speech waveforms for basically syntactic units. The processing of diphthongs, we preserved the parts of acutely changed waveforms in the middle, then processed others as vowels. The processing of consonant combined with vowels, we need retain consonants and a little parts of adjacent waveform of vowels, then imitate others as the above two classes. As long as the fundamental frequency mode of natural speech is used to control these periodic waves, fluent speech can be synthesized.

References

Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*, Prentice-Hall.

Hsieh, C. T., and Imai, S. (1988). *A Pitch Pattern Model of Continuous Speech of Chinese* (pp. 1101-1117) Trans. on IEICE A Vol. J71-A No. 5.

Lee, L. S., Tseng, C. Y. and Ouh-Young, M. (1989). *The Synthesis Rules in Chinese Text-to-Speech System* (pp. 1309-1320). IEEE Trans. On ASSP, Vol. 37, No. 9.

Chen, W. C., Hsieh, C. T., and Hsu, C. H. (2007). *Two-stage vector quantization based multi-band models for speaker identification* (pp.2336-2341). In Proceedings of International Conference on Convergence Information Technology, Gyeongju, Korea.

Luo, T. H. and Hsu, C. H. (2014). *Signal Analysis for "Kagaya Miyamoto Shiki" Music Therapy* (pp. 283-284). In Proceedings of International Scientific Conference on Engineering and Applied Sciences, Singapore.

Luo, T. H. and Hsu, C. H. (2018). *Attitudes towards Mental Health Nursing Students* (pp.718-719). Int. Conf. of Asian Conference on Engineering and Natural Sciences, Osaka, Japan.