ATLANTIS PRESS

Advances in Social Science, Education and Humanities Research, volume 196

2nd International Conference on Social Science, Public Health and Education (SSPHE 2018)

Chinese Audio Book Platform Design Based on Time Domain Analysis Technology

Xiang Bi
School of Computer Science
Wuhan University
Wuhan, China
250825180@qq.com

Cunchen Tang*
School of Computer Science
Wuhan University
Wuhan, China
cctang@whu.edu.cn

Yue Tang
Technology R&D Department
Wuhan Core Technology Co., Ltd.
Wuhan, China
595806988@qq.com

Abstract—With the advancement of building a learning society in the new era, nationwide reading has become a way of life today. The fast-paced and highly efficient "fragmented" reading has failed to meet the needs of the audience. The "acoustic" life is beginning, and a large number of audio books are flocking to the digital reading market, resulting in a sharp increase in jobs related to professional voice service. Using speech synthesis technology to achieve audio reading is a convenient, efficient, and low-cost way nowadays. This article focuses on the research of Chinese speech synthesis technology, explores the law of audio frequency change through time domain analysis technology and designs the audiobook platform to solve the problem of poor naturalness, weak fluency, and single mode in speech synthesis.

Keywords—Time Domain Analysis, Speech Synthesis Technology, Naturalness and Fluency, Chinese Audio Books, Platform Design

I. INTRODUCTION

After entering the second decade of the 21st century, building a learning society based on the "Internet+" framework has been steadily progressing. The scale of reading activities for the entire nation continues to expand, content is constantly enriched, methods are continuously innovated, and the impact is also increasing. The heavy work pressure and busy life rhythm brought new changes to the reading styles of modern people. People are no longer spending time in bookstores or libraries as they did in the 1980s and 1990s-holding a thick book, reading it alone, and giving a review. Nowadays, they prefer relying on smart devices to read a key or general content for obtaining knowledge. As a wide array of goods like "Micro" reading, handheld reading, fragmentary reading pop up through the Internet, the audience passively accepts these promotional and catapult-like information, but they actively read it, and sometimes they can learn from it. However, these readings also brought with it some drawbacks. Those who only read the title and have the wrong interpretation of content start to spread rumors. People only focus on contradictions and conflicts for fun rather than finding solutions, which lays pressure on public opinion. In addition to the gradual emergence of the media environment, readers who read via WeChat, Weibo, and other apps can quickly and widely forward the information so that it can be read by wider range of audience. The way of "fast food" reading makes in-depth reading unnecessary, and the effect of fast learning is not so good. As a result, audio books have been promoted as a new product.

Audio books (AB: audio books) refer to audio products that use computers, mobile terminals, optical discs, tapes, flash memories, and other media as the main carrier to store or play sounds. In this process, we use the way of listening to reach the goal of reading. In the 20th century, audio books were first produced in the United States, ensuring that audience had a transition from two-dimensional reading to three-dimensional reading. Combining words or pictures with the description of the content that the author needs to express, and adding them into speech can help audience liberating their vision by "listening to the book", which is no longer limited to the interaction of the hand and the eye. Therefore, you can engage in work and business while listening to a book. At the same time, you can also make full use of your leisure time, enrich yourself by learning under any circumstances, and enjoy the "listening" process. However, new problems arose. Audio books need readers or narrators to delivered information in vocal language. A large number of audio books needed to undergo artificial reading, digital recording, post-synthesis, uploading, and transmission. The increase in labor costs makes the development of books in the "Internet+" era lag behind. It also makes some newly-launched books wait several months before they enter the "sound reading" market, which increases the circulation period of the books so that information cannot be transferred quickly and effectively. Of course, in recent years, with the gradual emergence of speech synthesis softwares and text-to-speech technology, e-books have entered the state of audio reading. Due to the use of electronic synthesized speech technology, the cognition and language expression of words are not as realistic or natural as real readers, resulting in almost the same phonetic tone, rhythm pause, and stress rhythm after the synthesis of different textual contents. Listeners have the same feeling and charm and audibility of Chinese language are lost. More and more audiences are reluctant to accept audio books in this way, which affects the further development of audio reading.

For these problems brought about by audio synthesis technology to audio books, scientists at home and abroad have started to study this. Germany and Japan have studied this technology earlier, from the study of mechanical synthesis of speech to the development of digital speech synthesis technology, and the use of digital formats, waveform splicing and synthesis techniques to complete the audio-visual translation of audio books. In the 1980s, China began to engage in speech to speech (TTS: Text to Speech) and speech recognition (ASR: Automatic Speech Recognition) technologies. The Institute of Acoustics, Chinese University of Science and Technology, and University of Science and Technology Co., Ltd. were the first institutions to study speech synthesis. In 1998, the National Library submitted an application to the Ministry of

Culture. The State initiated the "China Digital Library Project" and started the preparation and construction of the China Digital Library Project. This opened up the development of audio books in China [1]. In 2013, speech synthesis (TTS) technology and speech recognition (ASR) technology began to be applied to the audiobook platform.

II. Review of the Development Status of Audiobooks and Speech Synthesis

As the most popular and convenient way of reading and learning, nowadays audio books have quietly entered ordinary people. There are 300 audio book platforms based on mobile Internet or APP software. According to the "Analysis and Development Prospects of China's Audio Book Industry in 2018-2024" published in 2017, nearly 70% of digital readers used "listening" (Audiobooks) in 2016, and monthly active users reached 10 million, 65.3% of whom are willing to pay, and users who use more than 10 times per month account for 24.2% of the total. The market has great potential, and it has become a new mode of knowledge service [2].The speech synthesis technology is mainly based on the electronic speech synthesis accomplished by the computer information system. Through the analysis of texts and the adjustment of the content, a technical model conforming to the Chinese pronunciation rules is established, and then certain algorithm rules are applied according to the rhythm characteristics of Chinese language, and the speech is stitched. Currently, there are two types of Chinese phonetic splicing methods, encoding splicing technology and waveform splicing technology. In addition, highly natural audio composite content generated from a large number of corpora as the original support material is currently used more widely.

A. *Analysis of the Status Quo of Audio Books and Their Problems*

According to the data in the 2016 Digital Reading White Paper, the 2016 China audio reading market grew by 48.3% with the total volume of ￥2.91 billion. According to a survey conducted by the China Press and Publication Research Institute, the listening rate of adults in our country reached 17%, and the per capita listening consumption was ￥6.81 [3]. There have been more than 200 mobile platforms with listening functions in China. The well-known leader in audiobook platforms such as the Himalayas FM and Handan FM has emerged and a competitive landscape has taken shape. However, as the market share continues to increase, audio books are faced with the copyright issue of source text acquisition and sound acquisition, which endures a large increase in the cost of labor, facing the risk of reshuffling the digital publishing industry in China based on policy orientation and supply and demand.

1) Development status of audiobooks: From the perspective of digital publishing industry development, audio books are a combination of the digital publishing industry chain and the information technology industry chain. With the help of the Internet information platform and relying on the book editing and publishing industry, it satisfies the audience's needs for higher levels of reading and realizes optimal allocation and establishment of new production mechanisms. At present, the mainstream text reading websites in China are all transforming into audio books. Tianfang Listening Net, founded in 2004, is the first audiobook platform established in China based on internet technology. In November 2016, JD.com launched the "Co-Reading Project", and readers took the opportunity to complete the birth of an audiobook. The participants of this program were nearly 60,000 people. They accumulated a large volume of sound material for JD books and increased public participation in reading. In September 2017, Dangdang.com announced that the "listen to listen to books" project was formally launched and took the first step toward challenging the "future industry". At the same time, Dangdang also announced that the listening industry will gradually establish itself as a strategic core product. In October 2017, the international digital book sales brand Amazon platform launched the full range of Kindle products, all with an audio reading function. From the earliest launched listening program-Tianfang -to the gradual maturity of different kinds of FM; from relying on the PC browser web page to building a mobile terminal APP, to WeChat reading, we can know that it is the flash of ideas or the innovation of technology that has promoted the development of audio books in China from the popularity of the service objects to the customization of private individuals. It has also promoted readers to read network literature rather than paper books. Some readers even promote storytelling via we-media. This is also the change in the form and function of audio books after the expansion of the media.

2) Development bottleneck and analysis of audio books: From the day when the audiobook reading platform was established, many problems such as content production, quality of reading materials, copyright regulations, and profit model were accompanied. Over time, issues such as copyright management, diversification, and optimization of cooperation models based on social science research such as law and economic management have gradually begun to be solved and have been more effective. After China's audiobooks reached the peak development period in 2016, the most obvious problem was the quality of audio books. Audio is the main element of audio books . The problem was mainly caused by the following three aspects: First, converting content into sound files is mostly manual recording and compositing. The labor cost is relatively high. Second, the level of recording personnel is uneven, and the quality of recording equipment varies. Third, the audience of audiobooks is responsible for the content of the work. Single recording way or content can not meet the difficult needs and situations.

At present, some mobile APP terminals have also adopted speech recognition and speech synthesis technology to produce audiobooks. After a reader completes the selection of the books, the "literary and linguistic transformation" and audio files are formed through the background technology, and the reader can download the promoted audio. This method makes up for the higher cost of labor, recording level, and the quality of recording equipment, but for readers, due to the use of electronic synthesis technology, they can only have the same feeling toward words rather than the charm of Chinese language. It cannot meet the diversification of sound effects. At the same time, due to the limited reading materials involved in the platform, it is difficult

for readers to read for whatever they want, which limits the breadth of reading.

B. Overview of Speech Synthesis Technology

Since voice synthesis technology entered China in the 1980s, it has been developing rapidly. Especially in the year of 2017, the first year of artificial intelligence, speech synthesis technology to complete the machine translation, text reading, spoken language output and other functions are continuously refreshing, creating a number of unknown possibilities.

1) Digital formant speech synthesis technology: The so-called digital formant voice synthesizing technology realizes the voice output through hardware. The resonators of inductance and capacitance are constructed by digital circuits or analog circuits, and the sound output is achieved according to the difference between the circuit-controlled resonant peak resonators and the frequency. Klatt, a speech and hearing scientist in the United States, completed the synthesis of seven different sounds in 1980 and defined roles for these seven sounds [4]. Although this synthesis technology was at the top international level at that time and it was promoted on the market, there were problems such as dullness, dullness, and poor naturalness. Therefore, the technology is currently used less frequently.

2) Waveform mosaic synthesis technology: In order to solve the drawbacks of the digital formant technology, some "outsourced" voice products appear, which is based on waveform splicing speech synthesis technology, also known as waveform overlay technology. This technical principle is mainly to use the existing waveform material to delete or insert in a complete cycle, without changing the timbre and pitch of the voice. Through the overlap between the start and end points of the waveform between words, the initial suffix is realized. This method is widely used in the text-to-speech system of French, German, and Japanese. In Chinese, the method of splicing and synthesizing waveforms is mainly based on the natural spelling of Pinyin. The combination of initials and vowels forms the Chinese character waveforms.

Another way to increase the naturalness of the waveform splicing synthesis technology is to use a large-scale voice library to complete the waveform splicing synthesis. That is, a large amount of real human voices are pre-recorded, a speech corpus system is established, and the corresponding content in the sound library is called as a splicing element during synthesis, and speech synthesis is realized according to the text content. Through a large number of synthetic experiments, it is proved that the number of splicing is small and the quality of speech synthesis is better. At the same time, a speech library with more than 10 hours can satisfy the synthesis of daily sentences. At present, the mature technology is the CHATR technology developed by ATRI (Advanced Telecommunications Research Institute) in Japan [4]. Waveform mosaic synthesis technology is currently the mainstream speech synthesis technology. In January 2018, a documentary titled "Innovative China" was broadcast on CCTV 9 sets. The result was very good. It introduced the R&D history, thinking model and research and development of cutting-edge technologies based on China's independent intellectual property rights. The interpretation of the documentary is a late voice actor Li Yi, who died in July 2013. His once-voiced works were preserved forever. The authors of the documentary made an analysis of his works, and realized the restoration of his original soundtrack in this documentary using the waveform mosaic synthesis technology of large-scale speech libraries in speech recognition and speech synthesis.

III. Time Domain Analysis Technology Architecture Chinese Audio Book Platform

After many years of development of audio books, the use of speech synthesis to achieve “listening” has become no problem. Especially in the face of some open-source online speech synthesis technologies, the audio reading experience is no longer limited to the APP or website platform, and the electronic reading content can be freely selected. Voice reading can be formed through simple operations. However, the naturalness, fluency, and pleasing feelings need to be improved, and some paper media readings need to obtain an electronic version in order to achieve online “listening” and cannot satisfy the audience’s freedom of choice in multi-formation. Focusing audio time domain analysis technology aims to solve the problem of difficulties in accessing to audio books, less content, and poor quality through algorithm analysis of the broadcast reading features of Chinese speech and the inherent information carried in the audio content itself.

A. Time Domain Analysis Speech Synthesis Technology

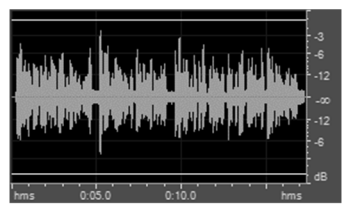


Fig. 1. Audio waveforms, time, volume coordinates

The sound is generated by vibration, and the transmission of sound is based on media. The time domain is an index that can measure the time characteristics of audio relative to the frequency domain. The time domain is derived from the mathematical concept, which means that time is the axis and time is the benchmark. In the audio, the length of the waveform is the basic display mode. In the audio playback process, the main function of the waveform is to visually display the form of the sound after being converted into an electrical signal through vibration. The most natural form of vibration is also called a sine wave. The sine wave has two parameters - one is the frequency and the other is the amplitude. The frequency represents pitch of the sound, and the amplitude represents volume. According to experimental calculations, the sine wave range that people can normally sense is between 20 Hz (low frequency) and 20,000 Hz (high frequency). In other words, the audio synthesis frequency of an audiobook is controlled within the range and can be perceived by people. Therefore, the audio volume change can be determined by the waveform, and the audio waveform can be judged by the relationship between time and volume. (Fig. 1) Thus, by changing the time domain, the naturalness and fluency of the audio content can be improved, and the auditory change

from “listening to reading” to “reading for pleasure” can be achieved.

1) The relativity of time domain and frequency domain: The time domain is used to describe the relationship of time, and the audio range over time can be expressed through an intuitive waveform. The frequency domain is used to represent the content of the frequency, and the spectrum shows the relationship between frequency and amplitude. Therefore, the time domain and frequency domain are a pair of corresponding relations (Fig. 2). In general, the description or analysis of sound uses the frequency meter to process its frequency spectrum because a section of the waveform with same or time domain content may have different frequency signals, which requires a further analysis of frequency structure. However, in order to change the naturalness of audiobooks, the basic elements of speech synthesis have been completed through analysis and synthesis of corpora before the implementation process. There is no operation to perform speech synthesis or speech analysis by frequency, and can be achieved only by according to the Chinese broadcast reading habits. The processing of voice information can be achieved through time domain analysis and establishment of corresponding algorithms. In order to better study the sound signal, researchers mostly regard short-term sound transmission as a “quasi-steady” process, that is, the characteristics of the sound signal are basically unchanged in a short time. At the same time, the speech signal realized by the time domain change is more intuitively reflected in the waveform information, and it is easier to handle the post-audio standardization and the judgment of the sound rhythm.

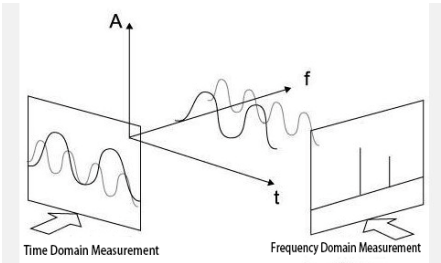


Fig. 2. Relative relation between time domain and frequency domain [5]

2) Intuition of time and level: Time is a visual representation of the response time domain. The time length, the accent, and the pause between syllables in the speech are explained more succinctly by the length of time and the speed. In the process of expressing the speech, the duration of the entire speech can also be judged by the progress of the time, which facilitates later compression or addition of speech content. The level intuitively reflects the level of the volume in the time domain display. Whether or not the level can increase or decrease can affect the processing of the audio content. In the process of increasing the reverberation or pitch change, the value of the level between different positions or the gap measurement of the level data in a period of time determines the audio fluency.

3) Regularity and instability of Chinese pronunciation: According to UNESCO statistics, as the most used language in the world, the number of Chinese users worldwide is 1.6 billion, ranking first in the world's population (accounting for one fifth of the world's population). It is the second most widely used in the world (more than 50 million foreigners use Chinese as a second language). The Chinese phonetic language is mainly based on the “Chinese Phonetic Alphabet Project” promulgated by the State in 1958. It uses the natural spelling of 21 initials and 39 finals of Pinyin and controls the use of “yin, yang, up, and down” and light tones.[6]. The natural spelling and pronunciation of Chinese characters can be found regularly, and the tone of pronunciation has standard norms, so Chinese pronunciation has a certain regularity. However, since Chinese is the same as other languages, it has a polysemy phenomenon and judges the basic semantics through the tone of pronunciation or polyphony. For example, the word “兄弟” pronounced as “xiōng dì” refers to two men. If the pronunciation is “xiōng di”, it only refers to the younger brother. The word “曝” in“一曝十年(yī pù shí nián)” is pronounced as “pù”, but it is pronounced as "bào guāng" in the phrase “曝光”. Therefore, Chinese pronunciation has the characteristics of instability. However, through analysis, we can find that the instability of Chinese speech is based on polysyllabic words. When there are disyllabic or polysyllabic words, we can use the basic semantics to determine the pronunciation of speech, so that we can correctly read the words. This content will be discussed in detail in this chapter’s section 3.

B. Frame Design of Chinese Audio Book Platform by Adopting Time Domain Analysis Technology

Because modern speech synthesis technology is based on electronic processing, it follows the principle of vibration and sound generation, and generates digital waveforms by constructing spectrum and signals to restore sound time domain information. As the early speech synthesis technology is relatively mature, the functions of spatial change and role selection in the vocal range have been basically realized, but the syllable, tone, and rhythm of speech synthesis, especially the processing of naturalness and fluency, have not yet been effectively solved. Currently speech synthesis is a single mode with more structured processing, and it can only provide speech output after text-to-speech conversion. There is no complete system for non-structural processing and personalized customization platforms for audio books. After the initial conversion of the textual information, the time domain analysis can be performed on the synthesized speech content. By adjusting the information of the amplitude, the sound phase, and the level in the time domain, the construction of the naturalness and fluency can be achieved, and the platform of audio books can be established.

The Chinese audio book platform based on time domain analysis technology is established based on waveform analysis and amplitude transformation. The speech information content is measured and analyzed using the breadth of time domain and Chinese specific speech principles, and the waveform information without syllables is extracted. The algorithm is

processed to analyze the prosody and semantics to change the amplitude and the level, to assist the sound phase and the effect, and to realize the construction of a personalized voice for the audio. Its framework is shown in Fig. 3.

Before any audiobook is recommended to audience, it requires the audience to select the appropriate content according to their individual needs. These contents are either paper texts or electronic documents. The paper texts are uploaded to the platform through photographs or scanning. The OCR recognition technology is used to convert textual content into electronic texts. According to the reader level, the reading sounds are divided into four categories: infants, children, youth, and middle-aged people. The sound ranges are divided into male and female voices. The reader determines the speech sound according to his/her age level. The user's needs and selected features are processed. After the user completes the selection, he enters the speech synthesis stage. The synthesized speech enters the platform server in the form of audio data packets, and adjusts the gap between syllables and stops through the syllable analysis algorithm. The rhythm algorithm solves voice tones and accents; and the effect analysis algorithms operates to process audio characters, reverberations, and environments, and an audio standardization algorithm detects amplitudes and levels to ensure that the processed audio is free of noise, pop noise, and silence. After completing the above algorithm processing, audio is transmitted to the audiobook platform for user selection requirement verification in the form of data packets. According to the comparison and confirmation with the sample in the system, the comparison is successfully recommended to the user. The platform operation is completed, and the user begins to experience audio books. If the process is not well performed, they will return to the speech synthesis step and perform algorithm analysis again until the verification succeeds and send the synthesized file to the user.

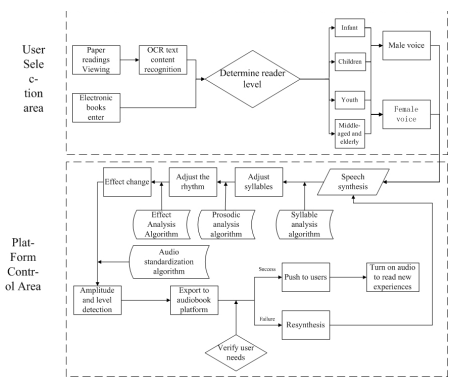


Fig. 3. Framework design of Chinese audiobook platform using time domain analysis technology

C. Syllable Extraction Synthesis And Chinese Rhythm Control

Each pronunciation of a Chinese character is called a syllable. Completion of the electronic speech synthesis by the computer only completes the compilation of the pronunciation of the words in the speech information from the individual to the whole, and does not affect the processing of the speech information. Fig. 4(a) shows the synthesized waveform file. It can be seen intuitively from the figure that the pronunciation of each text exists as an independent syllable, and a relatively low waveform signal appears in the unvoiced space. According to the audio broadcasting standard of professional broadcast institutions, the maximum measurement of digital devices is 0 dB, and the minimum is -60 dB. The volume standard of digital devices is between -24 dB and -12 dB [7]. Such an audio signal that is not higher than −24 dB when it is not sounded may be referred to as a noise signal, and the volume value in the k-time domain is set. When the k-value is higher than -24 dB, the noise calculation approaches the minimum value of $-\infty$, and noise reduction processing is performed on the entire audio file waveform by reducing the overall volume reference. Fig. 4(b) shows the waveform graph processed by the noise analysis algorithm. The volume reference decreases as a whole, and the content of the unvoiced part tends to $-\infty$. The circled part of the figure is the most obvious. Certainly, after the method is used, some waveform sounds have a phenomenon that the volume reference falls below −24 dB. There are generally two ways to solve this problem. First, the voice synthesis volume reference value is set in the range of −12 dB to 0 dB during the initial stage of speech synthesis. The second is to analyze and process through the audio standardization algorithm, and such methods will be described in this chapter's Section 4.

After the noise reduction process is completed, the syllable waveforms in the audio file need to be analyzed and judged, and the pronunciation of the vocal syllables is pronounced as a word pronunciation. Fig. 4 shows the pronunciation of each single syllable word intuitively, that is, each single syllable word is not connected and is presented separately. This has caused the phenomenon of "squeaking" in the process of "listening to the book", which has caused problems such as continuous but inconsecutive speech and obvious pronouncement of the machine, which is the case of poor fluency mentioned above. Fig. 5 and Fig. 4 are the same artificial recording waveforms with the same pronunciation content. Through the analysis of the waveform, it can be found that there is a connection relationship between the former syllable tail and the latter syllable head, and the connection relationship is related to Chinese speech rhythm, sound control of breath and other factors.

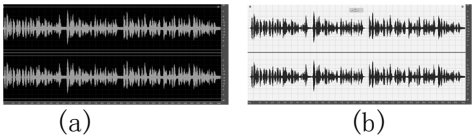


Fig. 4. Comparison of speech synthesis and noise analysis algorithms

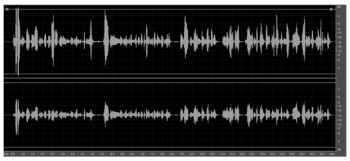


Fig. 5. Manual recording waveform

As a human-specific communicative function, language needs to pass through human's own feelings, judgments and other thoughts before forming speech. In the process of language expression, it is necessary to pay attention to words, rhetoric, and other norms. The pronunciation of Chinese is usually in the form of two-syllable words or polysyllabic words. Monosyllabic words appear less frequently in consecutive text descriptions,

and there are fewer occurrences of words above four syllables. Two-syllable, three-syllable and four-syllable pronunciation frequently happen. Therefore, the Chinese syllables can be grouped into phrases such as "1 (single syllable)," "1+1," "1+2," "2+1," "2+2," and "1+3." There are nine kinds of continuous reading methods such as "3+1", "4" and "4 (syllabic continuous)". After the electronic text is processed into a speech synthesis data packet, the contents of the Chinese characters corresponding to each syllable in the speech waveform can be judged against the contents of the electronic text. In conjunction with the analysis of the semantics of the electronic text, we can know whether or not adjacent syllables are words .Then we use the above nine kinds of continuous reading methods to complete the processing of disyllabic or polysyllabic words.

According to the “Chinese Phonetic Alphabet Project”, there are 21 initial phonetic consonants and 39 finals in Chinese Mandarin. Another is Yinping, Yangping, Shangsheng, Desheng (abbreviation: Yin, Yang, Up, and Down) and 5 minor tones. The biggest difference between the pronunciation of Chinese and English and other Western languages is the complexity of the tone control of the voice. "Yin, Yang, Up, and Down " were marked by the linguist Mr. Zhao Yuanren as the fifth-degree transcription method according to the adjustment value, followed by (Chinese phonetic transcription tone symbol): Yin 5-5 (~), Yang 3-5 (ˇ), 2-1-4 (`), and 5-1 (´). Light tone can be marked as 0-0. In order to facilitate the use and identification in the process of speech processing, it is usually marked with five codes of 1, 2, 3, 4, and 0. When these pronunciations appear on each individual syllable, they are pronounced as the above-mentioned tone; when there are disyllabic or polysyllabic words, they will change according to the rules of speech sound change. The so-called linguistic tone change refers to the change of the pronunciation of some syllables due to the mutual influence of adjacent syllables or the need for expression. The main changes in the linguistic tone change are reflected in the changes in the vocalizations such as upswings, whispers, erhua, etc. For example, "火(huo)" in the "火车(huo che)" is originally 2-1-4, and it needs to be changed to 2-1-1 for continuous reading. The "粉(fen)" in "粉笔(fen bi)" must be changed to 3-5. And these contents can be obtained through the analysis of the data of the existing manual recording samples to obtain the sounding rules of speech flow. These contents can be obtained by analyzing the data of existing manual recording samples. The paper lists the speech changes of the first syllable by analyzing the speech data in 40 manual recording audio files. As shown in Table I. This shows that similar sound analysis methods can be used to summarize the rules of other speech sounds and be applied to speech synthesis.

TABLE I. LINGUISTIC VARIATION ANALYSIS OF SPOKEN SYLLABLES

Post-syllable tone	Syllable pattern		
	1+1	*1+2*	*2+1*
Yin ping	2−1−1+5−5	2−1−1+5−5+5−5	2−1−1+5−5+5−5
Yang Ping	2−1−1+3−5	2−1−1+3−5+3−5	2−1−1+3−5+3−5
Sound	3−5+2−1−4	2−1−1+3−5+2−1−4	3−5+3−5+2−1−4
Devoicing	2−1−1+5−1	2−1−1+5−1+5−1	2−1−1+5−1+5−1
Softly	3-5+0-0 or 2-1-1+0-0	2-1-1+ Original tone +0-0 or 3-5+ Original tone +0-0	2-1-1+0-0+ Original tone or 3-5+0-0+ Original tone

TABLE II. CORRESPONDENCE BETWEEN TONE PITCH VALUE AND TIME DOMAIN THRESHOLD (UNIT: MS)

No.	Adjustment (d)	Total Time Domain Threshold (Z)	Header Time Domain Threshold (t)	Word Belly Time Domain Threshold (f)	Suffix Time Domain Threshold (w)
1	5-5	258-703	40-168	167-296	78-239
2	3-5	503-631	100-149	293-280	110-184
3	2-1-4	734-925	84-194	436-521	214-210
4	2-1-1	576-721	73-179	334-342	169-200
5	5-1	317-435	45-101	144-162	128-172
6	0-0	206-421	29-38	115-306	62-77

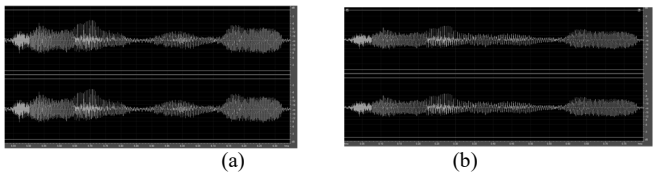
After analyzing and summarizing all the linguistic variations, it can be concluded that the vocal changes are different under different syllable combinations. In particular, the vocal changes from 2-1-4 to 2-1-1 or 3-. After 5, the syllable waveform shows a significant change in the time domain. Thus, according to the shortest syllable "a" and the longest syllable "shuang" of the pronunciation, the time domain threshold of each syllable is summed up under the case of different tone values, and at the same time, the head, the abdomen and the suffix of each syllable are measured. Domain thresholds, as shown in Table II.

The table vividly shows that in the case of the same syllable, the time domain changes are smaller except for the light tone, and the time domain of the word abdomen and suffix changes greatly. This is because the word belly determines the spelling part of Chinese syllables. The suffix contains the syllable tail and the scent of the speaker, and the prefix is only the beginning of syllable pronunciation. This can use the method of coincidence of the header and suffix waveforms to achieve "1+1", "2+2", "1+2", "2+1", and "2+2" speech connections and complete fluency. Define the speech synthesis syllable time domain as “L”, the previous syllable adjustment value as “d”, the previous syllable total time domain threshold as “Z”, the head time domain threshold as “t”, the word belly time domain threshold as “f”, the suffix time domain threshold as“w” ,the value of the last syllable as “d`”, the total time domain threshold of the previous syllable as “Z`”, the time domain threshold as

"t`", the time domain threshold as "f`", and the end time domain threshold as "w`" :

$$L = t + f + w \vee t` + f` + w` \quad (1)$$

According to this time domain threshold, the coincidence experiment was performed on the syllable waveforms in Fig. 4. Select two polysyllabic words "智能(zhi neng)" and "语音(yu yin)" to verify the verbal fluency after the prefix and suffix coincide. Fig. 6 is a comparison of the speech waveforms before and after the two syllables coincide. The tuning value of "能" is yangping 3-5, the threshold of suffix time domain should be 110-184; the tuning value of "语(yu)" is 2-1-1 after the change of upper voice, and the time domain threshold of prefix should be 73- 179.Judging from the time domain, (a) is 935 ms, (b) is 800 ms, when overlapping, the time domain is 135 ms, which is within the range of two thresholds. Judging from the auditory aspect, the audio after reconciliation has better fluency, consistent content, and clear and complete speech.



(a) (b)

Fig. 6. Comparison of waveforms in the "智能" "语音" "2+2" mode

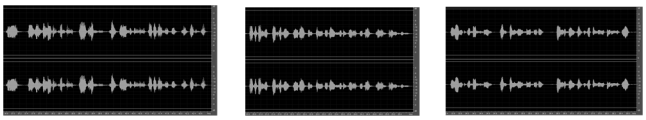
Note: Figure (a) refers to eclipse and Figure (b) refers to overlapping portion.

As mentioned above, the grammatical semantics of Chinese is pronounced as a basic unit of rhythm with two-syllable words. Voice after the analysis of the time domain only has basic fluency, and it is difficult to compare with the real voice. This requires an analysis of the phonological and textual rhythms of synthesized audiobooks that have been completed through the above steps. The "listening" process of audiobooks is not only a process of receiving auditory and awareness information, but also a process of understanding, digesting textual content and thus guiding thinking. Therefore, the audio synthesis of audio books is aimed at texts with textual features rather than phrases. The rhythm of discourse is mainly focused on stress.

Stress refers to syllables that should be re-read according to the content of the article and the semantics of the sentence. In Chinese, stress is generally divided into "stress" and "secondary stress". A novel full of ups and downs, on the audiobook platform, can be better by reading aloud the text. It can not only restore the novel's original appearance, respect the author's creation, but also enable the listeners to feel the description and the fantastic mood depicted by the sound from the text. This requires that some locations, time, and characters need to be reflected in stress. Since there is an accent, the content of the text should also be weakened, and the beginning and end of the text should be treated with strength and weakness. In the following section, we will select an article published on Xinhua.net on March 27, 2018 to illustrate the practical application of textual rhythms by manually annotating stress ( ), sub stress ( ), starting (⺈), and gaining (⺊).

"... In recent years, China's express and take-away industry has developed rapidly. The number of express delivery companies nationwide has grown to more than 20,000, and the average daily express service users exceed 200 million. However, some express delivery companies ignore the traffic safety, and the delivery staff ignore red lights and retrograde. Traffic violations such as speeding are more common, which disrupts the order of traffic in the city and leads to traffic accidents...."

It is not difficult to find through the annotation of textual rhythms that the subjects (ie, people, things, objects), quantifiers, results, etc. are emphasized in the form of stress or sub-stress. According to Chinese phonetic pronunciation and expression habits, the format of stress can be divided into two categories: "middle weight" and "heavy light". After the speech analysis of Kang Hui, Li Zimeng, and Hai Xia's news broadcast works in CCTV "News Broadcasting", we found that the accent of each double-syllable or multi-syllable phrase mostly stays on the belly of the last syllable in the phrase. The end of the phrase will appear to show signs of recovery. This phenomenon is reflected intuitively in the waveform by the upper and lower equal increments of the amplitude, as shown in Fig. 7.



(a) Kang Hui (b) Li Zimeng (c) Haixia

Fig. 7. CCTV News Three announcer news broadcast works

Through the comparison of the syllables before and after the content of the phrase, it can be known that the current syllables are overtone words, such as "星星(xing xing)"、 "爷爷(ye ye)"、 "银闪闪(yin shanshan)". The syllables are presented in a heavy light format, that is, the first syllable waveform is significantly stronger than the latter syllable waveform. The polysyllabic words gradually decrease in relation. In the current postsyllable non-overlapping words, the waveforms of the first syllable and syllable are significantly weakened and increase in relationship. The syllable stress falls on the word and suffix of the last syllable, such as the last syllable. At the end of a paragraph or chapter, only the stress falls on the word abdomen, and the suffix shows the trend. After a simple measurement, the highest amplitude in the time domain under stress is not more than -12 dB. In the sub-accents, that is, in the mid-acoustic state, the highest amplitude in the time domain does not exceed -17 dB and is higher than -22 dB. The minimum amplitude in the time domain in the light tone state is not less than - 24dB. The measurement of this result complies with the requirements of audio volume standards for broadcasting and television broadcast institutions. Combined with this phenomenon and principle, the syllable time domain analysis rhythm has the following conditions:

1) Heavy light format: Define the time domain of the first syllable as Q, the time domain of the second and subsequent syllables as E, and adjust the Q amplitude to an enhanced state, and the enhancement limit does not exceed -12 dB. We found that the E time domain gradually weakens, and the final attenuation value is not less than - 24dB. For example, if the first syllable has reached -12 dB after the speech synthesis, the Q time domain may not be adjusted and the Q time domain may be adjusted after the attenuation.

2) Medium format: Define the time domain from the first syllable to the second last syllable as R, the last syllable as d, the total time domain as Z, the head time domain as t, the word belly time domain as f, and the suffix time domain as w. The pitch value of the syllable is first judged, and the adjustment value may only appear in the “1, 2, 3, 5, and 6” situations in Table II.

Condition A is a phrase not ending in a sentence or paragraph. The overall adjustment of R is -17dB to -22dB, and the overall adjustment of the time domain Z at the end of the syllable is -16dB to -12dB, and the time domain t and w show a fading and fading tendency.

Condition B is the phrase at the end of a sentence or paragraph. The overall adjustment of R is -17dB to -22dB. The head t in the time domain of the last syllable coincides with the syllable of the previous syllable in the previous speech adjustment. Therefore, only the word abdomen needs to be adjusted. The domain f is -12dB, and the end-point time domain w is gradually weakened, and is not lower than -24dB.

Fig. 8 shows the waveform effect of Fig. 6(b) after adjusting the stress to the condition of “re-format condition A” in the voice of “smart speech”.

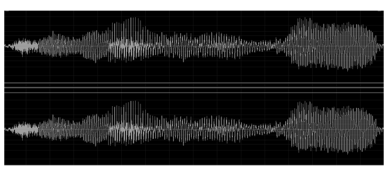


Fig. 8. Medium format condition a waveform effect

D. The "Elasticity" of Audio and Audio Standardization

“Elasticized” sounds generally refer to the state of an announcer or a free reader who is free to use sound during the manuscript processing, which can make the sound free to change the sound range, follow the level of the emotional processing syllable tone, and control the sound modality through resonance. Audio files that have undergone electronic speech synthesis technology do not have the sound range changes and resonance state of the sound, and it is impossible to talk about the influence of emotions on tones. How can the audio of audio books be acoustically eliminated from "dry sound" and enter the state of "wet sound"? Can it be changed after time domain analysis? After analyzing and experimenting with speech synthesis audio files, a positive answer was obtained. The biggest difference between an artificial reading audiobook and a speech synthesized audiobook is that the latter's sound pitch and range are stable in the same state. Although the syllables are stressed, they still cannot change the audio auditory experience alone. The influence of the vocal "resonance" effect is a direct factor. The vocal resonance of analog audio is mainly reflected in "reverberation" in electronic audio, and this "reverberation" effect is different from "room reverberation" or "convolution reverberation" and other electronic added to the audio in the recording studio. The effect is that it only needs to enhance the audio and analog thickness to achieve "reverberation" effect. According to this principle, a file is generated from the file that has completed the above algorithm to enhance the speech effect. After decoding the two preceding and following data packets, the waveform content is overlapped in the time domain.The error time threshold 1.5ms-2.5ms means audio enhancement, which produces "resonance" effect. Error time threshold 2.5ms-4.0ms refers to "room reverberation" effect.

The standardization of audio processing refers to the processing of reference data for multiple-effects or audio synthesis of different paragraphs to ensure that the output audio can achieve a uniform volume, avoid “pop”, “tearing sound”, too loud or too soft and reduce the audience's auditory bias on the same audiobook. The audio reference is known to be -24dB to -12dB, and the overall audio time domain can be set to the threshold dispersion. It should be noted that not all syllables need to be adjusted in the audio normalization process, and it is only necessary to detect and adjust syllables that exceed or fall below the threshold in the audio. In addition, given that the audio is too low or too high, the data may be caused by algorithms such as stress effects and reverberation effects, it is only necessary to ensure that these syllables can be standardized within -24 dB to -12 dB to guarantee the processing effect of the previous algorithm. Define the syllable in the time domain below - 24 dB as t, so that it approaches to -24 dB and define the syllable in the time domain higher than -12 dB as g, making it approach to -12 dB.

The completed speech synthesis audio will enter the platform for detection in the form of data packets, occupying different data spaces according to different durations. Before sending to the user end, in order to ensure that the audiobook is subjected to algorithm analysis and technical processing quality, effect detection will also be performed. Specific steps are as follows:

Step1: Use the user's choice of content to check whether the processing steps of the platform are complete;

Step2: Check whether the overall time domain reference p in the data packet is within the threshold range.

Step3: Extract the time domain of any two-syllable or multi-syllable phrase and detect whether it is a "middle-heavy" or "heavy light" time-domain change;

Step4: According to the numerical relationship between the adjustment value and the syllable threshold, use the embedded algorithm technology data to calculate whether the word F in the time domain and the threshold f in Table II are consistent:

(1) If F = f, and = -24dB $\leqslant$ p $\leqslant$ -12dB, the audiobook audio platform effect processing is successful, and can be sent to the user side;

(2) If F$\neq$f or -24dB $\geqslant$ p or p $\geqslant$ -12dB, then there will be a platform effect processing error in the audiobook. Return to the initial stage of speech synthesis in the platform control area, and reprocess to the user's end after the detection is successful.

IV. Summary and Prospect

Audio books have been widely used as an indispensable part of digital library construction. The intervention in speech synthesis technology has also shifted from theoretical research to applied research. However, the effects of naturalness, fluency, etc. have not been dealt with yet. When "Internet +" and "mobile network" have entered the era of economic development as the

main driving force, the new industrial revolution of audiobooks is poised to take off, and the subsequent cost, efficiency, and other factors affecting the balance have gradually emerged. The development of artificial intelligence has enhanced the unique "destructiveness" of the Internet: It establishes a thriving new mechanism for digital publishing and digital reading industry while dissolving old things through reintegrating resources and optimizing data. The audiobook platform based on time-domain analysis will be able to realize the transition from artificial production to artificial intelligence by virtue of its selective diversification, high anthropomorphic naturalness, rhythm, and fluency of syllables.

A. Summary

This article takes the audiobook platform as the research object, based on the time domain analysis techniques and methods, combines the laws and characteristics of Chinese linguistics, and solves the issues of poor naturalness, weak fluency, and single mode in the reading process of audiobooks. Data analysis and experimental results verify the feasibility of this solution. At the same time, it analyzes the relationship between the current speech synthesis technology and audio books and the problems at the theoretical and technical level.

Aiming at the theory of Chinese linguistics and the technical principle of acoustics, a system scheme for enhancing the audible reading effect of audiobooks using time domain technology was proposed, and a relatively complete audiobook platform was constructed. Using syllable extraction and prosody control methods, on the basis of ensuring the integrity of speech syllables, the threshold value of the word domain is calculated, the algorithm is embedded, and the effect is achieved. Audio benchmark is detected, and the data sent to the user end is achieved, thus achieving user selection and voice transformation, effects enhancement, achieving a smooth, natural sense of hearing.

B. Outlook

With the innovation of technology, the focus of the work of the library has shifted from a single text job to a multimedia job. The implementation of the digital library project has also led the construction of the library to an information-based high-speed channel. The gradual decline in the book publishing industry and the sales industry has also accelerated the pace of the “listening to the book” era. The future of audio books will change not only the way humans read, but also the conception of writing paradigms, emotional expression, and sound reduction, and the emergence of phenomena.

For language reading, “stop” is the most obvious way to control the pace of text. In the writing, the length of the pause is divided by some punctuation such as ",", ".", ",", ";", but in the text, the stop of some sentences seems to be impossible to complete by writing. For example: "他把几个团的负责干部叫到一起" If the severing sentence is between “个(ge)” and “团(tuan)”, it means the cadres of the Communist Youth League; if the severing sentence is between “ 的(de)” and “ 负(fu)”, it means that the regiment-level cadres . These are not marked in the general text and can only be understood based on the reader's understanding of the text context. Using text content big data analysis and comparing it with many sample content may solve this problem. The expression of emotions in biological species is greatly affected by psychological factors and they are different. Understanding of the same sentence or the same word may be very different. How to locate the audience's emotional thinking and how to make reasonable use of speech synthesis technology for realizing the release, grasp, and exposure of textual emotions are also worth discussing. The development of speech recognition and translation synthesis technology can translate the language people describe in another language according to customized requirements. In the process of audio reading, can we synthesize the audience's own speech through a certain technical means so that the reader can listen to their own version? These are topics for the future development of digital library construction and digital publishing.

The future reading era will be led by artificial intelligence, and will continue to develop with people’s improving living conditions.“Listening to books”, “reading pictures”, or “reading films” will gradually change people’s reading habits of paper reading materials, the writing habits of the text, the emotional expression habits, and redefining the word “reading”.

REFERENCES

[1] Sogou Encyclopedia, China Digital Library.(2017-03-20) [2018-03-02]. http://baike.sogou.com/v7676710.htm

[2] Intelligent Research Consulting, 2018-2024 Audio Book Industry Analysis and Development Forecast Report. (2018-02-24) [2018-03-12]. http://www.ibaogao.com. Com

[3] China Audiovisual and Digital Publishing Association. 2016 China Digital Reading White Paper.(2017-04-21) [2018-03-12]. http://www.sohu.com/a/135438957_500643

[4] Lu Shinan, Chu Min, Xu Jieping, He Lin, Chinese Speech Synthesis - Principles and Techniques. Beijing: Science Press, 2012.

[5] Sogou Encyclopedia. Frequency Domain.(2017-03-7) [2018-03-08]. http://baike.sogou.com/v7676710.htm http://baike.sogou. Com/v16431.htm?fromTitle=Frequency domain

[6] Ye Yusheng, Xu Tongyu, Linguistics Outline. Beijing: Peking University Press, 2010.

[7] National Press, Publication, Radio, Film and Television Administration of the People's Republic of China, Radio and Television Center Standard.(2015-11-30) [2018-03-29]. http://www.sarft.gov.cn/art /2015/11/30/art_151_29129.html