TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

School **School of Computer Science & Robotics**
Academic program **09.04.01 Computer Science and Engineering**
Division **Division for Information Technology**

MASTER'S GRADUATION THESIS

Topic of research work
Intelligent voice transcription based on iFLYTEK WEBAP

UDC 004.934.8.1

Student

Group	Full name	Signature	Date
8BM03	Yile Liu		

Scientific supervisor

Position	Full name	Academic degree, academic rank	Signature	Date
Associate Professor	Botygin I.A.	PhD		

SECTION ADVISERS:

Section «**Financial Management, Resource Efficiency and Resource Saving**»

Position	Full name	Academic degree, academic rank	Signature	Date
Associate Professor	Bylkova T.V.	PhD		

Section «**Social Responsibility**»

Position	Full name	Academic degree, academic rank	Signature	Date
Full Professor	Fedorenko O.U.	PhD		

ADMITTED TO DEFENSE:

Director of program	Full name	Academic degree, academic rank	Signature	Date
Artificial intelligence and machine learning	Spitsyn V.G.	PhD		

Tomsk – 2022

LEARNING OUTCOMES

Expected learning outcomes

Code competencies	Learning outcome (a graduate should be ready)
Universal competencies	
UK(U)-1	Able to critically analyze problematic situations using a systematic approach, to develop a strategy of action
UK(U)-2	Able to manage a project through all stages of its life cycle
UK(U)-3	Able to organize and manage a team, develop a team strategy to reach the set target
UK(U)-4	Able to use modern communication technologies, also in foreign language(s), for academic and professional interactions
UK(U)-5	Able to analyze and take into account the diversity of cultures in the process of intercultural interaction
UK(U)-6	Able to identify and implement priorities of their own activities and ways to improve them on the basis of self-assessment
General professional competencies	
GPC(U)-1	Able to independently acquire, develop and apply mathematical, natural-science, socio-economic and professional knowledge to solve non-standard tasks, including in a new or unfamiliar environment and in an interdisciplinary context
GPC(U)-2	Able to develop original algorithms and software tools, including those using modern intellectual technologies, to solve professional tasks
GPC(U)-3	Capable of analyzing professional information, summarizing, structuring, presenting in analytical reviews with substantiated conclusions and recommendations
GPC(U)-4	Capable of applying new scientific principles and research methods

	in practice
GPC(U)-5	Capable of developing and upgrading software and hardware for information and automated systems
GPC(U)-6	Capable of developing components of hardware-software complexes for information processing and computer-aided design
GPC(U)-7	Capable of adapting foreign data processing and CAD systems to the needs of domestic enterprises
GPC(U)-8	Capable of managing effectively the development of software tools and designs
Professional competencies	
PC(U)-1	Capable of creating software for analysis, recognition and processing of information, digital signal processing systems (06.042 "Big Data Specialist", 06.001 "Programmer")
PC(U)-2	Capable of designing complex user interfaces (06.025 "Graphic and User Interface Designer")
PC(U)-3	Capable of managing processes and projects for creation (modification) of information resources (06.017 "Software Development Manager")
PC(U)-4	Capable of managing the development of complex projects at all stages and phases of work (40.008 "Specialist in the organization and management of scientific is able to manage complex projects at all stages of work performance (40.008 "Specialist in the organization and management of research and development activities")
PC(U)-5	Capable of designing and organizing the educational process of the educational programmes with the use of modern educational technologies (01.002 "Educational psychologist (psychologist in education)")

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

School **School of Computer Science & Robotics**
Academic program **09.04.01 Computer Science and Engineering**
Division **Division for Information Technology**

APPROVED BY:
Director of program
_____ Spitsyn V.G.
«_____»_____2022 г.

ASSIGNMENT
for the Graduation Thesis completion

In the form:

Master thesis

For a student:

Group	Full name
8BM03	Liu Yile

Topic of research work:

Intelligent voice transcription based on iFLYTEK WEBAP	
Approved by the rector's order (date, ID)	No. 34-63/c of 03.02.2022

Deadline for completion of Master's Graduation Thesis:	10.06.2022

TERMS OF REFERENCE:

Initial data for research work *(name of the object of study or design; productivity or load; mode of operation (continuous, periodic, cyclical, etc.); type of raw material or product material; requirements for the product, product or process; special requirements for the specific functioning (operation) of the object or product in terms of operational safety, environmental impact, energy costs; economic analysis, etc.).*	The research and development object is the deep convolutional neural network algorithm based on iFLYTEK intelligent speech recognition.

List of the issues to be investigated, designed and developed *(analytical review of the literature in order to find out the achievements of world engineering science in the field in question; statement of the research, design, construction task; content of the research, design, construction procedure; discussion of the results of the work performed; names of additional sections to be developed; conclusion on the work).*	1.Research review 2.Introduction to End-to-End Model-Based Speech Synthesis Methods 3. Corpus Construction and Preprocessing 4. Dongxiang Speech Synthesis Based on End-to-End Model 5. Financial management, resource efficiency, and resource saving. 6. Social responsibility.

Advisors to the sections of the Master's Graduation Thesis

Section	Advisor
Financial Management, Resource Efficiency, and Resource Saving	Associate Professor Bylkova T.V.
Social Responsibility	Full Professor Fedorenko O.Yu.
English Language	Senior Lecturer Anufrieva T.N.

Date of issuance of the assignment for Master's Graduation Thesis completion according to the linear schedule	10.03.2022

The assignment was given by the scientific supervisor:

Position	Full name	Academic degree, academic rank	Signature	Date
Associate Professor	Botygin I.A.	PhD		

The assignment was accepted for execution by the student:

Group	Full name	Signature	Date
8BM03	Liu Yile		

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

School **School of Computer Science & Robotics**
Academic program **09.04.01 Computer Science and Engineering**
Division **Division for Information Technology**

Form of presenting the work:

Master's Thesis

SCHEDULED ASSESSMENT CALENDAR
for the Master's Graduation Thesis completion

Student:

Group	Full name
8BM01	Liu Yile

Topic of research work:

Intelligent voice transcription based on iFLYTEK WEBAP

Deadline for completion of Master's Graduation Thesis:	10.06.2022

Assessment date	Title of the section (module) / type of work (research)	Maximum score of a section (module)
1.03.2022	Drawing up and approving the terms of reference	10
10.03.2022	Selection and study of materials on the topic	10
30.03.2022	Research of subject area	15
20.04.2022	Conducting experiments	25
15.05.2022	Analysis and description of results	25
01.06.2022	Preparing for thesis defense	15

COMPILED BY:
Scientific supervisor

Position	Full name	Academic degree, academic rank	Signature	Date
Associate Professor	Botygin I.A.	PhD		

AGREED BY:
Director of program

Position	Full name	Academic degree, academic rank	Signature	Date
Full Professor	Spitsyn V.G.	PhD		

Group	Full name	Signature	Date
8BM03	Liu Yile		

TASK FOR
«FINANCIAL MANAGEMENT, RESOURSE EFFICIENCY AND RESOURSE SAVING» SECTION

To student:

Group	Name
8BM03	Liu Yile

Institute	School of Computer Science & Robotics	**Department**	Division for Information Technology
Education level	Master Degree Program	**Academic program**	Computer Science and Engineering

Initial data to «Financial management, resource efficiency and resource saving » chapter:	
1. Costs of research, including technical, financial, energy, information and human costs	Cost of material resources determined by the average cost of Tomsk
2. Norms of expenditure of resources	District coefficient - 1.3...
3. The taxation system used, the rates of taxes, discounting and lending	The coefficient of deductions for payment in off-budget funds - 27.1%
List of tasks:	
1. Evaluation of commercial and innovative potential	1. Analysis of potential consumers. And competitive solutions. 3. SWOT analysis. 4. Assessment of project readiness for commercialization
2. Development of the charter of the technical project	1. Stakeholders of the project. 2. Goals and results of the project. 3. Organizational structure of the project. 4. Constraints and assumptions of the project.
3. Planning of management process: structure and schedule, budget, and risks	1. Hierarchical structure of project work. 2. Project plan. 3. Research budget.
4. Estimation of resource, financial and economic efficiency	1. Integral indicator of efficiency.
List of graphical data:	
1. "Portrait" of the consumer of NTI results *2. Market segmentation* *3. Assessment of the competitiveness of technical solutions* *4. FAST chart* *5. SWOT matrix* *6. Schedule and budget for NTI* *7. Assessment of resource, financial and economic efficiency of NTI* *8. Potential risks*	

Date of task obtaining	

The task was given by the adviser:

Position	Name	Academic degree	Signature	Date
Associate Professor	Bylkova Tatyana Vasilievna	PhD		28.02.2022

The task was accepted by the student:

Group	Name	Signature	Date
8BM03	Liu Yile		28.02.2022

TASK FOR
«SOCIAL RESPONSIBILITY» SECTION

To student:

Group		Name	
8BM03		Liu Yile	
Institute	School of Computer Science & Robotics	**Department**	Division for Information Technology
Educational level	Master Degree Program	**Academic program**	Computer science and engineering

Subject BKP:

Intelligent voice transcription based on iFLYTEK WEBAP	
Initial data to «Social responsibility» chapter:	
Introduction - Characteristics of the object of study (substance, material, device, algorithm, technique) and the scope of its application. - Description of the working area (workplace) when developing a design solution / during operation	*Object of study iFLYTEK WEBAPI* *Scope of Speech Recognition* *Working area: office, laboratory* *Room dimensions: Two radiators. Two ventilation windows, 8 fluorescent lamps meet the needs of the human body in the working environment.* *Quantity and name of work area equipment high powered computer* *Workflows related to the object of study, carried out in the working area:_Modernization of equipment and writing new control codes and their debuggin*
List of items to be investigated and to be developed:	
1. Legal and organizational issues to provide safety when developing a design solution: - Special (specific for operation of objects of investigation, designed workplace) legal rules of labor legislation; - Organizational activities for layout of workplace.	TOI R-45-084-01 "Typical instructions for labor protection when working on a personal computer" GOST 12.2.032-78 SSBT "Workplace when performing work while sitting. General ergonomic requirements» СанПиН 1.2.3685-21 Hygienic standards and requirements for ensuring the safety and (or) harmlessness of environmental factors for humans SANITARY NORMS SN 2.2.4 / 2.1.8.562-96 "Noise at workplaces, in the premises of residential, public buildings and in residential areas" GOST 12.1.005-88 System of labor safety standards (SSBT). General sanitary and hygienic requirements for the air of the working area GOST 12.1.004-91 "SSBT Fire Safety" SP52.13330.2016 Natural and artificial lighting. Updated edition of SNiP 23-05-95 GOST 12.1.006-84 SSBT. "Electromagnetic fields of radio frequencies. Permissible levels at workplaces and requirements for control" Labor Code of the Russian Federation of December 30, 2001 No. 197-FZ N197-FZ (as amended on March 9, 2021):

	SanPiN 2.2.1/2.1.1.1200-03 "Sanitary protection zones and sanitary classification of enterprises, structures and other objects" GOST 17.1.3.07-82 Nature Protection (SSOP). Hydrosphere. Water quality control rules for reservoirs and streams GOST 17.2.3.01-86 Nature Protection (SSOP). Atmosphere. Rules for air quality control in settlements GOST 17.4.3.04-85 Nature Protection (SSOP). Soils. General requirements for control and protection against pollution
2. Work Safety when developing a design solution: -Analysis of identified harmful and dangerous factors. -Justification of measures to reduce probability of harmful and dangerous factors	**Harmful factors:** 1. Insufficient illumination; 2. Violations of the microclimate, 3. Exceeding the noise level. 4. Increased level of electromagnetic radiation **Dangerous factors:** - Increased voltage value in the electrical circuit, the closure of which can occur through the human body. Required means of collective and individual protection against the identified factors: headphones, devices for ventilation and air purification, light sources, devices for trapping and purifying air and liquids, protective grounding devices, protective coatings from electromagnetic radiation. Calculation: calculation of the artificial lighting system.
3. Ecological safety when developing a design solution	The impact of the object on the residential area, atmosphere, hydrosphere does not occur. Negative impact on the lithosphere during the disposal of computers and peripheral devices (printers, MFPs, webcams, headphones, speakers, telephones), fluorescent lamps, waste paper.
4. Safety in emergency situations when developing a design solution:	Possible emergencies: natural disasters, for example, a hurricane; geological impacts (earthquakes, landslides, landslides, collapses of the territory, etc.); technogenic accidents (sabotage, fires).
Assignment data for section according to schedule	

The task was issued by consultant:

Position	Full name	Scientific degree, rank	Signature	date
Full Professor	Fedorenko O.Yu.	PhD		

The task was accepted by student:

Group	Full name	Signature	date
8BM03	Liu Yile		

Summary

The Master's Graduation Thesis contains 104 pages, 25 images, 38 tables, 53 references.

Keywords: Intelligent speech recognition; iFLYTEK; Dong-Xiang speech synthesis; Non-text speech generation; Autoregressive end-to-end models； National common language； Non-autoregressive end-to-end model

The research object is the transcription and generation of Chinese dialects.

The aim of this research is to design an algorithm for intelligent speech recognition transcription and generation through an end-to-end speech synthesis model.

Before training the model, the speech in the body is improved and pre-processed. Improving the quality of the speech in the housing improves the quality of the synthesized speech. Finally, accurate acoustic features are extracted manually and the model is improved by introducing non-autoregressive end-to-end models (FastSpeech, FastSpeech2).

In model training, the autoregressive end-to-end model (Tacotron2, Transformer) is first used to train the model. However, due to the lack of training corpus and the defects of the autoregressive model, the problems of missing words, skipping words and slow synthesis speed occur, thus affecting the generalization ability of the model. To this end, this paper manually extracts accurate acoustic features and proposes a non-autoregressive end-to-end model (FastSpeech, FastSpeech2) as an improved method. The experimental results show that the FastSpeech2 model has the best performance in the Dongxiang language speech synthesis task.

List of abbreviations

TTS	Text-To-Speech
HMM	Hidden Markov Model
LSTM	Long short-term memory
Pre-Net	post-processing network
MPD	Multi-Period Discriminator
MSD	Multi-Scale Discriminator
FFN	Feed Forward network
FFT	Feed-forward Transformer
LR	Length Regulator
CWT	Continuous Wavelet Transform
iCWT	Inverse Continuous Wavelet Transform
STFT	Short-time Fourier Transform Frame
PESQ	Perceptual evaluation of speech quality
DMOS	Degradation Mean Opinion Score
MOS	Mean Opinion Score
SEGAN	Speech Enhancement Generative Adversarial Network
DP	Dynamic Programming
G	Generative Mode
D	Discriminative Model
RMSE	Root Mean Square Error

Table of contents

INTRODUCTION

China is a country with a vast territory, numerous ethnic minorities, and rich dialects. In addition, local ethnic minorities have their local dialects. At present, many ethnic minorities dialects have been listed as national endangered languages. Therefore, China advocates vigorously protecting the intangible cultural heritage of ethnic minority dialects. As the national commonly-used language, Mandarin needs to be widely disseminated among ethnic minorities. However, ethnic minorities areas lack bilingual teachers to teach Mandarin. Therefore, studying the speech synthesis technology of ethnic minority dialects can protect the ethnic dialects and has crucial significance for the national commonly-used language education in ethnic minorities areas. Many ethnic minority dialects in our country have no written expression. For example, the Dong-Xiang language, which is studied in this thesis, is not only a language without written expression but also endangered in the country. However, the current speech synthesis is all Text-To-Speech synthesis (TTS). The text is firstly processed by linguistics such as text norm, text segmentation, and grammatical analysis. Then the speech is synthesized by a vocoder from a trained acoustic model. This makes it challenging to synthesize Dong-Xiang speech without text. Therefore, this thesis takes the Dong-Xiang language as the research goal. We adopt Chinese characters as the expression of the Dong-Xiang language and use the Pinyin to mark Dong-Xiang phonetic symbols. The low-resource Dong-Xiang speech synthesis is implemented using an end-to-end speech synthesis method.

1 Research review

1.1 Research background and significance

Voice is the most commonly used communication method for human beings. People can quickly collect and exchange information based on voice. With the continuous development of information technology, human beings pay more and more

attention to the convenience of human-computer interaction. Obviously, voice is the best choice for establishing human-computer interaction 1. Speech synthesis is a hot topic of research in recent years, and the most commonly used method is text-to-speech conversion. Speech synthesis has a wide range of applications in daily life, such as translators, smart cars, smart homes, and smart reading. Not only that, in the education industry, it can also be widely used in bilingual teaching 2, which greatly improves the teaching quality and convenience.

China has a vast territory and many ethnic groups, resulting in a rich variety of dialects. The diverse dialects are intangible cultural heritage that our country vigorously protects. With the establishment of the "Law of the People's Republic of China on the Standard Language of the People's Republic of China" [3], Putonghua not only promotes the standardization and standardization of the national standard language, but also promotes economic and cultural exchanges among various ethnic groups and regions. Therefore, the study of ethnic minority dialect synthesis can not only protect ethnic dialects, but also alleviate the problem of lack of bilingual teachers and promote the widespread spread of Mandarin among different ethnic groups, which is of great significance to minority Mandarin education.

The unique national language of the Dongxiang people is Dongxiang language, which is rich in and inherited the history and culture of the Dongxiang people. Dongxiang language is an endangered language in China, and it has attracted much attention and research by domestic and foreign scholars due to the integration of various language components. It is different from most languages in that, firstly, it has no written expression, which makes it impossible to synthesize it by conventional text-to-speech conversion, and secondly, the corpus is extremely scarce and very difficult to obtain.

However, with the advancement of modernization and urbanization, the language and cultural resources of the Dongxiang language have been seriously lost, and it has been listed as an endangered language in China 4. On May 14, 2015, the Ministry of Education and the State Language Commission announced the official launch of the National Language Resources Protection Project 5. At the same time, the

promotion of Putonghua among the Dongxiang people is also being implemented, but the Dongxiang people lack bilingual teachers who can teach Putonghua. Therefore, the study of Dongxiang speech synthesis can not only protect the Dongxiang language well, but also promote the teaching of Putonghua among the Dongxiang people. Since Dongxiang is a language without written expression and has very few corpus resources, the traditional text-to-language conversion and training methods cannot be directly applied to the synthesis of Dongxiang, but need to be combined with a large number of accurate manual annotations and other means. Improve the generalization ability of the trained model. Therefore, this paper will take Dongxiang language as the research object, and use the end-to-end speech synthesis method to realize text-free speech synthesis of Dongxiang language under the premise of low resources.

1.2 Research status

1.2.1 Status Quo of Speech Synthesis

Speech synthesis is a technology that converts text into speech. It is similar to the human mouth and uses different tones to express different content. Speech synthesis technology 6 is mainly composed of language analysis part and acoustic system part, and these two parts are also called front-end and back-end. The language analysis part is to analyze the context according to the input text, and generate the corresponding language specification, so that the machine can plan how to read it. The main process is shown in Figure 1.1:

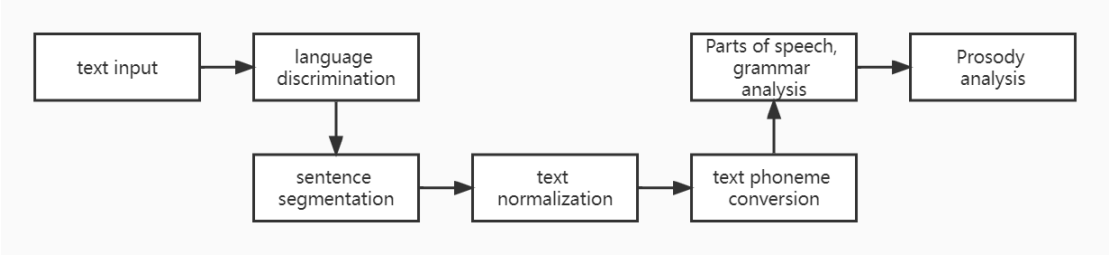


Figure 1.1 Linguistic Analysis

First, the language of the input text is judged, such as Chinese, English, etc., and then the entire text is divided into separate sentences according to the grammatical analysis of the corresponding language, and the text is standardized 7. Text

standardization is the process of converting different text variants into standard forms. For example, the Arabic numeral "123" in the text in Chinese speech synthesis is converted into the Chinese character "one two three" according to the set format rules in the corresponding scene. The next step is to convert the text into a sequence of phonemes. For example, in the training of Chinese speech synthesis, the text data is basically marked with Chinese pinyin, so it is necessary to convert the text into its corresponding pinyin. For example, "the weather is sunny" requires Converted to "tian1 qi4 qing2 lang3", since there are many polysyllabic words in Chinese, how to distinguish pronunciation requires word segmentation and part-of-speech analysis. The last is prosody analysis. Since human beings are rich in tone and emotion when expressing language, the synthesized speech needs to imitate the real human voice, so it is necessary to perform prosodic analysis on the text, such as where to pause and which words need to be lightened. Reading, etc., to achieve the cadence, twists, and turns of the voice.

At present, there are three main technical implementation methods in the acoustic system part, namely, waveform splicing technology, statistical parameter method and deep learning method. Waveform splicing and statistical parameters are also called traditional synthesis methods. Waveform splicing technology is to synthesize the target sentence through the splicing of syllables in the corpus 8. In the early stage, it is necessary to prepare a large number of recorded audios to build a corpus. The corpus in the corpus should contain as much as possible all the syllables and phonemes of the synthesized language 9 and other language features, and then splicing and synthesizing the corresponding text and speech in a large corpus based on statistical rules. The larger the size, the better the effect, which leads to the fact that the technology needs to consume a lot of manpower and material resources, and the synthesized speech will be affected by the environment. The words are too stiff and the naturalness is low, but the speech quality is good 10.

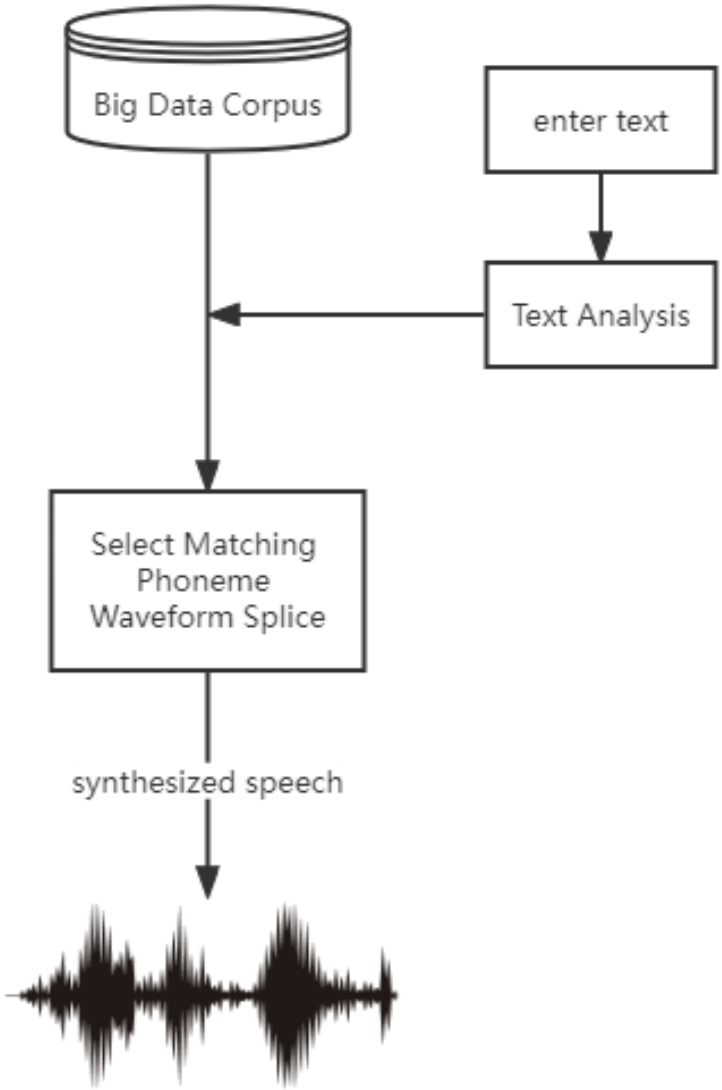


Figure 1.2 Waveform Stitching Technology

Subsequently, the speech synthesis method based on statistical parameters was proposed 11, among which, the statistical parameter method based on Hidden Markov Model (HMM) was widely used in the early stage 12. It mainly uses mathematical methods to model the parameters of the speech spectrum of the corpus, to construct the mapping relationship between text sequences and acoustic features, and generate parametric synthesizers.

When the target text is introduced, the text sequence is first matched with the corresponding acoustic characteristic, and then the acoustic features are converted into speech through the acoustic model 13. It does not require many manpower and material resources, and the synergy between words is too natural and smooth, but the

synthesized voice quality will be mechanically heavy and noisy.

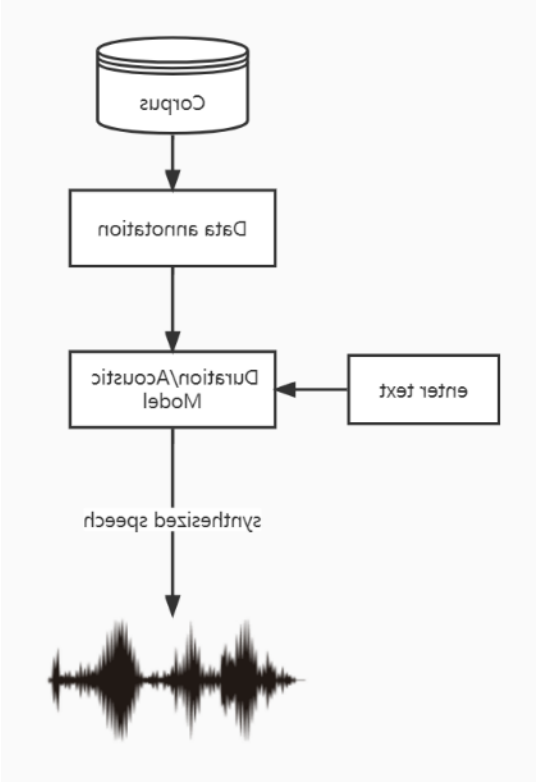


Figure 1.3 Speech synthesis based on statistical parameters

With the improvement of computer performance, the deep learning method has become the current mainstream method, and supervised or unsupervised learning is its characteristic and advantage. The hierarchical extraction of features makes the model have stronger generalization ability. In 2017, Google first proposed a true end-to-end speech synthesis model Tacotron in the paper 14. Compared with traditional speech synthesis methods, it has no complex engineering, acoustic modules, and only uses text sequences and speech spectrum. Training on paired datasets simplifies many processes. Since then, with the iterative update of technology, more and more end-to-end models have emerged, such as Tacotron2 15, Glow-TTS 16, etc., the synthesized speech quality is higher, and the emotion is better.

1.3 Research purpose

Due to the scarcity of corpus resources and the limitations of traditional methods, the synthesized Dongxiang speech quality is not high, the speech synthesis speed is slow, the generalization ability of the speech synthesis model is poor, and the synthesized speech cannot be controlled in a fine-grained manner. It can't be used well

in practical scenarios. In response to the above problems, this paper proposes an end-to-end Dongxiang speech synthesis method, which uses a large number of hand-extracted accurate acoustic features to solve the problems caused by low resources, to achieve high quality and high-efficiency Dongxiang speech synthesis with low resources.

1.4 Summary of this chapter

This chapter introduces the research background and significance of the paper, briefly describes the development and current situation of speech synthesis, and describes the research status of dialect synthesis and Dongxiang synthesis.

2 Introduction to End-to-End Model-Based Speech Synthesis Methods

At this stage, the end-to-end method is the mainstream method of speech synthesis, so this chapter introduces the framework and principle of the end-to-end speech synthesis model required for the experiment in detail. The content of this chapter is mainly divided into three parts: First, the structure and principle of the two autoregressive models Tacotron2 and Transformer are introduced. Secondly, the structure and principle of the two non-autoregressive models FastSpeech and FastSpeech2 are introduced. Finally, the content of this chapter is summarized.

2.1 Autoregressive speech synthesis model

The meaning of an autoregressive model is to use itself as a model of a regression variable, that is, a linear regression model that expresses a random variable at a later time in the form of a linear combination of random variables at multiple times in the early stage. The popular explanation is to arrange the predicted objects in the order of time to form a time series, and then infer the future change law and change trend from the change law of the formed time series. Autoregressive speech synthesis then consists in generating the speech spectrum according to the autoregressive definition, that is, through a step-by-step iteration in the network. The input of each step is the output of the previous step. Thus, the speech spectrum is generated frame by frame. Finally, the spectrum is converted into speech form by a vocoder.

2.1.1 Tacotron2-based speech synthesis method

Tacotron2 is mainly composed of an encoder and a decoder including an attention mechanism. The encoder converts the Dongxiang phoneme sequence into a latent state representation through the neural network, the decoder predicts the mel spectrogram according to the latent state representation output by the encoder, and finally converts the mel spectrum into a speech waveform through the vocoder to complete the whole process. The process of speech synthesis. The network structure is shown in Figure 2.1.

In the encoding stage, it consists of a Character Embedding layer, a convolutional layer and a bidirectional 20 Long Short-Term Memory (LSTM) network 21. First, the Dongxiang phoneme input sequence is encoded into a 512-dimensional phoneme vector in the character-embedding layer. It then passes through three convolution layers. Each convolution layer consists of 512 5 × 1 cores. The input character sequence is contextualized, which is similar to N-grams in natural language processing. This is followed by normalization and use of the ReLU activation function. The output of the final convolutional layer is fed into a bidirectional LSTM to generate encoded features, and the LSTM contains 512 units, i.e. 256 units in each direction.

The output of the final convolutional layer is fed into a bidirectional LSTM to generate encoded features, and the LSTM contains 512 units, i.e. 256 units in each direction.

$$f_e = ReLU(F_3 \times ReLU(F_2 \times ReLU(F_1 \times E(X))))$$

$$H = EncoderRecurrency(f_e)$$

In the formula, F1, F2, and F3 represent the three convolutional layers respectively, ReLU is the nonlinear activation on each layer of convolution, E represents the character embedding of the input Dongxiang phoneme sequence, and EncoderRecurrency represents the bidirectional LSTM.

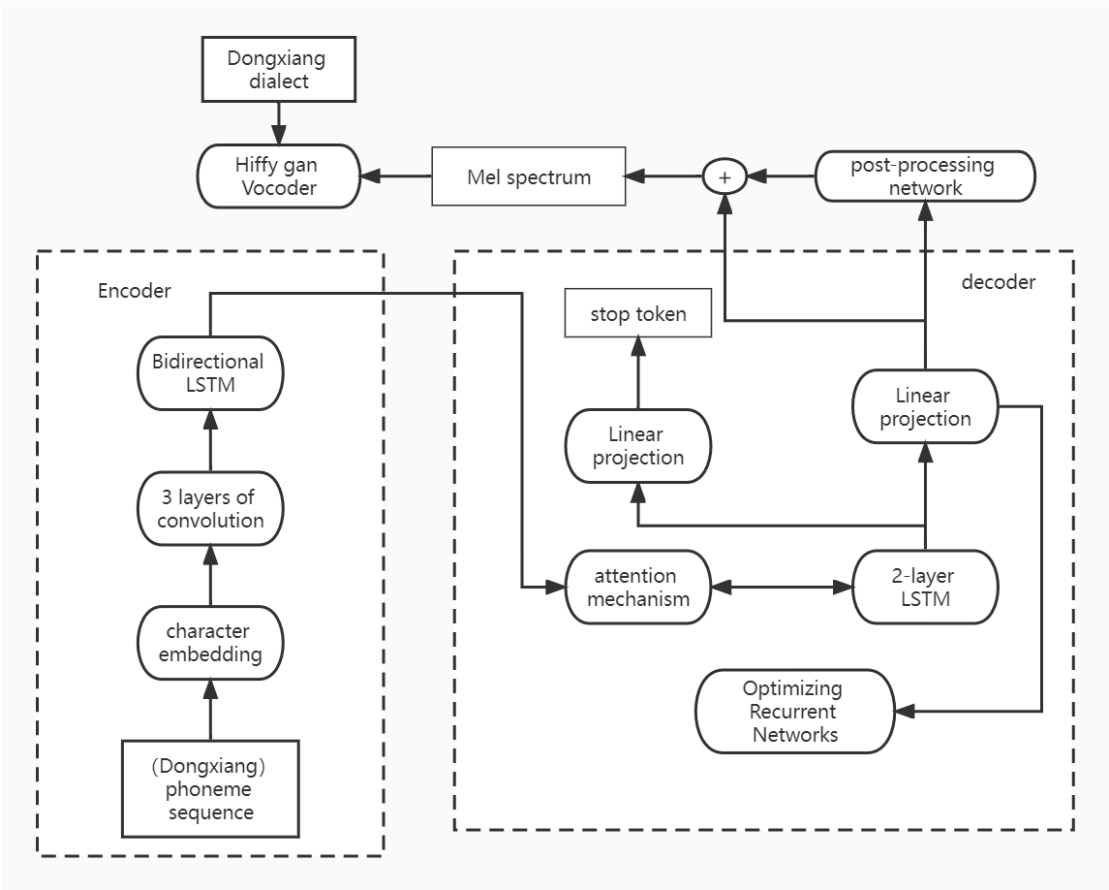


Figure 2.1 Tacotron2+Hifi gan network structure

The structure of the decoder is an autoregressive neural network, which predicts the mel-spectrogram from the input sequence of the encoder, frame by frame according to the order. The network uses a location-sensitive attention mechanism (Attention-Based Models for Speech Recognition) 22. Unlike previous attention mechanisms, it treats the attention weights calculated at previous time points as additional features, which allows the model to remain consistent when moving along the sequence, the purpose of which is to reduce the occurrence of duplication or omission errors in subsequences. The location feature is calculated by 32 1-dimensional convolutions of length 31, and then the input sequence and the corresponding location feature are projected to the 128-dimensional hidden layer representation to calculate the attention weight.

$$f_i = F \times ca_{i-1}$$

$$e_{i.j} = v_a^T Tanh(W_{S_j} + V_{h_j} + U_{f_{i,j}} + b)$$

Where v_a, W, V, U and b are all training parameters, s_j is the current hidden state of

the decoder, W_{S_j} is the current hidden state of the encoder, and V_{h_j} is the previously accumulated attention weight calculated by convolution Positional encoding, which has the advantage of considering both the content and the position of the input element.

In the decoding step, the spectrum predicted in the previous step is transferred to a progressive recurrent network (Pre-Net). This network is a two-layer layer. Each layer of the layer consists of 256 ReLUs. The output of the Pre-Net network is then concatenated with the context vector obtained from the attention part and transmitted to a two layer stacked single layer LSTM with 1024 units. The output of the LSTM is concatenated with the context vector again, and a linear projection is performed to predict the target spectral frame. At the same time, a scalar is projected which is passed to the sigmoid activation function to predict the probability of whether an output sequence has been completed. The predicted target spectrum frame goes through a 5-layer post-processing network (Post-Net), each layer is composed of 512 5×1 convolution kernels, which are used to predict a residual and superimpose the residual to the spectrum frame. The purpose is to improve the process of spectral reconstruction, and finally through the normalization process, except for the last layer of convolution, each layer is activated by the Tanh activation function.

Vocoder experiments use $Hifi\ gan$ 23. Compared with autoregressive vocoders such as Wave net 24, it has higher efficiency to synthesize higher quality audio. The main reason is that its discriminator consists of multi-scale The discriminator (Multi-Period Discriminator, MPD) and the multi-period discriminator (Multi-Scale Discriminator, MSD) are composed of two sub-discriminators. Their main function is to identify the dependence of phoneme duration and corresponding speech waveform. MPD mainly deals with speech data of different periods, while MSD is to evaluate speech samples at different scales. In the experiment, the trained Hifi gan pre-training model is used to synthesize the speech waveform from the Mel spectrogram predicted by Tacotron2.

2.1.2 Transformer-based speech synthesis method

Transformer is still an encoder-decoder structure 25. First, the Dongxiang

phoneme and its corresponding Mel spectrum are used as input, character embedding is performed for each phoneme, and then a positional encoding (Positional Encoding) is added to each character vector. Because each word and the position of the word in the sentence convey different meanings, such as "I stole his pen" and "He stole my pen" in the sentences "I" and "He "Different locations have different meanings. Since Transformer's self-attention mechanism cannot obtain the position information of words, it is necessary to add position encoding to ensure that words are in the correct position. Then, the corresponding relationship between phoneme and mel spectrum is trained by the codec, and finally the mel spectrum is predicted to synthesize speech through the vocoder. The specific network structure is shown in Figure 2.2.

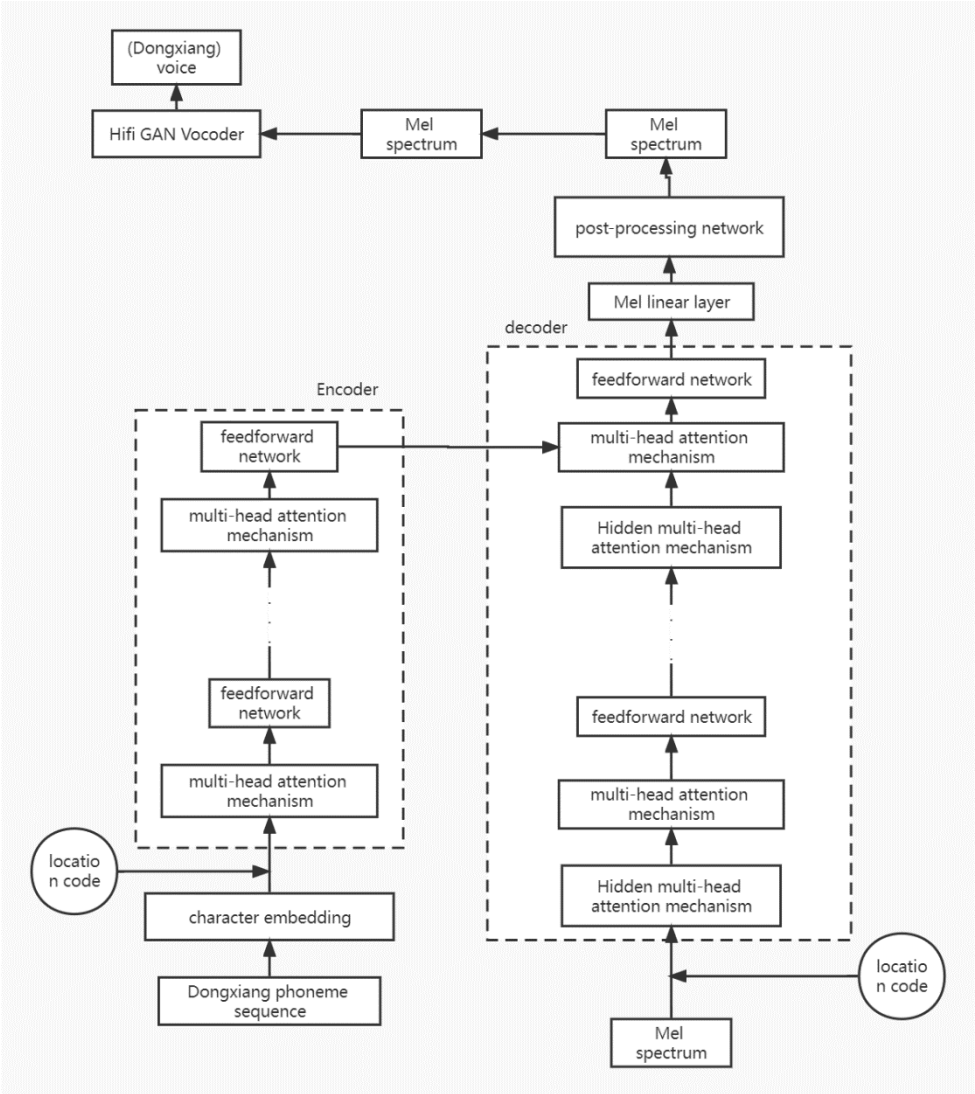


Figure 2.2 Tacotron2+$Hifi\ gan$ network structure

The encoder part is mainly composed of six identical sub-encoders. Each sub-

encoder is composed of a multi-head attention mechanism and a Feed Forward network (FFN). Before the last sub-encoder, the outputs are all inputs to the next sub-encoder, and the last sub-encoder is the input to the decoder. The output representation formula of a single sub-encoder is:

$$Output = LayerNorm(X + Layer(x))$$

Layer(x) represents the output of the current sub-encoder, and X represents the input of the encoder.

The calculation method of Attention is to use Scaled dot-product to calculate the dot product of Q and K, and to obtain the weight of V by dividing each dot product by dk through the softmax function. The calculation formula is:

$$Attention(Q、K、V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

In the formula, Q, K, and V are the input of the attention mechanism layer, and $d_k$is the dimension of feature Q and feature K. They can obtain the similarity between Q and V through the combined inner product of the softmax function, and finally calculate the similarity by weighting. In addition, to obtain a vector of dimension d_v 26.

Multi-head attention maps Q, K, and V by using multiple different linear transformations, then brings in, splices them in parallel, and finally obtains the final attention value through linear transformation. The calculation formula is as follows:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

$$MultiHead(Q、K、V) = Concat(head_1, \dots, head_n)W^O$$

where $W_i^Q \in R^{d_{model} \times d_k}$,$W_i^K \in R^{d_{model} \times d_k}$, $W_i^V \in R^{d_{model} \times d_v}$,$W^O \in R^{d_{model} \times hd_v}$.

In each sub-encoder, in addition to having multi-head attention, there is also a feed-forward network, which has two linear transformations and is activated using the activation function ReLU after the first linear transformation. The linear transformation formula is as follows:

$$FEN(x) = \max(0, xW_1 + b_1)\, W_2 + b_2$$

The encoder part is also composed of 6 identical sub-decoder stacks. Compared

with the sub-encoder structure, it has an additional Masked Multi-Head Attention (Masked Multi-Head Attention), and other structures are the same as the sub-encoder. The calculation principle of concealed multi-head attention is the same as that of multi-head attention, the difference is that a masking mechanism is added, which means that certain values are masked, so that the parameter update will not have an effect. There are two masking methods in the network, namely Padding Mask and Sequence Mask. The first is padding concealment. At the beginning of training, the sentence length of a batch cannot be the same. Therefore, in order to ensure that all input sequences have the same length, it is necessary to extract the longest sequence of this batch as the standard length, and the rest are smaller than it. The short sequence is filled with 0 to the standard length at the back. In order to prevent the attention mechanism from focusing on these 0s, these positions will be given very large negative numbers, so that the probability of these positions will be infinitely close to 0 when passing through $softmax$. The sequence concealment is to prevent the decoder from seeing future information. For example, at a certain training time t, the output of the decoder should only depend on the output before time t instead of after t, so sequence concealment will hide the output after time t.

It can be seen that the position coding of a symbol or spectrum frame position is critical for both the encoder and the decoder [27]. The way Transformer obtains the position code is to use the sine and cosine position code, each code corresponds to the sine and cosine curve, and the wavelength is composed of a geometric series from 2π to N×2 π. Then add it to the corresponding vector, and the calculation formula is as follows:

$$PE(pos, 2i) = \sin\left(\frac{pos}{1000^{2i/d_{modle}}}\right)$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{1000^{2i/d_{modle}}}\right)$$

where pos represents the absolute position of a character in the sentence.

Taking a sentence in Dongxiang as an example, the input phoneme sequence is "zh e4 sh $i4$ m $u4$ t $ou4$", then the pos of the phoneme "sh" in this sequence is 2;

d_{model} represents the dimension of the word vector; 2i and 2i+1 The parity is represented, and i represents the first dimension in the word vector. For example, the dimension of the word vector in the experiment is 512, then $i = 0,1,\ldots,255$.

2.2 Non-autoregressive speech synthesis model

In the introduction of the previous section, it can be found that the autoregressive generative model in the decoding stage synthesizes the speech recursively frame by frame, which cannot be parallelized, while the non-autoregressive model can parallelize the speech synthesis during the decoding process 28, It can greatly improve the speed of model training and speech synthesis 29.

2.2.1 Speech synthesis method based on FastSpeech

The model structure of $FastSpeech$ 30 is different from the conventional sequence-based encoder-attention mechanism-decoder structure, which mainly consists of Feed-forward Transformer (FFT), Length Regulator (LR) and The phoneme duration predictor (Duration Predictor) consists of three parts, and its network structure model is shown in Figure 2.3.

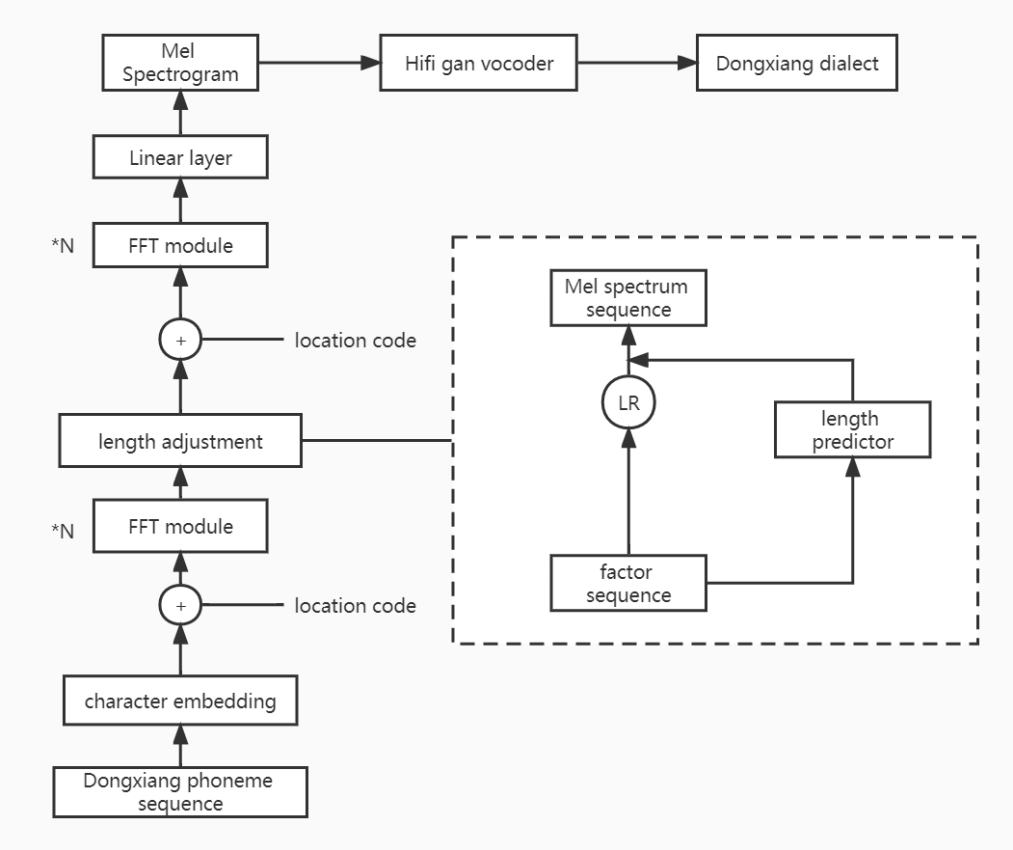


Figure 2.3 $FastSpeech + Hifi\ gan$ network structure

The structure of the feedforward transformer is a feedforward structure based on Transformer and one-dimensional convolution. As shown in the figure several FFT modules are used to build the relationship between the phoneme sequence and the chalk spectrogram. There are N FFT modules on each side of the phoneme and each side of the chalk spectrum, and there is a length corrector between the modules on either side to fill in the length gap between the phoneme and the chalk spectrum. Each of these FFT modules has a multi-headed attention and a two layer one-dimensional convolutional network with ReLU activation function. Since phoneme position, information is encoded by position in the transformer model, all multi-headed attention is used in the FFT to extract position information. The role of the one-dimensional convolutional network is to make the relationship between two adjacent latent states similar to the relationship between the phoneme and the chalk spectrum.

The function of the length adjuster is to solve the problem of length mismatch between the phoneme sequence and the Mel spectrum sequence in the feedforward transformer, and it can adjust the speech rate to increase prosody. The length of the phoneme sequence of a speech is less than the length of the Mel spectrum sequence, and a single phoneme may correspond to several segments of Mel spectrum, then the Mel spectrum length corresponding to a phoneme is called the phoneme duration. According to the phoneme duration d, the length regulator can copy the hidden state of the phoneme by d times to make the length of the hidden state of the phoneme equal to the length of the Mel spectrum. The calculation formula of the length adjuster is expressed as:

$$H_{mel} = LR(H_{pho}, D, a)$$

where H_{pho} is the hidden state of the phoneme sequence. Can be $[h1, h2, \ldots, hn]$, n represents the length of the phoneme sequence, D represents the phoneme duration of the corresponding phoneme in the phoneme sequence, can be $[d_1, d_2, \ldots, d_n]$, n represents the corresponding phoneme, H_{mel} represents the corresponding phoneme sequence the total length of the Mel spectrum sequence, α is a hyperparameter, and its function is to control the speech rate. It is specified that α=1

is the normal speed, and the value can be changed according to the rate that needs to be adjusted.

Assuming $H_{pho} = [h_1, h_2, h_3]$, , this is the phoneme duration $D = [2,1,3]$ corresponding to the phoneme sequence, then based on the normal speech rate, the phoneme sequence corresponds to $H_{pho} = [h_1, h_2, h_3, h_3, h_3]$, .

In training, in order to predict more accurate phoneme durations, an important part of the length regulator is the phoneme duration predictor, which consists of two layers of one-dimensional convolutions containing the $ReLU$ activation function, each layer contains normalized Dropout layers that normalize and prevent overfitting, and linear layers that can transform the output to a scalar. The phoneme duration predictor is jointly trained with the $FastSpeech$ model to predict the length of the Mel spectrum corresponding to each phoneme, and the mean square error (MSE) will be calculated between the prediction and the actual error.

2.2.2 Speech synthesis method based on FastSpeech2

The network model structure of FastSpeech2 is generally consistent with that of $FastSpeech$ 31. Based on it, the feature predictor part is mainly added. This part is mainly composed of three small parts, namely the duration predictor, the fundamental frequency predictor and the volume predictor, and its network structure is shown in Figure 2.4.

The purpose of the fundamental frequency predictor is to better predict changes in the fundamental frequency contour 32. The fundamental frequency of each syllable in a sentence can clearly express different emotional states, so the fundamental frequency feature is added to conduct experiments to It is very important to add emotion to the synthesized speech. The method consists in using the Continuous Wavelet Transform (CWT) to decompose a continuous fundamental frequency sequence into a fundamental spectrogram 33. The rhythm of the synthesized speech is not obvious, so CWT is used to convert the one-dimensional time domain fundamental frequency curve into a two-dimensional fundamental frequency spectrum. By fitting to the dimensional spectrum, the dynamics will be much better. The information lost in

forward and reverse CWT conversion is very small. Therefore, for speech synthesis with Inverse Continuous Wavelet Transform The target fundamental spectrogram is converted into a fundamental frequency curve. The MSE loss is used for optimization when training the fundamental frequency predictor.

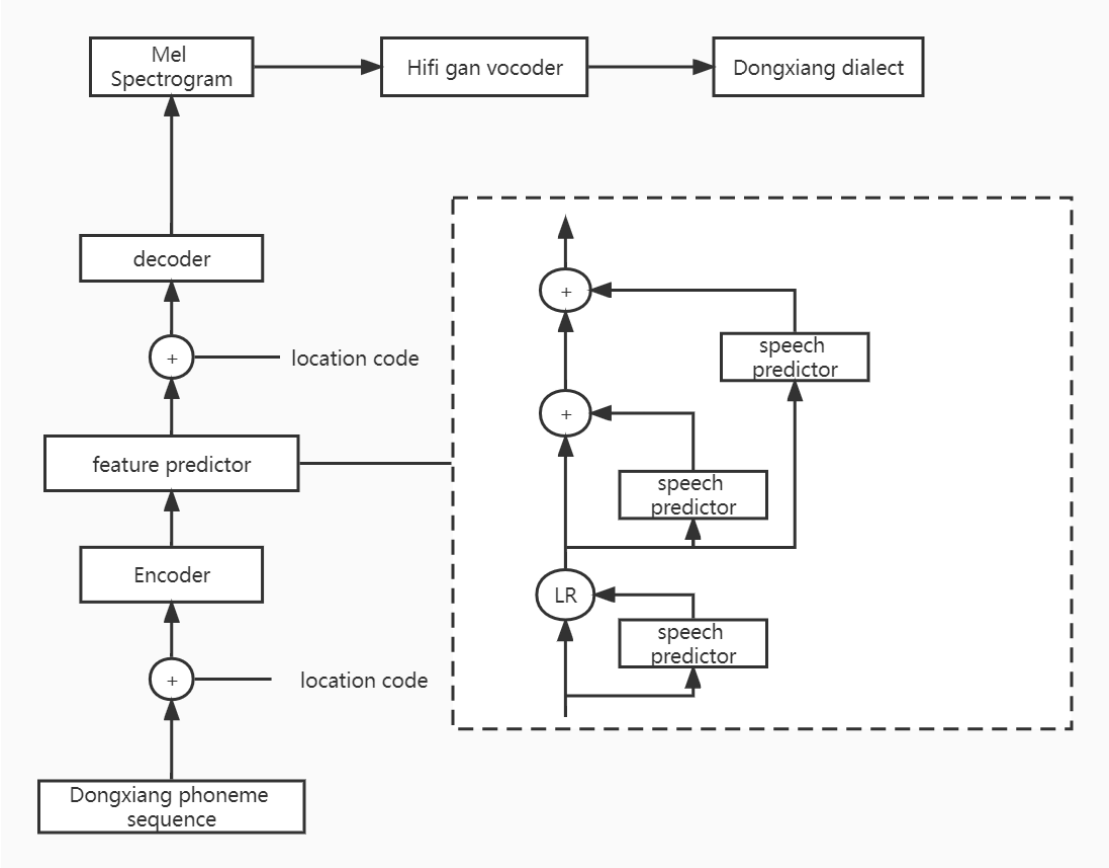


Figure 2.4 FastSpeech2+Hifi gan network structure

The volume predictor calculates the L2 norm of a Short-Term Fourier Transform (STFT) frame. Calculated as the volume during speech processing. The predictor then quantizes the volume of each frame by 256. Continuous value variation is also optimised with MSE loss. The main frequency remains unchanged. The main frequency remains unchanged. In the encoder stage, the phoneme duration information, fundamental frequency information and volume information are simultaneously encoded in the hidden sequence. Mel spectrum sequence for training.

2.3 Summary of this chapter

This chapter mainly introduces the synthesis methods that need to be used in the research, which are mainly divided into autoregressive models and non-

autoregressive models. In the case of large corpus training, the autoregressive model can make the model have good generalization ability due to the existence of the attention mechanism, while the non-autoregressive model has the advantage of high speed due to the fully parallel structure. The comparison experiment of the methods selects the most suitable synthesis method under the premise of low resources.

3 Corpus Construction and Preprocessing

3.1 Corpus establishment

3.1.1 Text Design of Corpus

Text design is the key to corpus construction, and scientifically selecting reasonable texts will make research more effective. There are 3200 sentences of bus station announcement corpus in the previous research, but due to the deep influence of modern Chinese loanwords on Dongxiang, the pronunciation of many emerging words is not much different from that of Chinese, such as computer, network, bus, etc. Cars, etc., and many words that have been used since ancient times, such as wood, feather, moon, etc., are pronounced in ancient Dongxiang language. Therefore, in order to make the corpus cover as many language characteristics of Dongxiang language as possible, the research is based on the analysis results in Chapter 3, Combined with the book 34, 400 authentic Dongxiang sentences were designed, which included all the phonemic features, lexical features and sentence pattern features of Dongxiang. Since there is no written expression in Dongxiang language, Chinese characters are used as the way of expression in the design, and each sentence has its corresponding Dongxiang International Phonetic Alphabet to facilitate subsequent annotation operations and experiments. The final Dongxiang language corpus contains 3200 bus station announcement corpora and 400 authentic Dongxiang sentences, a total of 3600 sentences.

3.1.2 Audio recording of the corpus

For the recording of Dongxiang dialect, we first searched for a number of Dongxiang students who master authentic Dongxiang dialect, and then the recording work was carried out in a recording studio, which could avoid a lot of environmental

noise. The recorded audio will be reviewed to ensure that every the validity of the audio, the inappropriate audio will be supplemented, and finally processed by the professional audio processing software Adobe Audition. The voice is saved in mono WAV format with 16kHz sampling rate and 16bit quantization precision.

3.1.3 Annotation of Corpus

In order to facilitate training, this paper uses Chinese characters as the expression text of Dongxiang language, and uses *Hanyu Pinyin* to mark Dongxiang phonemes. According to the Chinese notation system studied by predecessors, a set of Dongxiang language machine-reading phonetic transcription scheme (SAMPA-DX) 41 was designed, as shown in Table 3.1 below.

Table 3.1 Dongxiang language machine reading phonetic symbol scheme

Mandarin Chinese International Phonetic Alphabet	Dongxiang International Phonetic Alphabet	Dongxiang dialect
a	a	a
ə	ə	e
i	i	i
	ɯ	I
o	o	o
u	u	u
ai	ai	ai
əi	əi	əi
ao	ao	ao
əu	əu	əu
ia	ia	ia
iə	iə	iə
iu	iu	iu
ua	ua	ua
ui	ui	ui
iao	iao	iao
ia	ia	ia
	ang	AN

	eng	EN
	uan	uan
	on	on
	ou	ou
	un	un
	uo	uo
	ue	ue
p	b	b
p^h	p	p

Continued Table 3.1 Dongxiang language machine reading phonetic symbol scheme

Mandarin Chinese International Phonetic Alphabet	Dongxiang International Phonetic Alphabet	Dongxiang dialect
m	m	m
f	f	f
t	d	d
t^h	t	t
n	n	n
l	l	l
	r	r
	dz	z
ts^h	ts	c
s	s	s
	dʑ	J
$tɕ^h$	tɕ	q
	dʐ	zh
ʐ	ʐ	R
ɕ	ɕ	x
$tʂ^h$	tʂ	ch
ʂ	ʂ	sh
	g	g

k^h	k	k
x	x	H
	ɣ	y
	G	G
	q	K
	h	h
	j	j
	w	w

Pinyin transcription of Dongxiang phonetic symbols in the corpus was performed using SAMPA-DX. First, it is necessary to compare whether the Dongxiang phonetic symbols of the target sentence are consistent with the Hanyu Pinyin. If they are consistent, keep them. If they are inconsistent, search for the corresponding Hanyu Pinyin according to SAMPA-DX for replacement. The specific transcription process is shown in Figure 3.2 below.

Table 3.2 Dongxiang phonetic transcription example sentences

这	桌	是	木	头
dz̩ ə4	dz̩ uo1	ʂ i4	m u4	t ou4
zh e4	zh uo1	sh i4	m u4	t ou4

Table 3.2 is the transliteration example. It can be seen that each character is separated according to the phonetic and final of the Chinese character, and then the Dongxiang phonetic symbol and the phonetic final of the pinyin are compared. Transcribe all the corpus accordingly.

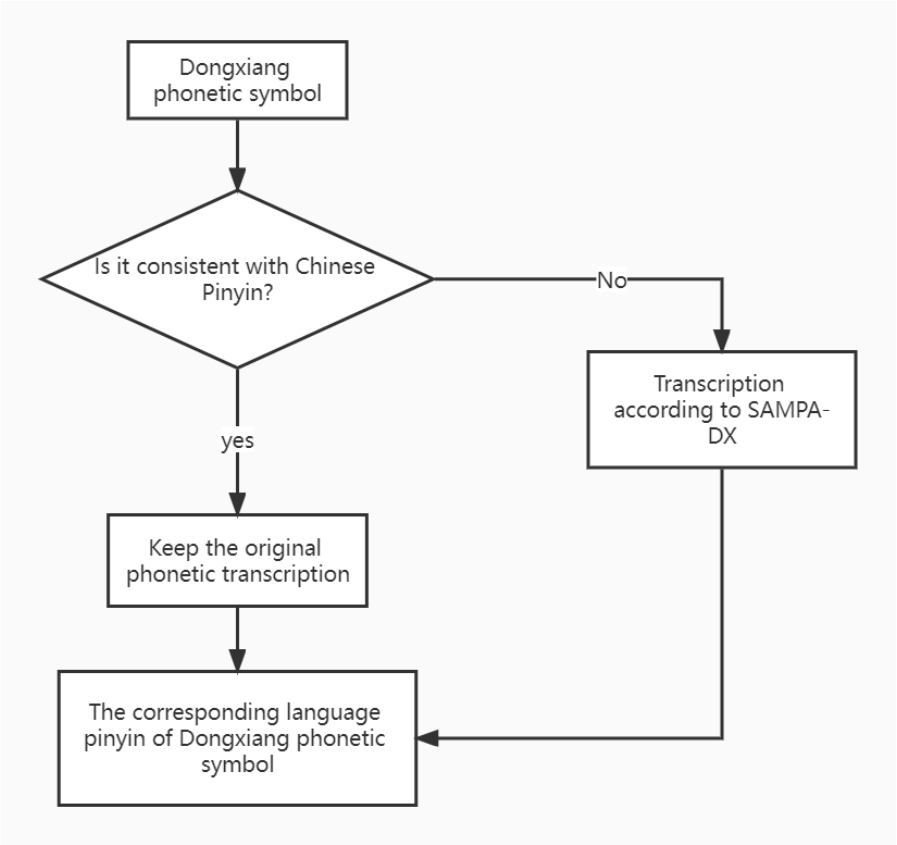


Figure 3.1 Dongxiang phonetic transcription pinyin flow chart

The research also needs to build a language dictionary suitable for model character encoding, refer to the Dongxiang phonetic symbols in SAMPA-DX, and then classify them according to the phonetic consonants. Each final corresponds to four tones, and the tones are marked after the final according to 1234. For example, uan1, uan2, uan3, uan4, as shown in Table 3.3, each phoneme corresponds to a number to ensure that different phonemes and different tones correspond to different numbers, so as to facilitate the conversion of phonemes into numbers for character embedding during training.

Table 3.3 Example of Dongxiang Character Encoding Dictionary

Code number	phoneme
31	Ou1
32	Ou2
33	Ou3
34	Ou4

35	Uan1
36	Uan2
37	Uan3
38	Uan4

3.2 Raw Speech Enhancement Processing

3.2.1 Speech enhancement preprocessing based on SEGAN

The purpose of speech enhancement techniques is to improve the intelligibility and clarity of noisy audio 42. Although the corpus audio is recorded in a recording studio, the non-professional recording conditions make the recorded audio more or less carry environmental noise, which affects the final synthesized Dongxiang speech quality. Enhanced preprocessing, using a pure and noise-free corpus and Dongxiang corpus for adversarial training, so that the voice quality of the Dongxiang corpus is as close to the pure corpus as possible. The purpose of preprocessing is to improve the original Dongxiang corpus. The voice quality, thereby indirectly improving the synthesis the biggest advantage of using generative adversarial network is that it automatically learns the data distribution of the original sample set 43.

The SEGAN network mainly includes two models 44 : the first is the Generative Mode (G), which mainly uses a joint distribution probability to generate the same distribution probability as the training sample. According to the input Dongxiang speech, the corresponding denoised pure Dongxiang speech is output. The second is the discriminative model (D), which will compare the existing pure voice with the pure voice generated by the generative model to see if the quality of the voice is consistent. If it is consistent, it will be output. Feed back to G, and adjust the parameters to regenerate until the generated voice quality reaches the standard. The flow chart is shown in Figure 3.2.

The advantages of using the SEGAN model are as follows:

1. There is no recursive operation like RNN, so the processing speed is fast.

2. It is an end-to-end model and is processed based on the original audio. Since no features are extracted, no changes are made to the original data.

3. Model training learns from different speakers and noise types and shares their parameters, which makes the whole system relatively simple.

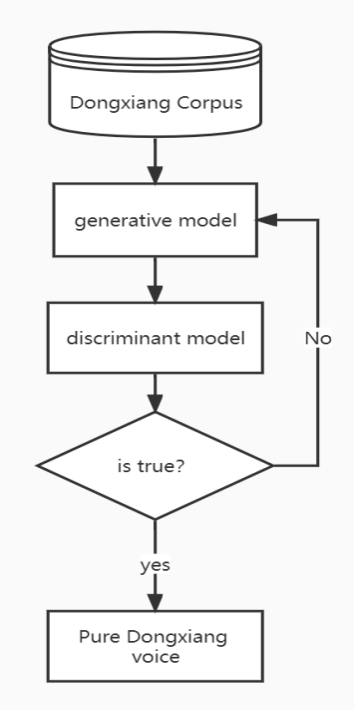


Figure 3.2 SEGAN processing flow chart

The way G learns the mapping is adversarial training, the other model D is usually a binary classifier, and the input of D is the imitation data and the real data output by G. The adversarial feature comes from the fact that D must classify the sample distribution of the real data as the real sample, and the imitation distribution generated by G is classified as the fake sample, which can create a situation where G is trying to fool D. When D identifies the data generated by G as a fake sample, D will back-propagate the information to G, so that G will continuously update the parameters to move towards the real sample distribution, so that D can classify the data generated by G as a real sample. In the training process, G must try to deceive D, and D must prevent being deceived. This process can be described as a minimax game 45 , and its objective function is:

$$min_G max_D V(D,G)$$
$$= E_{X\sim Pdata(x,x_r)}[\log D(x,x_C)] + E_{X_C\sim Pdata(x_r),z\sim P_r(r)}[\log(1 - D(G(z,x_c)))]$$

In the formula, $D(x)$ is the discriminator, which outputs the probability that

the result is true according to the input data, and $G(z)$ is the generator, which generates specific distribution data that conforms to the input of $D(X)$ according to the input data, and x is the model input. Data, xc is the real data set, training needs to maximize $D(G(z,xc))$, in fact, minimize $log\ 1\ -\ D(G(z,xc))$. In order to stabilize the training and improve the quality of the samples generated by G, the objective function is slightly changed, and the least squares binary code (1 is true, 0 is false) is used to replace the objective function [46], and the formula is changed as follows:

$$min_D V(D,G) = \frac{1}{2} E_{X\sim Pdata(x,x_r)}[D(x,x_c-1)^2] + \frac{1}{2} E_{X\sim Pdata(x_r),z\sim P_Z(Z)}[D(G(z,x_c-1))^2]$$

$$min_D V(D,G) = \frac{1}{2} E_{X\sim Pdata(x_r),z\sim P_Z(Z)}[D(G(z,x_c)-1)^2]$$

Through alternating adversarial training, the network will eventually converge to a Nash equilibrium state, that is, a stable state.

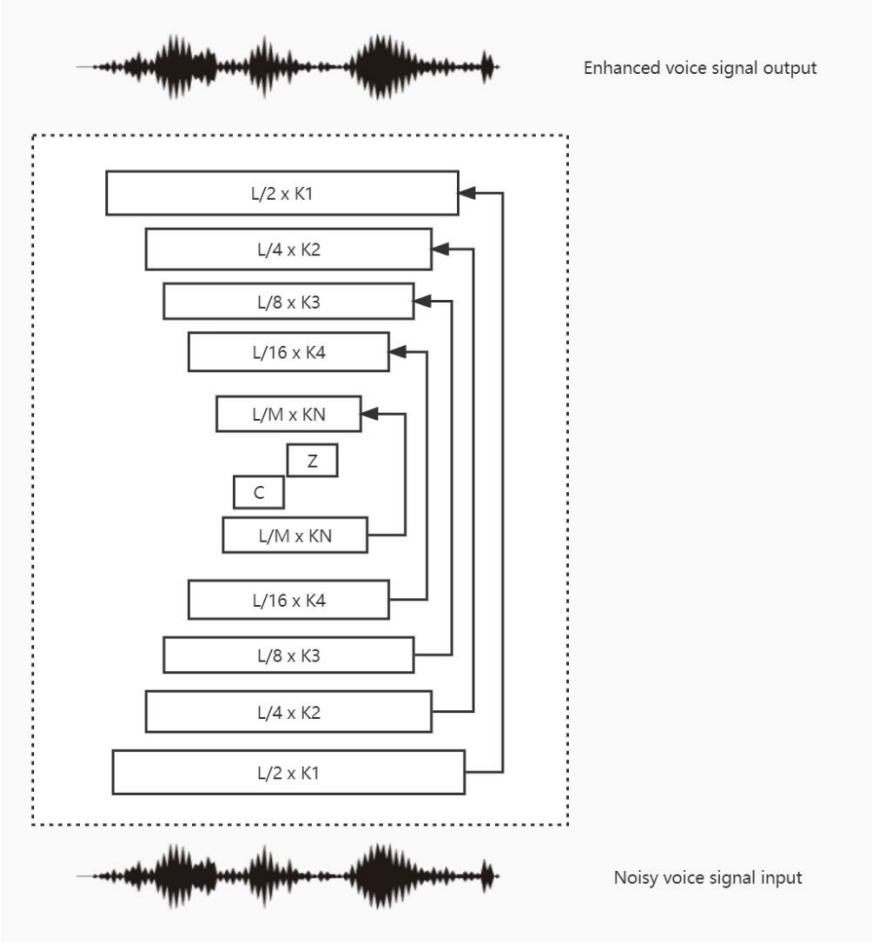


Figure 3.3 Generator G network structure diagram

Since the main function of D is to make two-category judgments, it will not be

elaborated. The main structure of the network is G, as shown in Figure 4.3, which is similar to an auto-encoder (Auto-encoder). In the encoding stage, the input signal passes through $Multi-layer$ stride convolutional compression, and then pass through the activation function $PReLUs$ 46 to obtain convolution results from each step of the filter. In the decoding stage, the strided convolution of the encoding process is transposed, followed by an activation function. The network has the characteristics of skip connection, that is, each individual encoding layer is connected to its corresponding decoding layer, bypassing the intermediate process. An important feature of the G network is that it is an end-to-end structure. When processing the original speech, all intermediate transformations are removed to obtain acoustic features. Here, it is necessary to pay attention to regression losses, such as mean absolute error or root mean square error. Then the solution to the problem is to use a generative adversarial setting to correct the output waveform according to the back-propagated information of D, remove the noise signal that D considers to be false, and finally generate qualified pure speech 48 .

3.2.2 Results and Evaluation of Preprocessing

The pure corpus used in the experiment uses the AISHELL-3 database. The corpus is recorded in a professional recording studio with a high-fidelity microphone (44.1 kHz, 16 bit), and has undergone strict quality inspection by professional voice proofreaders. Then use the original Dongxiang language corpus and AISHELL-3 for adversarial training, and set the parameters uniformly, as shown in Table 3.4:

Table 3.4 Parameter settings

parameter	Numerical value
G_learning rate	0.0002
D_learning rate	0.0002
Batch size	32
Epoch	1000
Optimzer	Adam
Label_smoothing	0.25

The evaluation adopts PESQ (Perceptual evaluation of speech quality) (perceptual evaluation of speech quality), which is an objective, full-reference speech quality evaluation method 49 . Computational comparison, here is to compare the original Dongxiang corpus with the enhanced corpus, which can objectively provide a MOS prediction value. PESQ scores range from 0.5-4.5, with higher scores indicating better quality. The evaluation results are shown in Table 3.5 below:

Table 3.5 Comparison of original corpus and SEGAN enhanced corpus

	original corpus	SEGAN enhanced corpus
PESQ	3.83	4.06

The evaluation results show that the speech quality of Dongxiang language preprocessed by SEGAN is higher than that of the original Dongxiang language corpus.

3.3 Sections of this chapter

This chapter first develops and extends the 400 authentic Dunxiang sentences based on the characteristics of the Dunxiang language, and completes the expansion of the entire Dongxiang language corpus through recording, Dongxiang phonetic transcription and pinyin, and then through the speech enhancement network SEGAN. The Dongxiang language corpus speech is enhanced and preprocessed, and the results show that the use of speech enhancement technology can improve the corpus speech quality and lay the foundation for the subsequent realization of high-quality Dongxiang language speech.

4 Dongxiang Speech Synthesis Based on End-to-End Model

Based on the content of the previous chapters, this chapter uses the end-to-end speech synthesis model to realize the speech synthesis of low-resource Dongxiang

language, and analyzes and compares the synthesis results of different models, and chooses the best one. The content of this chapter is mainly divided into six parts: The first part uses the $Tacotron2$ model to conduct synthetic experiments and summarize the problems. The second part uses the Transformer model to conduct experiments and summarize the problem. The third part uses the $FastSpeech$ model to improve the experiment to solve the problems in the first two parts. The fourth part uses the $FastSpeech2$ model to conduct optimization experiments to improve the synthesis quality. The fifth part evaluates all the experimental results subjectively and objectively. Section VI concludes this chapter.

4.1 Dongxiang Speech Synthesis Based on Tacotron2

The Tacotron2 model uses pre-processed Dongxiang speech data and text data transcribed by SAMPA-DX as input data for training. After the training, the Dongxiang speech is synthesized through the generated pre-training model. In the study, the vocoder is uniformly selected with $Hifi\ gan$, the reason is that the speech quality of the synthesized speech is smooth and clear and there is no noise. Finally, the evaluation and exploration are carried out through the synthesized results, and the experimental frame diagram is shown in Figure 4.1.

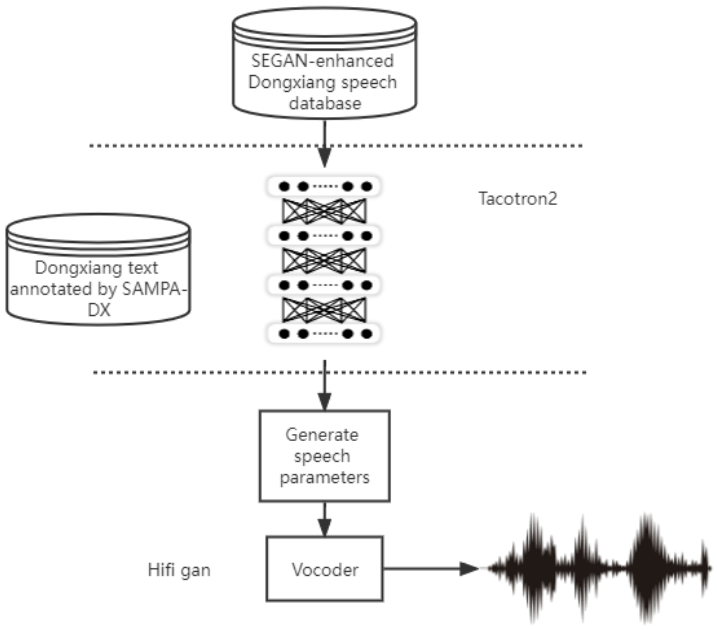


Figure 4.1 Experimental framework of Dongxiang speech synthesis based on Tacotron2

The experimental training data is divided according to the ratio of 9:1, that is, 3240 sentences are the training set, and 360 sentences are the test set. Experiments were conducted with 1000 batches of training in an environment where the CPU was Intel Xeon, the memory was 256GB, and the GPU was a single RTX2080Ti. The optimizer uses $Adam$ 50, the first-order estimated decay exponent is set to 0.9, the second-order estimated decay exponent is set to 0.999, and the learning rate is 0.001. This setting helps to optimize the model weights. Finally, before the training starts, the parameters are set uniformly, as shown in Table 4.1.

Table 4.1 Parameter settings

parameter	Numerical value
Learing rate	0.001
Dropout rate	0.5
Batch size	32
Epoch	1000
Optimizer	Adam

In the training process, it is found that due to the inherent order property of the recurrent network structure and the limitation of hardware memory, the batch processing of some long Dongxiang sentences will be hindered. And when the training is completed and tested with the test set, it is found that when encountering complex sentence patterns during speech synthesis, there will occasionally be a problem of repeated pronunciation of a certain word or missed reading of a certain sound. This problem is fatal in commercial use. It is easy to make the synthesized speech ambiguous or lose key information. Therefore, 100 sentences were extracted for testing in the experiment, and the robustness of the pre-training model was tested. The specific test conditions are shown in Table 4.2 below.

Table 4.2 Test statistics of wrong words, words and sentences

Model	skip word	omission	wrong sentence	Error rate

Tacotron2	10	13	10	33%

In terms of speech synthesis speed, the synthesis speed is slow due to the autoregressive model structure, which is unfavorable for some practical application scenarios. For example, when applied to a translator, when Dongxiang people are in dialogue with Han people, it takes too long for the sentences spoken by the Han people to be recognized, synthesized and played by the translator, which will greatly affect the efficiency of the dialogue. The experiment also made statistics on the synthesis rate. Select 20 Dongxiang sentences of different lengths to calculate the synthesis rate of the synthesized Mel spectrum and the overall rate of the synthesized speech with the vocoder, and take the average value. The specific synthesis rates are shown in Table 4.3.

Table 4.3 Speech synthesis rate

Model	Speech synthesis rate (ms)
Tacotron2	8.4±5
Tacotron2+Hifi gan	8.6±5

The speech quality of Dongxiang language synthesized by all experimental schemes will be assessed and compared in the last section of this chapter. Compared with traditional speech synthesis methods, the end-to-end speech synthesis model does not require complex acoustic feature data training, which greatly reduces the difficulty of training. And the voice quality is improved, but it still has some problems. In response to the problems in the experiment, the research has improved the experimental scheme. In the next section, the training model is changed to Transformer for experiments, because Transformer is a model with a non-cyclic structure. It completely relies on the influence of the attention mechanism on the input and output, and is also replaced with self-attention in the attention mechanism part, which allows the model to perform parallel computing, which greatly improves the computing efficiency. Coupled with the position encoding proposed by the model, it can be it is a good

solution to the processing problem of long sentences. Due to the improvement of the attention mechanism, whether it can solve the problem of missing words and skipping words remains to be studied, so after the end of the experiment, the same 100 Dongxiang sentences will still be used to test the robustness. The details will be introduced in the next section.

4.2 Dongxiang Speech Synthesis Based on Transformer

As mentioned in the previous section, since $LSTM$ needs to be calculated in order, due to the interdependent characteristics of long sentence sequences, it cannot effectively capture and cannot be calculated in parallel. In this section, Transformer is planned to be used for training and synthesis. Its structural advantage is that the multi-head attention mechanism (Multi-Head Attention) based on the self-attention mechanism in the model can capture the interdependent features in the long sequence, and the model can be trained in parallel with high speed efficient. The same training data is also used to train the network model, and the experimental frame diagram is shown in Figure 4.2.

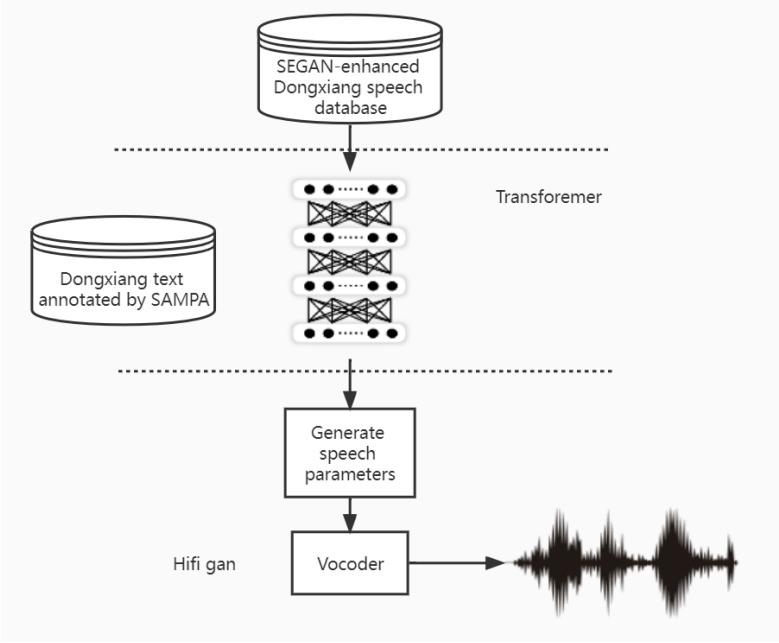


Figure 4.2 Experimental framework of Dongxiang speech synthesis based on Transformer

The experimental data is the same as the previous section. The training set and the test set are divided according to 9:1. The Dongxiang phoneme sequence is used as

the input of the encoder, and the Mel spectrogram is used as the input of the decoder. The two are in a one-to-one correspondence, that is, each Dongxiang phoneme A sequence has a Mel-spectrogram corresponding to it. The window length of each frame is 25ms, the frame shift is 10ms, and the mean and variance of each speech are normalized. The experiment is performed in an environment where the CPU is Intel Xeon, the memory is 256GB, and the GPU is a single RTX2080Ti. times of training. The specific training parameters are shown in Table 4.4 below:

Table 4.4 Parameter settings

parameter	Numerical value
Learing rate	0.0001
Dropout rate	0.5
Batch size	32
Epoch	1000
Optimizer	Adagrad
Dropout	0.1
Warmup steps	5000

During training, Transformer can process and train longer sentences in batches, which solves the problem of sentence length processing in Tacotron2. The Dongxiang language speech synthesis was performed on the trained Transformer model, and the same 100 sentences were used to test the problem of missing words and skipped words. The study found that although the use of Transformer reduced the error rate, it still could not solve this problem. The specific results are shown in the table below. As shown above, there are two main reasons for thinking deeply about this principle:

1. The amount of data in the Dongxiang language corpus is low, and the models trained through the attention mechanism under low resources have poor generalization ability and are prone to errors.

2. There is instability in the attention mechanism in the two autoregressive models, which leads to intolerable errors in the synthesized sentences.

Table 4.5 Test statistics of wrong words, words and sentences

Model	skip word	omission	wrong sentence	Error rate
Tacotron2	10	13	10	33%
Transformer	5	7	4	16%

In terms of speech synthesis speed, although the model structure of Transformer makes it possible to train in parallel, the speed is fast and efficient, but the speed of speech synthesis has not been greatly improved, but it is slightly better than Tacotron2, but it is far from practical. The specific results are shown in Table 4.6:

Table 4.6 Speech synthesis rate

Model	Speech synthesis rate (ms)
Tacotron2	8.4±5
Tacotron2+Hifi gan	8.6±5
Transformer	6.9±5
Transformer+Hifi gan	7.0±5

Although the improvement experiment has optimized the problems of Tacotron2, it has not been completely solved. After the experiments of the two schemes, there are mainly the following three problems that need to be solved:

1. There is instability in the attention mechanism, which leads to errors in synthetic sentences.

2. The generalization ability of the model trained by the network autonomous learning under the premise of low resources is poor.

3. Using the trained model to synthesize Dongxiang speech is slow, which affects the practicability.

4.3 Dongxiang speech translation based on FastSpeech

In order to solve the problem of autoregressive end-to-end speech synthesis model synthesizing Dongxiang, this section will use a large number of hand-extracted acoustic features and use $FastSpeech$ as an experimental model for research. The purpose of using manual extraction of acoustic features is to solve the instability of the model under low resources, and to use more accurate features instead of letting the model learn independently, to greatly improve the generalization ability of the model. The experimental block diagram is shown in Figure 4.3.

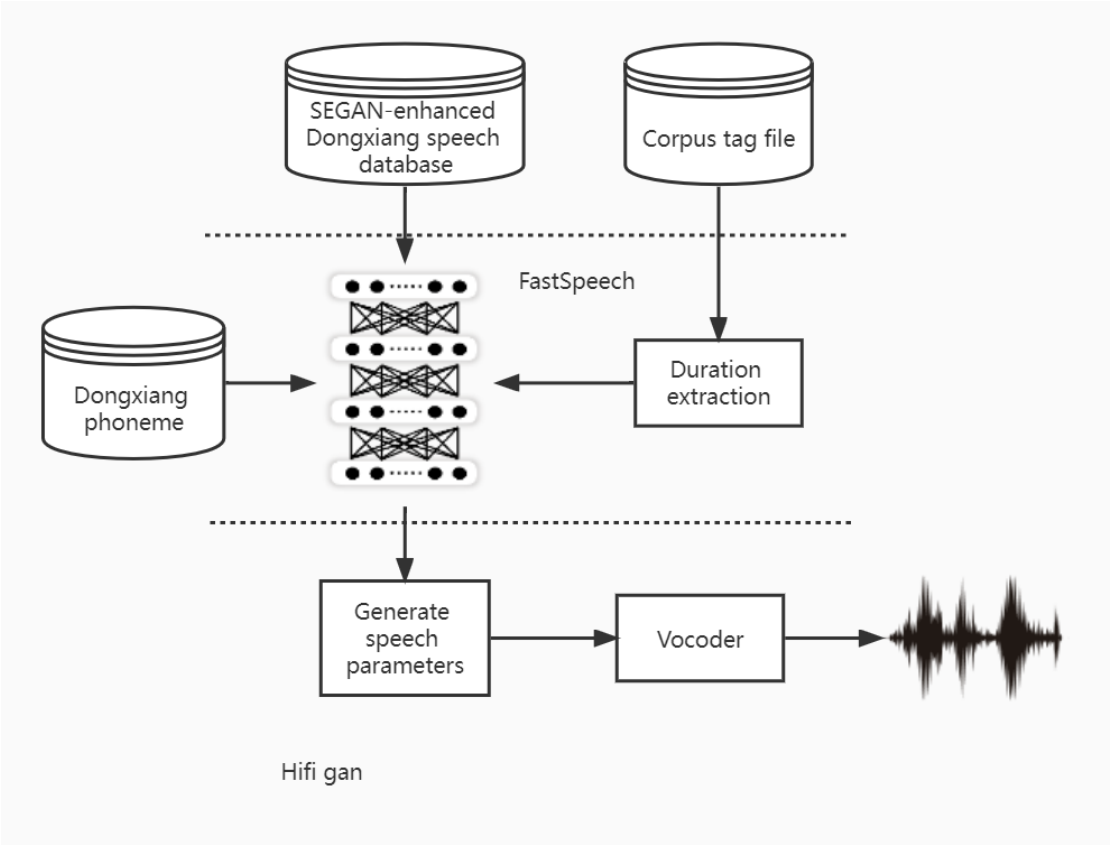


Figure 4.3 Experimental framework of Dongxiang speech synthesis based on $FastSpeech$

The main advantages of $FastSpeech$ are as follows:

1. The model is a parallel non-autoregressive structure, which can well solve the problem of slow speech synthesis rate.

2. In the experiment, the accurate phoneme duration will be manually extracted, and the phoneme and the Mel spectrum will be forced to be aligned according to the characteristics of the model, to solve the problem of missing words and skip words

caused by the instability of the attention mechanism.

3. The model can adjust the phoneme duration of the synthesized speech, that is, adjust the speed of the synthesized speech, to increase the rhythm of the synthesized speech.

$FastSpeech$ needs to extract the phoneme duration of the corpus in advance during training, while the conventional tool to extract the phoneme duration requires a large amount of speech data to train a model for extraction. The context information is extracted sentence by sentence, using the format of context annotation. The content of the extracted context information mainly includes phoneme duration, pronunciation information and prosody information, etc. According to the program design, the annotation file is generated. The duration information also includes the "mute" and "pause" parts as shown in Table 4.7. 100000 divide the extracted duration information. This is the frame number. 100000 also divide the difference between the end period and the start period of the monophonic. This is the frame number of that phoneme. For example, "sil" mute. Start time is 0, and end time is 4500000. The duration of "mute" is 45 frames. If the calculated number of frames is not an integer, such as 14.5 frames, then the final output will be 15 frames, because taking a small value may lose the voice. If the information is calculated according to 10 ms per frame, the extension of 5ms can be ignored. Finally, this Dongxiang language includes silence and pause, and outputs a phoneme duration array, namely [45, 3, 15, 12, 13, 8, 15].

Table 5.7 Phoneme Duration Array Extraction

duration information	**phoneme**	**Mono phone duration (frames)**
0-4500000	*sil*	45
4500000-4800000	*ch*	3
4800000-6250000	e	14.5
6250000-7400000	z	11.5
7400000-8700000	l	13
8700000-9500000	s	8
9500000-11000000	un	15

The experiment still uses a 9:1 division of the training set and the test set. All phoneme durations corresponding to the speech data in the Dongxiang corpus are extracted. The names are also in one-to-one correspondence as shown in Table 4.8. m01 represents male speaker #1 and 0411 represents the data number starting with 0411 and ending with 0811. It is guaranteed that each fragment of Dongxiang speech data in the preprocessing step for training matches the phoneme sequence and the phoneme durations array. When extracting phoneme durations, a more accurate duration can be extracted if calculated according to 1 frame of 5 ms, but when training the network, the sequence length will be doubled, affecting synthesis efficiency. Since an error of 5 ms has little effect on the experiment, it is calculated according to 1 frame of 10 ms.

Table 4.8 Pair nomenclature

Dongxiang language voice file name	Phoneme duration file name
DX_m01_0411.wav	DX_m01_0411_duration.npy
DX_m01_0412.wav	DX_m01_0412_duration.npy

During training, the network takes phoneme sequence, phoneme duration array and Mel spectrum sequence as input. The phoneme sequence is converted into digital code and input into the encoder through character embedding. In the controller, phoneme durations are expanded according to the phoneme durations array as shown in Table 4.9. If we take the Dongxiang sentence as an example, the synthesized Chinese is first converted into Dongxiang phonemes, then translated into Pinyin according to the phonetic character scheme. It is then split into consonants and endings, and then the phoneme durations of each consonant and endings are extracted according to the length regulator. After expansion, the length of the obtained sequence is consistent with the length of the corresponding Mel spectrum sequence. Finally, the linear layer outputs an 80-dimensional Mel spectrogram, and synthesizes the Dongxiang speech waveform through $Hifi\ gan$.

Table 4.9 Dongxiang phoneme duration extension

这(this)		桌(table)		是(yes)		木（wood）		头（head）	
$dz_{ɕ}$	ɘ4	$dz_{ɕ}$	uo1	ʂ	i4	m	u4	t	ou4
zh	e4	*zh*	uo1	*sh*	i4	m	u4	t	ou4
3	2	4	3	2	1	3	1	2	3
333	22	4444	333	22	1	333	1	22	333

Experiments were conducted with 2000 batches of training in an environment where the CPU was Intel Xeon, the memory was 256 GB, and the GPU was a single RTX2080Ti. Table 4.10 is the network parameter setting.

Table 4.10 Parameter settings

parameter	Numerical value
Learing rate	0.001
Dropout rate	0.1
Batch size	16
Epoch	2000
Optimizer	Adam
Vocab_size	200
Max_seq_len	1500

In order to verify whether the experiment can completely solve the problem of skipping or missing words in the autoregressive model, after training the model, use the same 100 sentences as the previous two sections for testing. The statistics are shown in Table 4.11. It can be seen that the forced alignment of the phoneme duration and the Mel spectrum can completely solve the errors in the synthesized Dongxiang language, and it also verifies that the learning of the sequence-to-spectrum mapping through the attention mechanism is not stable, especially on the premise of low resources. Down. Since the experiment uses the accurate duration of manual calculation and extraction,

the trained model has a high generalization ability.

Table 4.11 Test statistics of wrong words, words and sentences

Model	skip word	omission	wrong sentence	Error rate
Tacotron2	10	13	10	33%
Transformer	5	7	4	16%
FastSpeech	0	0	0	0%

In terms of speech synthesis rate, the $FastSpeech$ model does not need to predict and synthesize Dongxiang speech frame by frame due to the advantages of the fully parallel non-autoregressive structure, so the synthesis rate is hundreds of times faster than the autoregressive model. The specific results are as follows Table 4.12. This enables high communication efficiency in practical application scenarios.

Table 4.12 Speech synthesis rate

Model	Speech synthesis rate (ms)
Tacotron2+Hifi gan	8.6±5
Transformer + Hifi gan	7.0±5
Transformer+Hifi gan	**0.13±0.1**

The model can also change the speech rate parameter α of the length regulator to control the synthesized speech rate, and it can also adjust the duration of pauses in the middle of synthesized sentences to increase prosody. Express different emotions through different speech speeds and different pauses in a sentence. Figure 4.4 shows the comparison of the spectrograms of a sentence with normal speech rate and slowed down 1 times speech rate.

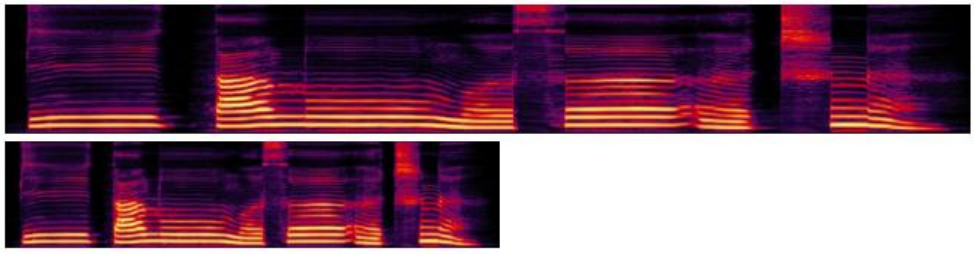


Figure 4.4 Comparison of normal speech rate and 0.5 times speech rate

After the $FastSpeech$ experiment is completed, it has basically realized the rapid synthesis of high-quality Dongxiang speech under low resources, but there is still room for improvement. In the $FastSpeech$ experiment, only the accurate phoneme duration was manually extracted, and the synthesized speech still lacks more. In order to make the synthesized speech both accurate and rich in emotion, in the next section, on the basis of extracting the accurate phoneme duration, the fundamental frequency and volume of the extracted speech will be used for experiments with $FastSpeech2$, and more features will be used. Synthesize more emotional Dongxiang speech under training.

4.4 Dongxiang Speech Synthesis Based on FastSpeech2

In order to make the synthesized Dongxiang voice more emotional, it is necessary to solve the problem of one-to-many mapping, that is, in the process of model training, a Dongxiang voice not only corresponds to its text, but also needs to correspond to the fundamental frequency of the voice. Acoustic features related to speech such as volume. Therefore, in this experiment, based on the accurate phoneme duration extracted in the previous experiment, the fundamental frequency and volume will be manually extracted as training data through $Praat$ speech processing software 51, so as to make the synthesized Dongxiang speech more accurate. Full of emotion, the experimental framework is shown in Figure 4.5.

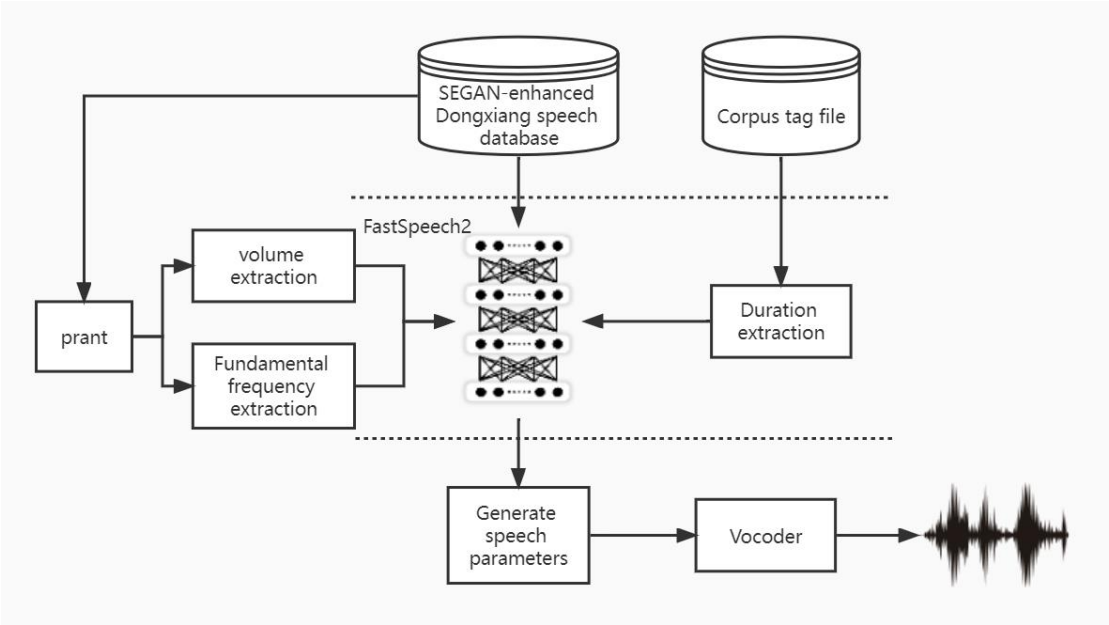


Figure 4.5 Experimental framework of Dongxiang speech synthesis based on FastSpeech2

Before the training starts, it is necessary to extract the fundamental frequency and volume of Dongxiang speech sentence by sentence. Use Praat speech processing software to first outline the fundamental frequency and volume curve, as shown in Figure 4.6, which is the corresponding diagram of the frequency spectrum and the speech waveform. The yellow line is the base frequency line, the blue line is the volume line.

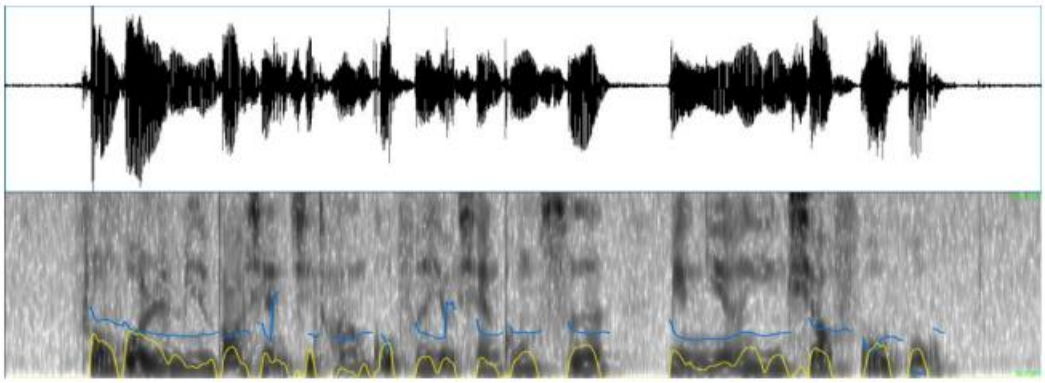


Figure 4.6 $Praat$'s analysis of speech feature map

Extract the value of the fundamental frequency according to 1 frame 10ms. Select all voices to get a list of fundamental frequencies, for example as shown in Table 4.13. Extract the fundamental frequency of each frame. Since the phoneme Dongxiang corresponds to more than one frame, the phoneme frequency must be tuned to the

fundamental frequency. Therefore, for the fundamental frequency of this phoneme, take the average of all its frequencies.

Table 4.13 Phoneme and fundamental frequency correspondence table

phoneme	phoneme duration	Baseband group	Average fundamental frequency
ch	3	237.3607 225.8213 213.0073	225.3964
e	2	205.7258 205.6729	205.6993
sp	2	0 0	0
z	1	209.0001	209.0001
i	3	209.1467 206.0690 208.1264	207.7807
sil	2	0 0	0

The volume will calculate the short-term Fourier energy extraction of each frame in the network speech preprocessing stage to calculate the average energy corresponding to each phoneme, and encode it into the hidden state sequence during training.

The experiment still uses a 9:1 division of the training set and the test set. According to the preparation work, the fundamental frequency and volume of the Dongxiang corpus speech are extracted, and the phoneme duration of the Dongxiang corpus extracted in the $FastSpeech$ experiment is combined. After preprocessing, each sentence of Dongxiang speech The Mel spectrum corresponds to its phoneme sequence, phoneme duration, fundamental frequency and volume file. The pairing

method is shown in Table 4.14. The training method, environment and parameters are the same as $FastSpeech$. The phoneme sequence, duration, fundamental frequency and volume array as input, the phoneme sequence is encoded in the latent state by character embedding in the encoder stage, and then the three features with corresponding embedding methods will be encoded into the latent state together during the training process. It is forced to align with the Mel spectrum sequence, and finally the Dongxiang speech is synthesized by the $Hifi\ gan$ vocoder according to the generated Mel spectrum.

Table 4.14 Pairing naming method

Dongxiang language voice file name	Phoneme duration file name
DX_m01_0411.wav	DX_m01_0411_duration.npy DX_m01_0411_pitch.npy DX_m01_0411_energy.npy
DX_m01_0412.wav	DX_m01_0412_duration.npy DX_m01_0412_pitch.npy DX_m01_0412_pitch.npy

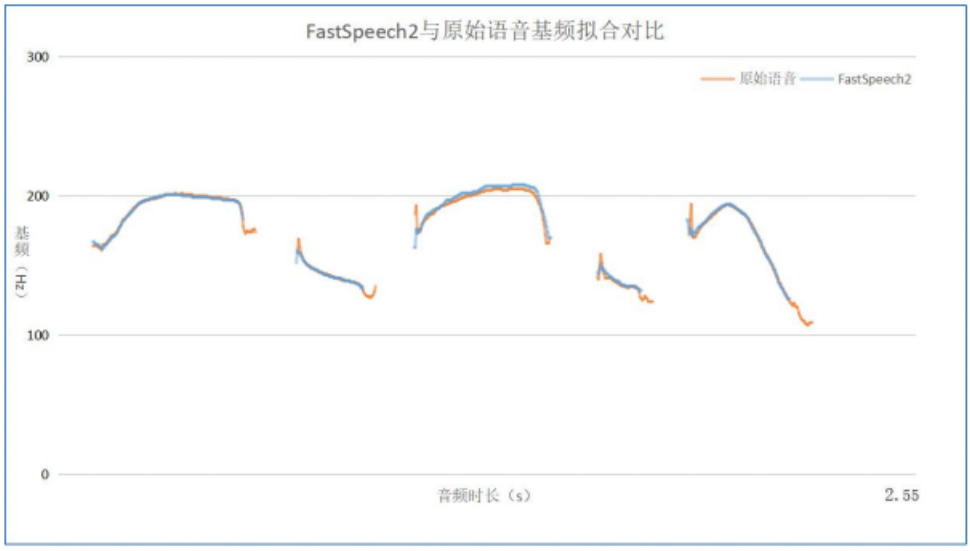
To evaluate whether the fundamental frequency of the synthesized Dongxiang language is accurate, the experiment will use the standard deviation, skewness coefficient and kurtosis of the fundamental frequency distribution of the synthesized speech and the original speech to test the accuracy, and also analyze the Dongxiang speech synthesized by the $FastSpeech$ model. For comparison, the results are shown in Table 4.15.

Table 4.15 Comparison of fundamental frequency parameters

Model	standard deviation	skewness coefficient	number of peaks
original voice	52.1	0.748	0.912
FastSpeech	48.2	0.623	1.213
FastSpeech2	**51.7**	**0.752**	**0.934**

It can be seen that the value of the main frequency parameter of Dongxiang speech synthesized by the FastSpeech2 model is closer to the value of the original speech than that of FastSpeech. The fundamental frequency contour of the synthesized speech is also compared and analysed (Figure 4.7). It can be seen that the main frequency curve set by FastSpeech2 fits more closely to the contour of the original speech. However, the fundamental frequency curve of $FastSpeech$ is quite different from the original speech, and the fitting effect is poor, and there will be problems such as fluctuation, instability and missing.

(a). $FastSpeech$ and original speech fundamental frequency fitting comparison



(b). FastSpeech2 and original speech fundamental frequency fitting comparison

Figure 4.7 Fundamental frequency fitting curve comparison

The volume fitting curves are also compared. As can be seen in Figure 4.8, in terms of volume fitting, the two models are not very good compared to the original speech, but $FastSpeech2$ is more similar. High, it is gentler in volume maintenance, while $FastSpeech$ fluctuates more and is not stable enough in volume maintenance.

(a). $FastSpeech$ and original speech volume fitting comparison

(b). FastSpeech2 and original voice volume fitting comparison

Figure 4.8 Volume Fitting Curve

To sum up, using a variety of hand-extracted acoustic features for training, the synthesized Dongxiang speech is more similar to the original speech and has emotion. The final experiment achieves high-quality and high-efficiency Dongxiang speech

synthesis with low resources.

4.5 Evaluation of Dongxiang Synthesized Voice Quality

4.5.1 Subjective evaluation

The subjective evaluation methods used in this study are Degradation Mean Opinion Score (DMOS) and Difference Mean Opinion Score (MOS) 52. 10 non-Dongxiang college students and 10 Dongxiang college students who master Dongxiang language were selected to form an evaluation team of 20 people to evaluate the Dongxiang language speech synthesized from the original speech of the Dongxiang language corpus and the experimental training model.

The evaluation process of DMOS is to play the original Dongxiang voice and the corresponding synthesized Dongxiang voice to the evaluation team, and inform the evaluation team in advance whether they are original voices. Scoring, the evaluation standard is shown in Table 4.16, and the final average is divided into the DMOS evaluation result.

Table 4.16 DMOS Evaluation Criteria

voice quality	similarity	Score
excellent	can't feel the difference	5
good	feel the difference but not resent it	4
middle	feel the difference but resent	3
bad	The difference is clearly felt but tolerable	2
extremely bad	intolerable differences	1

The evaluation will select 50 Dongxiang sentences and synthesize them through 4 experimental models and traditional methods respectively. The synthesized

Dongxiang speech and the original Dongxiang speech in the corresponding corpus form a total of 50 sets of evaluation data, and the synthesized speech of a single sentence and the original speech are one group. , each group is played in the order of the original voice, HMM, DNN and end-to-end speech synthesis model. The evaluation team scores subjective feelings according to the DMOS evaluation standard, separates Dongxiang students and non-Dongxiang students for statistical calculation, and finally takes the average value as the evaluation result, the DMOS evaluation result is shown in Figure 4.9.

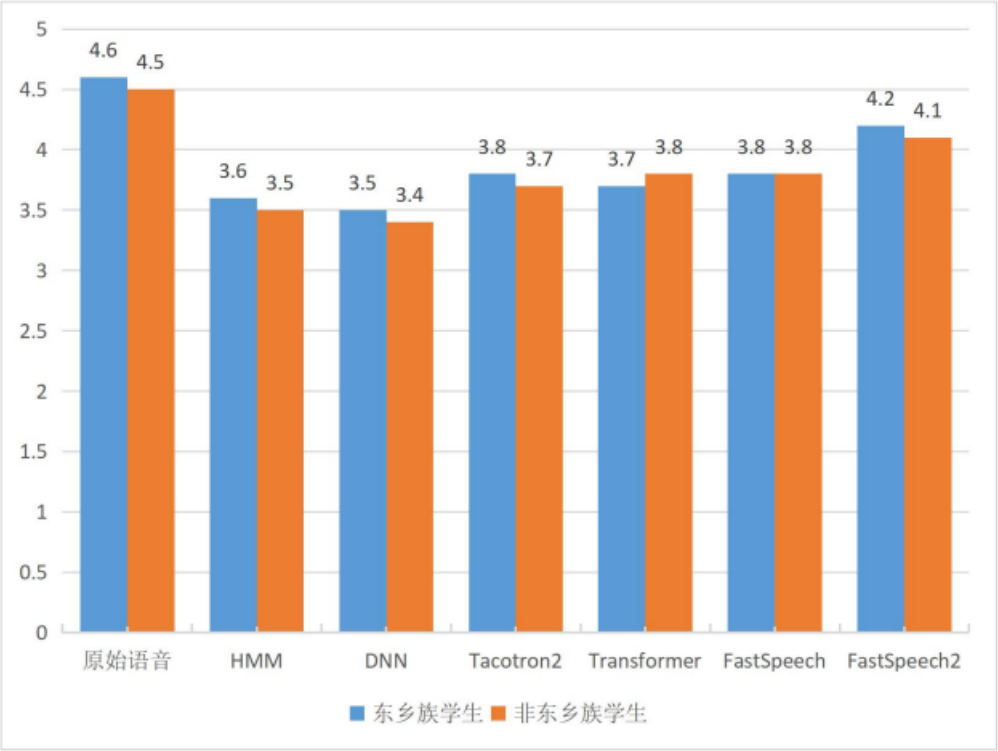


Figure 4.9 Dongxiang language DMOS evaluation

According to the results, it can be seen that the end-to-end model has better similarity compared with the traditional method under the premise of low resources, and among the four end-to-end models, due to the addition of more voice features to FastSpeech2, the similarity is greater, and the Dongxiang ethnic group has a higher similarity. of students think that the Dongxiang language synthesized by $FastSpeech2$ is more accurate in pronunciation, but there are still a small number of words with slightly changed pronunciation.

The evaluation process of MOS is to play the original Dongxiang language and

the corresponding synthesized Dongxiang language respectively without notifying whether they are original voices or not, then play the synthesized voices again, and then ask the evaluation team to evaluate the naturalness and feasibility of the synthesized voices. Distortion evaluation is performed on intelligibility, and the score is scored according to the MOS scoring standard.

Table 4.17 MOS Evaluation Criteria

voice quality	similarity	Score
excellent	can't feel the difference	5
good	Distorted but not disgusted	4
middle	Distorted but disgusted	3
bad	Distortion is noticeable but tolerable	2
extremely bad	intolerable distortion	1

The MOS evaluation also selects 50 pairs of speech data, mixes them, and finally tells which data is the original speech, and then plays the synthesized speech again. The evaluation team is required to score the naturalness and intelligibility of the synthesized speech according to the MOS evaluation standard. Finally, the statistical calculation is also separated, and the MOS evaluation result is shown in Figure 4.10.

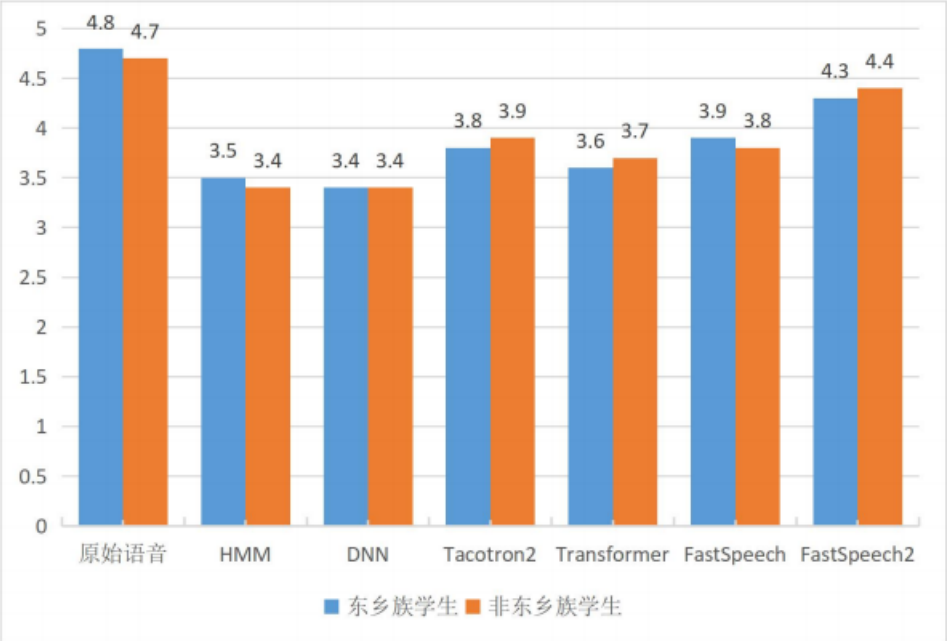


Figure 4.10 Dongxiang MOS evaluation

According to the results, the traditional method does not have good naturalness and intelligibility due to the problem of low resources. Some sentences still contain noise, while the end-to-end speech synthesis model has relatively significant improvement. The text synthesized by the FastSpeech2 model is the best.

In the final comparison, it can be concluded that under low resources, the subjective evaluation effect of the end-to-end speech synthesis model is better than that of the traditional method, and FastSeech2 makes the synthesized Dongxiang language rich in emotion and high through multi-feature training. Voice quality without synthesis errors.

4.5.2 Objective evaluation

The objective evaluation methods used in the study are the root mean square error of fundamental frequency, the root mean square error of phoneme duration, and the $PESQ$ evaluation.

The fundamental frequency is an important parameter of the speech signal. The experiment calculates the root mean square error (Root Mean Square Error, RMSE) 53 of the original Dongxiang speech in the corpus and the corresponding synthesized Dongxiang speech on the fundamental frequency to evaluate the relative performance

of the synthesized speech. The degree of distortion of the original speech, which is calculated as follows:

$$RMSE(F) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - \hat{F}_i)^2}$$

In the formula, $F_i$is the fundamental frequency value of a single phoneme in the synthesized speech, $\hat{F}_i$is the corresponding single element fundamental frequency value in the original speech, and N is the total number of phonemes.

The phoneme duration is the pronunciation duration of each phoneme. By calculating the $RMSE$ of the original Dongxiang speech and the synthesized Dongxiang speech in the phoneme duration, it is possible to evaluate whether the phoneme of the synthesized speech is accurate and the rhythm is good or not. The calculation method is related to the root mean square error of the fundamental frequency. the same, as follows:

$$RMSE(D) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(D_i - \widehat{D}_i)^2}$$

In the formula, $D_i$is the phoneme duration value of a single phoneme in the synthesized speech, $\widehat{D}_i$ is the phoneme duration value of the corresponding single element in the original speech, and N is the total number of phonemes.

In the objective evaluation, the 360-sentence test of the traditional method and the end-to-end model is combined into speech data and the original speech to calculate and compare the fundamental frequency RMSE, phoneme duration RMSE and PESQ. The results are shown in Table 4.18.

Table 4.18 Objective evaluation of Dongxiang language

Model	Fundamental RMSE	Phoneme duration RMSE	PESQ
HMM	8.69	0.56	3.24
DNN	4.82	0.52	3.13

Tacotron2	4.11	0.45	3.49
Transformer	4.09	0.42	3.54
FastSpeech	3.89	0.09	3.52
FastSpeech2	**1.82**	**0.04**	**3.81**

The first is the MRSE of the fundamental frequency. The table shows that without using the fundamental frequency features for training, even in the end-to-end model, there is still a large gap between the synthetic Dongxiang speech and the fundamental frequency of the original speech. In FastSpeech2, after and artificially extracting the exact fundamental frequency, the synthesized fundamental frequency of the speech is very close to the original speech. Then there is the phoneme duration, whether it is the traditional method or the autoregressive end-to-end model, due to the problem of low resources, the phoneme duration obtained by parameter modeling or the self-learning of the attention mechanism is not accurate, and the autoregressive model is extracted manually. Training with accurate phoneme duration can improve this problem very well. The PESQ results also show the advantages of the end-to-end model and the effectiveness of multi-feature training. The final result shows that the training of the FastSpeech2 model by manually extracting various speech features can synthesize emotional and high-quality Dongxiang speech.

4.6 Chapter Summary

Experiments on Dongxiang speech synthesis using different end-to-end models are conducted. An autoregressive end-to-end method of Dongxiang speech synthesis is proposed. The quality of the synthesized Dongxiang speech is higher than that of the traditional method. However, due to the structure of the autoregressive model and the instability of the attention mechanism, the quality of the synthesized Dongxiang speech is low. To improve the autoregressive speech synthesis model, a non-autoregressive end-to-end Dongxiang speech synthesis method is proposed. A large number of manually extracted parameters of duration, fundamental frequency and phoneme

loudness are used. It is shown that the FastSpeech2 model, using multifunctional learning, can provide efficient Dongxiang speech synthesis with better generalization ability.

5 Financial management, resource efficiency and resource saving

5.1 Assess commercial potential and prospects for conducting scientific research

5.1.1 QuaD Technology

QuaD (QUALITY ADVISOR) technology is a flexible tool for measuring characteristics that characterize the quality of a new development and its prospects in the market, and allows you to decide whether to invest money in a research project. In its content, this tool is close to the method for evaluating competitive technical solutions.

The basis of the QuaD technology is the search for the weighted average of the following groups of indicators:

1) Indicators for assessing the potential for commercial development:

- market prospects;
- prospects for design and production;
- financial efficiency;
- the impact of the new product on the performance of the company;
- availability for sale;
- legal protection, etc.

2) Indicators for evaluating the quality of development:

- an association;
- ergonomics;
- durability;
- energy efficiency;
- maintainability;
- the weight;

- dynamic range;
- the level of material consumption of development and others.

Indicators for assessing the quality and prospects of a new development are selected based on the selected object of study, taking into account its technical and economic characteristics of development, creation and commercialization.

To simplify the QuaD procedure, the evaluation is recommended in tabular form (Table 1). In accordance with QuaD technology, each indicator is evaluated by an expert on a one-point scale, where 1 is the weakest position and 100 is the strongest. Expertly determined weights should be 1 in total.

Scorecard for comparing competitive technical solutions (developments)

Criteria for evaluation	The weight criteria	Points	Maximum score	Relative value (3/4)	Weighted average (3x2)
1	2	3	4	5	
Development quality assessment indicators					
1. Noise immunity	0,18	30	100	0,3	5,4
2. Reliability	0,1	10	100	0,1	1
3. Ease of operation	0,12	80	100	0,8	9,6
4. The level of material consumption of development	0,15	15	100	0,15	2,25
5. energy efficiency	0,09	20	100	0,2	1,8
Indicators for assessing the commercial potential of development					
6. Availability of development certification	0,1	85	100	0,85	8,5
7. After-sales service	0,02	50	100	0,5	1
8. Price	0,09	63	100	0,63	5,67
9. Market penetration rate	0,07	97	100	0,97	6,79

10. Product Competitiveness	0,08	54	100	0,54	4,32
Total	1				46.33

We will assess the quality and prospects using QuaD technology using the formula:

$$К = \sum В_i \cdot Б_i ,$$

where Пср is the weighted average value of the indicator of quality and prospects of scientific development;

Bi – indicator weight (in fractions of a unit);

Бi – weighted average of the i-th indicator.

If the value of the PSR indicator is between 100 and 80, then this development is considered promising. If from 79 to 60 - then the prospects are above average. If from 69 to 40 - then the prospects are average. If from 39 to 20 - then the prospects are below average. If 19 and below - then the prospects are extremely low.

Based on the results of the assessment of quality and prospects, a conclusion is made about the amount of investment in current development and directions for its further improvement.

The technology can be used when conducting various marketing researches, which significantly reduces their labor intensity and increases the accuracy and reliability of the results.

5.1.2 SWOT analysis

SWOT - Strengths (strengths), Weaknesses (weaknesses), Opportunities (opportunities) and Threats (threats) - Is a comprehensive analysis of a research project. SWOT analysis is used to study the external and internal environment of the project.

At the third stage, the final SWOT-analysis matrix should be compiled, which is given in the postgraduate student's work (Table 2).

Table 2 - SWOT analysis

	strengths of the research project: C1. Environmental Safety. C2. Lower production cost compared to other technologies. C3. Resource efficiency. C8. Ease of use.	**Weaknesses of the research project:** Sl1. Lack of scientific development prototype Sl2. Lack of certification Sl3. Lack of promotion in the market. DC4. Financing DC5. There is no after-sales service.
Capabilities: IN 1. The emergence of additional demand for a new product. IN 2. Emergence of the cost of competitive developments. AT 3. The emergence of a new cash register.	development of a leak and spark complex that has higher quality indicators than those on the market (in particular and speed) in order to obtain a finished product with competitive advantages with optimal costs, high quality and engineering services.	1. Increasing the qualifications of personnel from potential consumers 2. Creation of engineering services for the purpose of training to work with the finished product 3. Purchase the necessary equipment for prototype testing 4. Shortage of supply or change of supplier
Threats: U1. Material (Inzka). U2. Lies in obtaining certification (high cost) U3. Lack of demand for new production technologies U4. Developed production technology competition U5. Technology Export Restrictions	1. Promoting the program to create demand 2. Creating a competitive advantage for the finished product 3. Product certification and standardization	1. professional development of personnel from potential consumers 2. Product certification and standardization 3. Purchase of necessary equipment for prototype testing 4. Promoting a program to create demand 5. Lack of supply or change of supplier 6. Creating competitive advantages for finished products 7. Creation of engineering services to learn how to work with the finished product

5.2 Identification of possible research alternatives

The morphological approach is based on a systematic study of all theoretically possible options arising from the laws of the structure (morphology) of the object of study. The synthesis covers both known and new, unusual options that can be missed with a simple search. By combining the options, a large number of different solutions are obtained, some of which are of practical interest.

The most successful combinations of this work were identified:

1. Universal; The first combination is universal. The resulting work will be done at no extra cost and will be of average quality.

2. Cheapest; The second combination is easy to implement, the least resources, cheapness of work. The result is cheap but poor quality work.

5.3 Project initiation

Charter of the scientific project of the master's work:

1. Goals and results of the project. Table 7 provides information on the hierarchy of project goals and criteria for achieving goals. The objectives of the project include goals in the field of resource efficiency and resource conservation.

Table 7 - Goals and results of the project

Project goals:	Intelligent voice transcription based on iFLYTEK
Expected results of the project:	It can realize basic speech transcription and improve the algorithm scheme to realize speech synthesis.

2. Organizational structure of the project. Table 8 provides members of the project working group.

Table 8 – Project working group

№ п/п	Name, main place of work, position	Role in the project	Functions
1	Botygin Igor Alexandrovich, Candidate of Technical Sciences, Associate Professor of the Department of Information Technology	Supervisor	Project planning, staff consulting

2	Yile Liu , student of TPU	Executor	Development of the whole project

5.4 Planning of research works.

Work plan. The list of works and the correspondence of the work to our performers, the duration of these works, are presented in Table 3.

Table 3 - List of works and duration of their implementation

the main topics of the project	№ раб	content of works	Position of the performer
Choice of research direction	1	Drafting and approval of the project theme	scientific adviser
	2	Analysis of the relevance of the topic	
Choice of research direction	3	Search and study of material on the topic	Student
	4	Choice of research direction	Scientific supervisor, student
	5	Work scheduling	
Theoretical studies	6	The study of literature on the topic	Student
	7	Selection of regulatory documents	
	8	Analysis of the means and cops used	
	9	Systematization and registration of information	
Evaluation of the results	10	Analysis of results	Scientific supervisor, student
	11	conclusion	scientific adviser

In most cases, labor costs account for the bulk of development costs, so it is important to determine the labor intensity of each of the participants in the study.

We determine the duration of the stages of work by an experimental-static method, which is implemented:

- analog method;

- probabilistically.

Определить ожидаемое значение продолжительности работы, используя

вероятностный метод - метод двух оценок t_{min} и t_{max}.

$$t_{ожс} = \frac{3 \cdot t_{\min} + 2 \cdot t_{\max}}{5}$$

where t_{max} – is the maximum falsity of work, person/day.

t_{min}–Minimum complexity of work, person/day;

The following specialists took part in the performance of the works listed in Table 5.1:

- engineer;
- scientific director.

Based on the expected labor intensity of the work, the duration of each work is determined in working days Tp, taking into account the parallel execution of the work by several performers. Such a calculation is necessary for a reasonable calculation of wages, since the share of wages in the total estimated cost of scientific research is about 65%.

$$T_{\text{р}i} = \frac{t_{\text{ожі}}}{\text{Ч}_i}$$

where $T_{\text{р}i}$ – Duration of one work, work. days;

$t_{\text{ожі}}$ – Expected labor intensity of one job, man-days

Ч_i –The number of performers performing the same work at the same time at this stage, pers.

For the convenience of the conspiracy, the duration of each of the stages of work from the working day should be translated into calendar days. To do this, use the following formula:

$$T_{\text{кі}} = T_{\text{рі}} \cdot k_{\text{кал}}$$

Where T_{ki}– Duration of youth work in calendar days;

T_{pi} – the duration of the i-th job in working days;

k_{kan}–calendar factor.

Table 5 shows the duration of the stages of work and the number of performers used in each stage.

Table 4 - Schedule of scientific research

Stage	Performers	Duration of work, days			Duration of work, person/day			
					$T_{РД}$		$T_{КД}$	
		t_{min}	t_{max}	$t_{ож}$	НР	И	НР	И
Formulation of the problem	НР	2	3	2,4	2,88	-	3,48	-
Development and approval of terms of reference (ТЗ)	НР, И	1	2	1,4	1,68	0,17	2,03	0,21
Selection and study of materials on topics	НР, И	15	18	13,2	4,75	14,26	5,75	17,25
Development of the calendar plan	НР, И	1	2	1,4	1,68	0,17	2,03	0,21
Selecting a block diagram of a device	НР, И	2	4	2,8	1,01	4,32	1,22	5,23
Choice of device circuit diagram	НР, И	8	10	8,8	3,17	10,56	3,84	12,78
Calculation of the circuit diagram of the device	И	3	5	3,8	-	4,56	-	5,52
Device layout development	И	20	22	20,8	-	24,96	-	30,2
Conducting experimental studies	НР, И	4	6	4,8	1,73	5,76	2,09	9,97
Correction of the parameters of the circuit diagram of the device	И	2	4	2,8	-	3,36	-	4,07
Registration of a settlement and explanatory note	И	7	10	8,2	-	9,84	-	11,91
Design of graphic material	И	2	4	2,8	-	3,36	-	4,07
Summarizing	НР, И	2	4	2,8	3,02	1,01	3,65	1,22
Total:				76,2	19,92	82,33	21,13	99,61

Table 5 - Calendar schedule for conducting R&D on the topic

Stage	Type of work	Performers	t_x	March	April	May	June
1	Formulation of the problem	НР	4				
2	Development and approval of terms of reference	НР,И	3				
3	Selection and study of materials on the subject	НР,И	18				
4	Development of the calendar plan	НР,И	3				
5	Selecting a block diagram of a device	НР,И	6				
6	Choice of device circuit diagram	НР,И	13				
7	Calculation of the circuit diagram of the device	И	6				
8	Device layout development	И	31				
9	Conducting experimental	НР,И	10				
10	Correction of the parameters of the circuit diagram of the device	И	5				
11	Registration of a settlement and explanatory note	И	12				
12	Design of graphic material	И	5				
13	Summarizing	НР,И	4				

-scientific director

-engineer

Based on Table 4, a calendar schedule was compiled. The schedule is built for the maximum duration of work within the framework of a research project based on Table 5 by months and decades (10 days) for the period of validity of the diploma. At the same time, work on the diagram should be highlighted with a different shade depending on the performers responsible for this or that work.

Scientific and technical research budget. Table 6 presents the cost of materials.

Table 6 - Material costs

Cd

Nr,

pcs

Tr,

Normo-h

Sm,

RR.

Name of detail	Cd	Nr, pcs	Tr, Normo-h	Sm, RUB
1. Piezo element (phased arrays)	*2*	*2*	*1*	*900*
2. Microcontroller	*1*	*1*	*0,3*	*300*
3. Indicator	*1*	*1*	*1,5*	*100*
4. Amplifier	*2*	*2*	*0,3*	*60*
5. Supply system	*1*	*1*	*1*	*650*
6. Pay	*2*	*2*	*2*	*20*
Total for materials				*2100*
Transport and procurement costs (15%)				*315*
Total for Wm				*2415*

Electricity is also included in the costs. The operating time of the equipment is

calculated based on the data for the TRD of Table 5.2 for the engineer, on the basis that the length of the working day is 8 hours.

The power consumed by the equipment is determined by the formula:

$$P_{ОБ} = P_{УСТ.ОБ} К_С$$

where $P_{ОБ}$ the installed capacity of the equipment, kW;

Kc - demand coefficient, depending on the number, loading of groups of electrical receivers.

For technological equipment of low power Kc=1. Electricity costs for technological purposes are shown in Table 8.

Table 8 - Electricity costs for technological purposes

equipment identification	**Equipment operation time tOB, hour**	**Power consumption $P_{ОБ}$, kW**	**Xpenses $Э_{ОБ}$, rub.**
Soldering Station	35	0,05	9.22
Source of power	5	0,4	10.54
Personal Computer	160	0,3	252,96
Total:			273.72

Э=273.72 RUB

Depreciation deductions are calculated for the period of use of the equipment according to the formula:

$$C_{AM} = \frac{H_A Ц_{ОБ}}{F_Д} t_{BT}^n$$

Where H_A – annual depreciation rate,

$Ц_{ОБ}$ – equipment price,

$F_Д$ – valid annual fund of working time,

$t_{ВТ}$ – operating time of computing equipment when creating a software product,

COMPUTER

$$H_A = 25\%;$$

$$Ц_{ОБ} = 26000\ RUB$$

$$F_Д = 2422\ hours$$

$$t_{BT} = 150\ hours$$

$$n = 1.$$

$$C_{AM2} = \frac{H_A Ц_{ОБ}}{F_Д} t_{BT} n = 413{,}9\ RUB$$

Soldering station:

$$H_A = 25\%;$$

$$Ц_{ОБ} = 1000\ RUB$$

$$F_Д = 2415\ hours$$

$$t_{BT} = 36\ hours$$

$$n = 1.$$

Power source:

$$H_A = 25\%;$$

$$Ц_{ОБ} = 25000\ RUB$$

$$F_Д = 2417\ hours$$

$$t_{BT} = 5\ hours$$

$$n = 1.$$

$$C_{AM3} = \frac{H_A Ц_{ОБ}}{F_Д} t_{BT} n = 12.93 руб$$

So the depreciation expense was:

$$C_{AM} = C_{AM1} + C_{AM2} + C_{AM3} = 413.9 + 3.62 + 12.93 = 430.45\,руб$$

The basic salary (Зосн) of the head (laboratory assistant, engineer) from the enterprise (if there is a head from the enterprise) is calculated by the formula:

$$Зосн = Здн \cdot Траб,$$

where Зосн – basic salary per employee;

Тр – duration of work performed by a scientific and technical worker, slave. days;

Здн – average daily wage of an employee, rub.

The average daily wage is calculated by the formula:

$$З_{ДН} = \frac{З_{М} \times М}{F_{Д}}$$

where $З_{М}$ – monthly salary of an employee, rub.;

$З_{ДН}$ – number of months of work without vacation during the year (with a six-day week 10.4);

$F_{Д}$ – actual annual fund of working hours of scientific and technical personnel ($F_{Д}$=1794 hours/year/person = 236 working days/year/person)

This item includes the amount of payments provided for labor legislation, for example, payment of regular and additional holidays; payment of time associated with the performance of state and public duties; payment of remuneration for long service, etc. (on average - 12% of the amount of the basic salary).

Employee's monthly salary:

$$Зм=Зтс \cdot Кр=26300 \cdot 1,3=34190 \text{ RUB.},$$

where Зтс – wages at the tariff rate, rub.;

кр – regional coefficient equal to 1.3 (for Tomsk).

And you can get the average daily salary of a supervisor:

$$З_{ДН} = \frac{З_{М} \times М}{F_{Д}} = \frac{34190 \times 10,4}{236} = 1506,68\,RUB$$

And the basic salary ($З_{осн}$) of the head:

Зосн=Здн · Траб=1506,68 · 38=57253,76 RUB

The student during the passage of undergraduate practice receives equal to 20,000 rubles / month. The average daily pay is:

$$З_{ДН} = \frac{З_М \times М}{F_Д} = \frac{20000 \times 10.4}{236} = 881.36\ RUB$$

The main income of a student, during pre-diploma practice, is equal to:

Зосн=Здн · Траб=881.36 · 38=33419.68 руб.

Additional wages are calculated on the basis of 12-20% of the basic wages of employees directly involved in the implementation of the topic:

Здоп=Кдоп · Зосн=57253,76 · 0,12=6870,36 RUB,

Where Здоп – additional salary, RUB;

kдоп – additional salary ratio;

Зосн – basic salary, rub.

The article includes deductions to off-budget funds:

Звнеб=Квнеб · (Зосн + Здоп)=0,302 · (57253,76 +6870,36)=19365,48 RUB

Where kвнеб – coefficient of deductions for payment to off-budget funds (extra-budget = 0.302.).

Overhead costs are calculated according to the following formula:

Снал=Кнал · (Зм+ Зобор + Зосн + Здоп + Звнеб + Знауч + Зонт),

Where kнакл – overhead ratio. The value of the overhead ratio can be taken in the amount of 16%.

Table 7 - Basic salary costs

Article title	Scientific supervisor RUB	Master RUB	Total RUB
The cost of the basic salary of the performers of the theme	57253	33420	65618

Costs for additional wages of theme performers	6870	1000	7870
Deductions to off-budget funds	19365	10395	19270
Overheads	18815	-	18815
Cost budget	102208	44815	111573

The item "Overhead" reflects the costs of project development, which are not taken into account in the previous articles.

$$C_{ПРОЧ} = (C_{ПОЛН} + C_{СОЦ})0,5$$

*Сн=(*74974.76+22492.428*)0,5=46733.584* RUB

5.5 Evaluation of the comparative effectiveness of the study

The determination of efficiency is based on the calculation of an integral indicator of research efficiency. Its location is linked to the determination of two weighted averages: financial efficiency and resource efficiency. The integral financial indicator of development is defined as:

$$I_{финр}^{исп.i} = \frac{Ф_{pi}}{Ф_{\max}},$$

where $I_{финр}^{исп.i}$ – integral financial indicator of development;

$Ф_{pi}$ – cost of the $i\text{-}th$ version;

$Ф_{\max}$ –The maximum cost of a research project (including analogues). (including analogues).

The integral indicator of the resource efficiency of the variants of the object of study can be determined as follows:

$$I_{pi} = \sum a^i \cdot b^i,$$

where: I_{pi} –Integral Resource Effectiveness Index for Youth Development Version;

a^i – Weight coefficient of the i-th development version;

b_i^a, b_i^p – The ball score for the i-th version of the development is set by expert evaluation of the selected evaluation scale;

n – Number of comparison parameters.

Table 9. Comparative evaluation of the characteristics of project implementation options

Object of study Criteria	Parameter weighting factor	Исп.1	Исп.2
1. Versatility	0,2	4	4
2. Reliability	0,2	5	4
3. Functional power	0,2	5	5
4. Material intensity level	0,15	4	4
5. Energy saving	0,1	5	5
6. Maintainability	0,15	4	4
TOTAL	1		

The integral indicator of the effectiveness of development options ($I_{ucni.}$) is determined on the basis of the integral indicator of the effectiveness of resources and the integral financial indicator according to the formula:

$$I_{ucni.} = \frac{I_{p-ucni}}{I_{финр}^{исп.i}},$$

Comparison of the integral indicator of the effectiveness of development options will allow determining the comparative effectiveness of the project (see Table 18) and choosing the most appropriate option from the proposed ones. Comparative project efficiency ($Э_{ср}$):

Table 10. Comparative development efficiency

No.n/a	Indicators	Ex.1	Ex.2
1	Integral indicator of development resource efficiency	0,9	0,62
2	Integral financial indicator of development	4,5	4,55
3	Comparative efficiency of variants	5,625	8,75
4	Integral efficiency indicator	0,64	1

Table 9 and Table 10 show (Where Ex.1 is a newly developed method and Ex.2 is a method of the same type. For example, Huffman coding uses more powerful single-chip microcomputers and other components when creating circuits for the Huffman algorithm. At the same time when processing data on a computer, this increases the CPU load and uses more resources.) that the complex indicator of resource efficiency of the new compression method for data received from ultrasonic sensors is higher than that of similar compression methods. Therefore, in comparison with similar methods, the developed new method saves resources and is easy to operate.

6 Social responsibility

6.1 Industrial safety and Introduction：

Social responsibility is the responsibility of an individual scientist and the scientific community to society. Of paramount importance is the safety of the use of technologies that are created on the basis of scientific achievements, the prevention or minimization of possible negative consequences of their use, and the provision of safe research for both the subjects and the environment.

In the course of this work, methods for compressing data obtained from ultrasonic sensors were developed and studied. This work is carried out in a laboratory and all work is

done with the help of computers. This section also includes an assessment of working conditions at the workplace, an analysis of harmful and harmful labor factors and the development of measures to protect against these factors.

6.1.1 Deviation of microclimate indicators in the room

Let's analyze the microclimate in the room where the workplace is located. The microclimate of industrial premises is determined by the following parameters: temperature, relative humidity, air velocity. These factors affect the human body, determining its well-being.

Optimal and permissible values of microclimate parameters are given in Tables 1 and 2

Table 1 - Optimal microclimate standards

Period of the year	Air temperature, C°	Relative humidity, %	Air speed, m/s
Cold	19-23	40-60	0.1
Warm	23-25		0.1

Table 2 - Permissible microclimate standards

Period of the year	Air temperature, C°		Relative humidity, %	Air speed, m/s
	Lower permissible limit	Upper permissible limit		
Cold	15	24	20-80	<0.5
Warm	22	28	20-80	<0.5

Warm Temperature in the warm season 23-25°C, in the cold season 19-23°C, relative humidity 40-60%, air speed 0.1 m/s.

The total area of the working room is 42m2, the volume is 147m3. According to Сан ПиН 2.2.2 / 2.4.1340-03, sanitary standards are 6.5 m2 and 20 m3 of volume per person. Based on the above data, we can say that the number of jobs corresponds to the size of the premises according to sanitary standards.

After analyzing the overall dimensions, consider the microclimate in this room. Consider temperature, air humidity, and wind speed as parameters of the microclimate.

The room is naturally ventilated through the presence of an easily opened window opening (windows), as well as a doorway. According to the coverage area, such ventilation is general exchange. The main disadvantage is that the supply air enters the room without preliminary cleaning and heating. According to SanPiN 2.2.2 / 2.4.1340-03, the volume of air required per person in a room without additional ventilation should be more than 40 m3 [1]. In our case, the volume of air per person is 42 m3, which means that additional ventilation is not required. The microclimate parameters are maintained in the cold season by water heating systems with water heating up to 100°C, and in the warm season by air conditioning, with parameters according to [2]. The normalized parameters of the microclimate, the ionic composition of the air, the content of harmful substances must comply with the requirements [3].

6.1.2. Exceeding noise levels

Noise is one of the most common hazards in manufacturing. It is created by working equipment, voltage converters, fluorescent work lamps, and penetrates from the outside. Noise causes headache, fatigue, insomnia or drowsiness, weakens attention, memory deteriorates, reaction decreases.

The main source of noise in the room are computer-cooling fans. The noise level varies from 35 to 42 dBA. According to SanPiN 2.2.2 / 2.4.1340-03, when performing basic work on a PC, the noise level at the workplace should not exceed 82 dBA [4].

At values above the permissible level, it is necessary to provide personal protective equipment (PPE) and collective protection equipment (CPE) against noise.

Means of collective protection:

1. elimination of the causes of noise or its significant weakening in the source of education;
2. isolation of noise sources from the environment (use of silencers, screens, sound-absorbing building materials);
3. the use of means that reduce noise and vibration in the way of their propagation;

Individual protection means;;

1. the use of overalls and hearing protection: headphones, ear plugs, antiphons.

6.1.3. Increased level of electromagnetic radiation

The source of electromagnetic radiation in our case are PC displays. The computer monitor includes radiation in the X-ray, ultraviolet and infrared regions, as well as a wide range of electromagnetic waves of other frequencies. According to SanPiN 2.2.2 / 2.4.1340-03, the electromagnetic field strength in terms of the electrical component at a distance of 50 cm around the ВДТ should not exceed 25 V/m in the range from 5 Hz to 2 kHz, 2.5 V/m in the range from 2 to 400 kHz [1] . The magnetic flux density should not exceed 250 nT in the range from 5 Hz to 2 kHz, and 25 nT in the range from 2 to 400 kHz. The surface electrostatic potential should not exceed 500 V [1]. In the course of the work, a PC of the Acer VN7-791 type was used with the following characteristics: electromagnetic field strength 2.5 V/m; the surface potential is 450 V (the basics of fire protection of enterprises GOST 12.1.004 and GOST 12.1.010 - 76) [5].

With long-term constant exposure to an electromagnetic field (EMF) of the radio frequency range when working on a PC, the human body has cardiovascular, respiratory and nervous disorders, headaches, fatigue, deterioration in health, hypotension, changes in cardiac muscle conduction. The thermal effect of EMF is characterized by an increase in body temperature, local selective heating of tissues, organs, cells due to the transition of EMF to warm energy.

Maximum permissible levels of exposure (according to *OST* 54 30013-83):

a) up to 10 μW/cm2, operating time (8 hours);

b) from 10 to 100 μW/cm2, operating time no more than 2 hours;

c) from 100 to 1000 μW/cm2, operating time no more than 20 minutes. on condition use of safety glasses;

d) for the population as a whole, the PPM should not exceed 1 μW/cm2.

Protection of a person from the dangerous effects of electromagnetic radiation is carried out in the following ways:

CPE (collective protective equipment)

1. time protection;

2. distance protection;

3. reduction of radiation intensity directly in the radiation source itself;

4. source shielding;

5. protection of the workplace from radiation;

PPE (personal protective equipment)

1. Glasses and special clothing made of metallized fabric (chain mail). It should be noted that the use of PPE is possible during short-term work and is an emergency measure. Daily protection of operating personnel must be ensured by other means.

2. Glass coated with a thin layer of gold or tin dioxide (SnO2) is used instead of ordinary glasses.

6.1.4. Electric shock

The dangerous factors include the presence in the room of a large number of equipment using a single-phase electric current with a voltage of 220 V and a frequency of 50 Hz. According to the danger of electric shock, the room belongs to the premises without increased danger, since there is no high humidity, high temperature, conductive dust and the possibility of simultaneous contact of current-carrying elements with grounded metal cases of equipment.

The laboratory refers to a room without an increased risk of electric shock. Safe ratings are: I < 0.1 A; U < (2-36) V; Rground < 4 ohm. The following protection measures against electric shock are applied indoors: inaccessibility of current-carrying parts for accidental contact, all current-carrying parts are isolated and fenced. Grounding, grounding of electrical equipment is used. The inaccessibility of current-carrying parts is achieved by their reliable

isolation, the use of protective fences (casings, covers, grids, etc.), the location of current-carrying parts at an inaccessible height.

Everyone needs to know the measures of medical care in case of electric shock. In any work place, it is necessary to have a first aid kit for first aid.

Careless handling of appliances, faulty electrical installations or damage to appliances most often causes electric shocks.

To release the victim from live parts, it is necessary to use non-conductive materials. If, after releasing the victim from stress, he does not breathe, or his breathing is weak, it is necessary to call an ambulance team and provide the victim with first aid:

- provide access to fresh air (remove tight clothing from the victim, unbutton the collar);
- clear the airways;
- start artificial ventilation of the lungs (artificial respiration);
- if necessary, start an indirect heart massage.

Any electrical appliance must be immediately de-energized in the event of:

- a threat to human life or health;
- the appearance of a smell characteristic of burning insulation or plastic;
- the appearance of smoke or fire;
- the appearance of sparks;
- detection of visible damage to power cables or switching devices.

PPE and CPE are used to protect against electric shock.

Means of collective protection:

1. Grounding of electric current sources;

2. Use of shields, barriers, cages, screens, as well as grounding and shunt rods, special signs and posters.

Individual protection means:

1. The use of dielectric gloves, insulating pliers and rods, metalwork tools with insulated handles, voltage gauges, galoshes, boots, stands and mats.

6.1.5 Fire hazard

According to the explosion and fire hazard, the premises are divided into categories A, B, C1-C4, D and D, and buildings into categories A, B, C, D and D.

According to FSR 105-03, the laboratory is classified as Category B. Characteristics: flammable and non-flammable liquids; solid flammable and non-flammable substances and materials; substances and materials, which can only burn by interaction with water, atmospheric oxygen or each other. If the premises in which it is situated are not in the most hazardous category A or B. According to the degree of fire resistance, this room belongs to the 1st degree of fire resistance according to SNiP 2.01.02-85 (made of brick, which belongs to slow-burning materials).

The occurrence of a fire when working with electronic equipment can be for reasons of both electrical and non-electrical nature.

Causes of a non-electrical fire:

a) negligent careless handling of fire (smoking, unattended heaters, use of open flames);

Causes of an electrical fire: short circuit, overcurrent, sparking and electric arcs, static electricity, etc. Primary fire-fighting equipment is used to contain or extinguish a fire in its initial stages. Primary fire-fighting equipment is usually used before the arrival of the fire brigade.

Water-foam fire extinguishers (OKHVP-10) are used to extinguish fires without the presence of electricity. Carbon dioxide (OU-2) and powder fire extinguishers are designed to extinguish electrical installations under voltage up to 1000V. To extinguish current-carrying parts and electrical installations, a portable powder fire extinguisher, such as OP-5, is used.

In public buildings and structures, at least two portable fire extinguishers must be placed on each floor. Fire extinguishers must be located in prominent places near room exits and not more than 1.35 m high. The placement of primary fire extinguishing equipment in corridors and passages should not interfere with the safe evacuation of people.

To prevent fire and explosion, it is necessary to provide:

* special isolated rooms for storage and spillage of flammable liquids (HFL), equipped

with blowing ventilation in explosion-proof version - according to GOST 12.4.021-75 and SNIP 2.04.05-86;

* special rooms (for storage in containers of dusty rosin) isolated from heating devices and heated parts of the equipment;

* primary fire-fighting equipment on the production floor (mobile carbon-dioxide fire-extinguishers GOST 9230-77, foam fire-extinguishers TU 22-4720-80, boxes of sand, felt, burlap or asbestos cloth);

* automatic alarms (type SVK-Z M 1) to indicate the presence of pre-explosive concentrations of flammable vapors of solvents and their mixtures in the air of the premises.

The laboratory fully complies with fire safety requirements, namely, the presence of a fire and security alarm, an evacuation plan shown in Figure 1, powder fire extinguishers with a certified stamp, signs indicating the direction to the emergency (evacuation) exit.

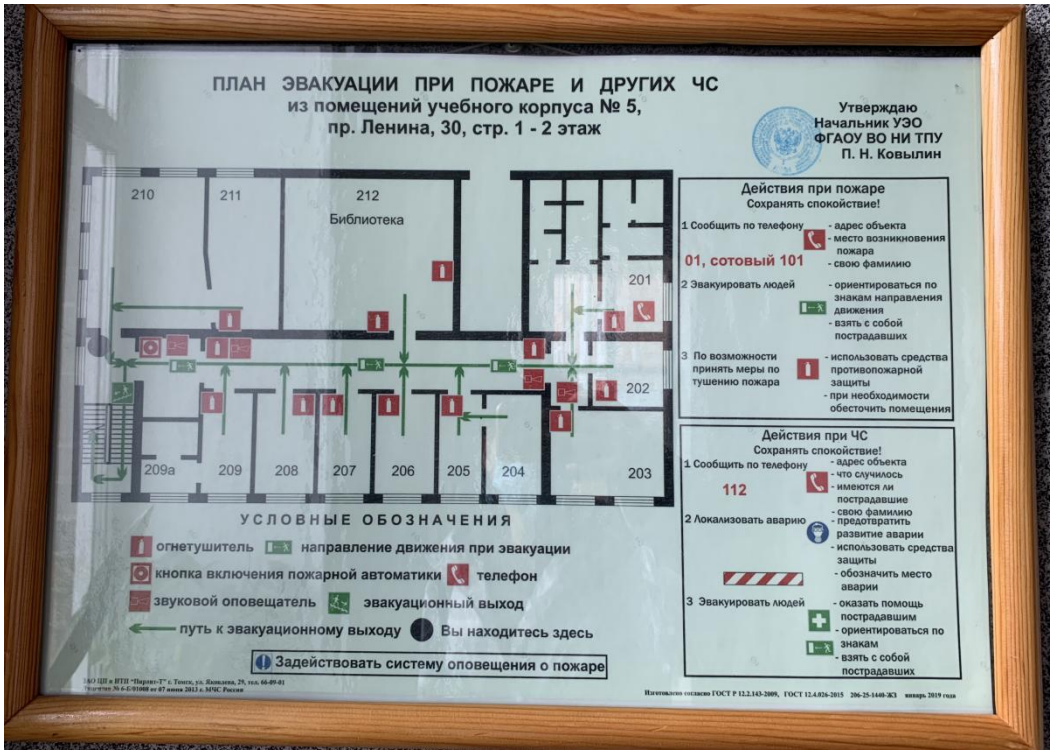


Figure 6.1 Evacuation plan

6.2. Environmental Safety

Computers have a huge number of components that contain toxic substances and pose a threat to both humans and the environment. These substances include:

- lead (accumulates in the body, affecting the kidneys, nervous system);

- mercury (affects the brain and nervous system);

- nickel and zinc (may cause dermatitis);

- alkalis (burn through the mucous membranes and skin);

Therefore, the computer requires special complex methods of disposal. This set of activities includes:

- separation of metal parts from non-metal;

- metal parts are melted down for subsequent production;

- non-metallic parts of the computer are specially processed;

Based on the above, before planning the purchase of a computer, you must:

- Take care in advance about how the existing equipment will be disposed of before buying a new one.

- Find out how new equipment complies with modern eco-standards and accept it for disposal after the end of its service life.

It is necessary to dispose of office equipment, and not just throw it in a "dump" for the following reasons:

Firstly, any computer and organizational equipment contains a certain amount of precious metals. Russian legislation provides for a clause according to which all organizations are required to keep records and movement of precious metals, including those that are part of fixed assets. For non-compliance with accounting rules, the organization may be fined in the amount of 20,000 to 30,000 rubles (according to Art. 19.14 of the Code of Administrative Offenses of the Russian Federation).

Secondly, the company can also be fined for unauthorised removal of machinery or equipment to a 'landfill'.

Recycling stage. By recycling machinery, we take care of the environment: the amount of non-recyclable waste is minimised and waste such as plastics, plastics, ferrous and non-ferrous metal scrap is recycled. Electronic boards containing precious metals are sent to a refinery after recycling, and the pure metals are then delivered to the State Fund instead of

ending up in landfills.

Thus, recycling a computer can be done as follows:

1. Use the services of a professional recycling company who can come and collect all the appliances that you plan to recycle.

2. You can contact your local municipality for electronics recycling.

6.3. Emergency Safety

Natural emergency - the situation in a certain territory or water area that has developed as a result of the occurrence of a source of a natural emergency that may cause or has caused human casualties, damage to human health and (or) the natural environment, significant material losses and violation of people's living conditions.

Production is located in the city of Tomsk with a continental cyclonic climate. Natural phenomena (earthquakes, floods, droughts, hurricanes, etc.) are absent in this city.

Possible emergencies at the facility in this case can be severe frosts and sabotage.

Frost is typical for Siberia in the winter season. Reaching critically low temperatures will lead to accidents in heating and life support systems, suspension of work, frostbite and even casualties among the population. In the event of pipe freezing, spare heaters should be provided. Their number and capacity should be enough to ensure that work in production does not stop.

In preparation for winter, you must perform the following activities:

- purchase and store a gasoline or diesel generator;
- purchase and store a gas heater with a cylinder in a warehouse;
- for the winter period, fulfill (1-3) daily supply of drinking and technical water;
- provide production with warm transport in case of an accident on municipal transport.

In the laboratory, the occurrence of emergency situations (ES) of a man-caused nature is most likely.

Technogenic emergencies are situations that arise as a result of industrial accidents and disasters at facilities, highways and product pipelines; fires, explosions at facilities.

To prevent the likelihood of sabotage, the enterprise must be equipped with a video

surveillance system, round-the-clock security, access control system, reliable communication system, as well as to exclude the dissemination of information about the security system of the facility, the location of premises and equipment in the premises, security systems, signaling devices, their installation locations and number. Officials once every six months conduct training to practice actions in case of emergency evacuation.

6.4. Legal and organizational issues of security:

6.4.1 Organizational security measures

To ensure safety, before starting work, it is necessary to prepare the necessary tools and fixtures for work, and prepare a desktop. It is also necessary to check the absence of external damage to electrical equipment, the presence and serviceability of control, measuring and signaling devices, a computer, toggle switches, switches, etc. Fault detection, it is not allowed to carry out repairs on your own, it is necessary to report to the head of the laboratory.

It is forbidden to start work if the personnel has not been trained and tested in accordance with the established procedure for labor safety knowledge.

Work with the machine must be carried out in a clean room, free from dust, fumes, acids and alkalis, corrosive gases and other harmful impurities that cause corrosion.

After finishing work with the installation, you must exit the program, turn off the power of the computer.

6.4.2 Legislative regulation of design solutions

The main task of regulating design decisions is resolved by complying with laws (tax laws, labor and civil codes). The head (responsible) assumes the obligation to implement and organize the evacuation rules and comply with the safety requirements in the premises.

A working room equipped with a computer and computer equipment should have the following parameters:

1. Protective ground.
2. Insulation, fencing and ensuring the inaccessibility of live parts.
3. Application of low voltage and double insulation.

The area per workplace for adult users must be at least 6 m^2, and the volume must

be at least 20 m³. For interior decoration of premises, diffusely reflective materials with a reflection coefficient of 0.7-0.8 (for the ceiling), 0.5-0.6 (for walls) and 0.3-0.5 (for the floor) should be used. The surface of the floor must be flat, without potholes, non-slip, easy to clean, clean, and have antistatic properties. Particular attention should be paid to fire safety, since fires in rooms with computer equipment are associated with a danger to human life and large material losses.

Working hours may not exceed 40 hours per week. Reduced working hours are possible. For workers under the age of 16, it is no more than 24 hours per week, and for those aged between 16 and 18, no more than 35 hours per week, the same as for persons with category I and II disabilities. In addition, working time depends on the working conditions: for employees working in jobs with hazardous living conditions, no more than 36 hours per week.

6.5 List of regulatory documentation

1. GOST 54 30013-83 Electromagnetic radiations from UHF. Maximum permissible levels of exposure. Safety requirements

2. GOST 12.4.154-85 "SSBT. Shielding devices for protection against electric fields of industrial frequency".

3. GN 2.2.5.1313-03 Maximum allowable concentrations (MAC) of harmful substances in the air of working area

4. SanPiN 2.2.4/2.1.8.055-96 "Electromagnetic radiations of radio frequency range (EMR RF)".

5. SanPiN 2.2.4.548-96. Hygiene requirements for microclimate of production premises.

6. SANPN 2.2.4/2.1.8.562-96. Noise at workplaces, in premises of residential and public buildings and in territory of residential buildings.

7. GOST 12.4.123-83. Means of collective protection against infra-red radiations. General technical requirements.

8. GOST R 12.1.019-2009. Electrical safety. General requirements and nomenclature

for types of protection.

9. GOST 12.1.030-81. Electrical safety. Protective earthing. Grounding.

10. GOST 12.1.004-91. Fire safety. General requirements.

GOST 12.2.037-78. Fire machinery. Safety requirements

11. SanPiN 2.1.6.1032-01. Hygiene requirements for ambient air quality

12. GOST 30775-2001 Resource Conservation. Waste management. Classification, identification and coding of wastes.

13. SNiP 21-01-97. Fire safety rules.

14. GOST 12.4.154. Occupational Safety Standards System. Shielding devices for protection against electric fields of industrial frequency. General technical requirements, main parameters and dimensions.

6.6 Illumination

According to SNiP 23-05-95 in a laboratory where periodic monitoring of the production process takes place with people constantly in the room, the illumination with a general lighting system should not be lower than 150 Lx.

Properly designed and executed lighting provides a high level of efficiency has a positive psychological effect on a person and contributes to an increase in labor productivity.

There should be no sharp shadows on the working surface, which create an uneven distribution of surfaces with different brightness in the field of view, distort the sizes and shapes of objects. As a result, fatigue increases and labor productivity decreases.

To protect against the blinding brightness of visible radiation (plasma torch in a chamber with a catalyst), goggles, shields, and helmets are used. Glasses should not restrict the field of vision, should be light, should not irritate the skin, fit well to the face and should not be covered with moisture.

The calculation of the general uniform artificial illumination of a horizontal work surface is carried out using the luminous flux coefficient method, which takes into account the luminous flux reflected from the ceiling and walls. The length of the room A = 7 m, width B = 6 m, height = 3.5 m. The height of the working surface above the floor hp = 1.0 m. category

of visual work.

Room area:

$$S = A \times B,$$

where A – length, m;

B – width, m.

$$S = 7 \times 6 = 42 \text{ m}^2$$

Reflection coefficient for freshly whitewashed walls with windows, without curtains C=50%, freshly whitewashed ceiling Ceiling P=70%. A safety factor that takes into account soiling of the luminaire, for spaces with low dust emission is KZ=1.5. Unevenness factor for fluorescent lamps Z= 1.1.

Reflection coefficient of freshly whitewashed walls with windows, without curtains ρ_C=50%, freshly whitewashed ceiling $\rho_П$=70%. The safety factor, taking into account the contamination of the luminaire, for rooms with low dust emission is SC = 1.5. Coefficient of unevenness for fluorescent lamps Z= 1.1.

We choose a fluorescent lamp LD-40, the luminous flux of which is equal to $Ф_{ЛД}$ = 2600 Lm.

We choose fixtures with fluorescent lamps of the ODOR-2-40 type. This luminaire has two lamps with a power of 40 W each, the length of the luminaire is 1227 mm, the width is 265 mm.

The integral criterion for the optimal location of luminaires is the value of λ, which for fluorescent luminaires with a protective grille lies in the range of 1.1–1.3. We accept λ=1.1, the distance of the fixtures from the ceiling (overhang) h_c = 0.3 m.

The height of the luminaire above the working surface is determined by the formula:

$h = h_n - h_p$,

where h_n –luminaire height above the floor, suspension height,

h_p – work surface height above the floor.

The smallest permissible suspension height above the floor for two-lamp luminaires ODOP: $h_n = 3,5$ m.

The height of the luminaire above the working surface is determined by the formula:

$$h = H - h_p - h_c = 3,5 - 1 - 0,5 = 2,0 \text{ м}.$$

The distance between adjacent lamps or rows is determined by the formula:

$$L = \lambda \cdot h = 1,1 \cdot 2 = 2,2 \text{ м}$$

The number of rows of lamps in the room:

$$Nb = \frac{B}{L} = \frac{6}{2,2} = 2,72 \approx 3$$

Number of fixtures in a row:

$$Na = \frac{A}{L} = \frac{7}{2,2} = 3,2 \approx 3$$

Total number of fixtures:

$$N = Na \cdot Nb = 3 \cdot 3 = 9$$

The distance from the extreme fixtures or rows to the wall is determined by the formula:

$$l = \frac{L}{3} = \frac{2,2}{3} = 0,7 \text{ м}$$

We place the lamps in three rows. The figure shows the plan of the room and the placement of luminaires with fluorescent lamps.

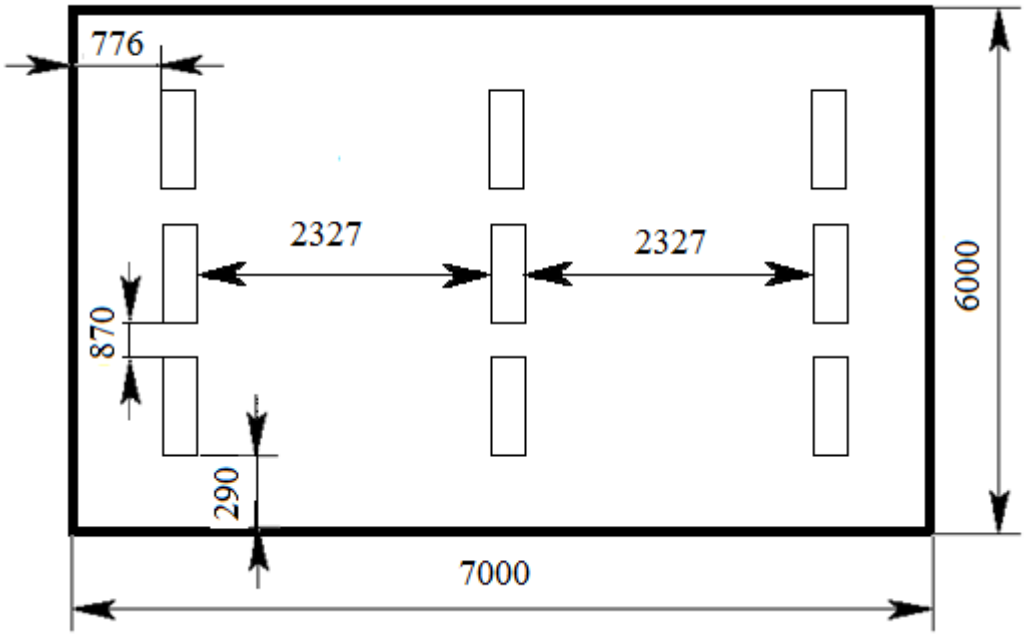


Figure 6.2 Plan of the premises and placement of luminaires with fluorescent lamps.

Achieving uniformity of illumination is achieved if the condition $L_1/3$ and $L_2/3$ is observed, i.e.

7000 = 2* L_1 + 2/3* L_1 + 3*265; 6205 = 8/3* L_1; L_1=2327; $L_1/3$ = 776;

6000 = 2* L_2 + 2/3* L_2 + 3*1227; 2319=8/3* L_2; L_2=870; $L_2/3$=290;

The room index is determined by the formula:

$$i = \frac{A \cdot B}{h \cdot (A + B)} = \frac{7 \cdot 6}{2{,}0 \cdot (7 + 6)} = 1{,}6$$

The coefficient of use of the luminous flux, showing what part of the luminous flux of the lamps falls on the working surface, for ODOR-type luminaires with fluorescent lamps at $\rho_П$ = 70%, $\rho_С$ = 50% and room index i = 1.6 is equal to η = 0.47.

The required luminous flux of a group of fluorescent lamps of a luminaire is determined by the formula:

$$\Phi_{\text{п}} = \frac{E \cdot A \cdot B \cdot К_З \cdot Z}{N \cdot \eta} = \frac{150 \cdot 7 \cdot 6 \cdot 1{,}5 \cdot 1{,}1}{9 \cdot 0{,}47} = 2457{,}44 \text{ лм}$$

Making a condition check:

$$-10\% \leq \frac{\Phi_{лд} - \Phi_{п}}{\Phi_{лд}} \cdot 100\% \leq 20\%;$$

$$\frac{\Phi_{лд} - \Phi_{п}}{\Phi_{лд}} \cdot 100\% = \frac{2600 - 2457{,}44}{2600} \cdot 100\% = 5{,}5\%.$$

Thus, we have obtained that the required luminous flux does not go beyond the required range. Now let's calculate the power of the lighting installation:

Poy = N * Podop

Poy = 9 * 40 = 360 W

6.7 Bibliography

1. SanPiN 2.2.2/2.4.1340-03. Sanitary and epidemiological rules and regulations 'Hygienic requirements for personal electronic computers and organisation of work'. Moscow: Goskomsanepidnadzor, 2003.

2. Safety Requirements SNB 4.02.01-03 "Heating, Ventilation and Air Conditioning".

3. SanPiN 2.2.4.548-96. "Hygienic requirements for microclimate of production premises".

4. SanPiN 2.2.4/2.1.8.562-96 "Noise at Work, in Residential and Public Buildings and in Residential Areas".

5. Basics of fire protection of enterprises GOST 12.1.004 and GOST 12.1.010 - 76

6. SanPiN 2.2.2.542-96 "Hygienic requirements for video and data display terminals, personal electronic computers and work organization".

Conclusion

The synthesis of high quality and emotional Dunxiang speech under resource-limited conditions has been implemented. The end-to-end autoregressive models Tacotron2 and Transformer were used to conduct speech synthesis experiments. It is found that the autoregressive model has poor generalization ability, low speech synthesis rate and loss of

synthesized speech in low resource condition. To improve the synthesis efficiency, the basic frequency information was extracted manually and synthesized using FastSpeech2. By comparing and analyzing the subjective and objective synthesis performance of all models, it was found that FastSpeech2 model can be used in resource-limited environments.

References

1. Zhang Bin, Quan Changqin, Ren Fuji. A Review of Speech Synthesis Methods and Development [J]. Small and Microcomputer Systems, 2016, 37(001): 186-192.
2. Wu Ping. A review of bilingual teaching research in the past five years [J]. China University Teaching, 2007, 000(001): 39-47.
3. Anonymous. Law of the People's Republic of China on the Standard Language and Characters of the People's Republic of China [J]. Publishing Science, 2001, 1(1): 2.
4. Xu Shixuan, Liao Qiaojing. A review of research on endangered languages [J]. Contemporary Linguistics, 2003, 5(2): 16.
5. Cao Zhiyun. The positioning, goals and tasks of the Chinese language resource protection project [J]. Language and Character Application, 2015(4): 8.
6. Yu Lei. A review of speech processing technology based on deep learning [J]. China Strategic Emerging Industries 2020, (14): 221.
7. Chen Zhigang, Hu Guoping, Wang Xifa. Text Standardization Method in Chinese Speech Synthesis System [J]. Journal of Chinese Information, 2003, 17(4): 46-52.
8. Wu Z, Cao G, Meng M, et al. A Unified Framework for Multilingual Text-to-Speech Synthesis with SSML Specification as Interface[J]. Tsinghua Science & Technology, 2009, 14(5): 623-630.
9. Moulines E, Charpentier F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones[J]. Speech Communication, 1990, 9(5-6): 453-467.
10. Valbret H, Moulines E, Tubach J P. Voice transformation using PSOLA technique[J]. Speech Communication, 1992, 11(2-3): 175-187
11. Zen H, Tokuda K, Black A W. Statistical parametric speech synthesis [J]. Speech Communication, 2009, 51(11): 1039-1064.
12. Tokuda K, Zen H, Black A W. An HMM-based speech synthesis system applied to English [C] // Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002: 227-230.
13. Yamagishi J, Tamura M, Masuko T, et al. A Training Method of Average Voice Model for

HMM-Based Speech Synthesis[J]. Ieice Transactions on Fundamentals of Electronics Communications & Computer Sciences, 2003, 86(8): 1956-1963.

14. Wang Y, Skerry-Ryan R J, Stanton D, et al. Tacotron: Towards end-to-end speech synthesis[C] // Procedings of Interspeech 2017. 2017, 4006-4010.
15. Shen J, Pang R, Weiss R J, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]//2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018: 4779 -4783.
16. Kim J, Kim S, Kong J, et al. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search [J] // NeurIPS 2020[C]. 2020, 33: 8067-8077.
17. Hu Qiong. Tianjin Dialect Speech Synthesis Based on Hidden Markov Model [D]. Shanghai: Shanghai Jiaotong University, 2012.
18. Bellman, R. Dynamic Programming[J]. Science, 1966, 153(3731):34-37.
19. Qi Fangkun. Research on phonetic generation in Dongxiang dialect [D]. Lanzhou: Northwest Normal University, 2018.
20. Schuster M, Paliwal K K. Bidirectional recurrent neural networks [J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.
21. Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-1780.
22. Chorowski J, Bahdanau D, Serdyuk D, et al. Attention-Based Models for Speech Recognition[J]. Computer ence, 2015, 10(4): 429-439.
23. Su J, Jin Z, Finkelstein A. HiFi-GAN: High fidelity denoising and dereverberation based on speech deep features in adversarial networks [J]. arXiv preprint arXiv:2006.05694, 2020.
24. Van Den Oord A, Dieleman S, Zen H, et al. WaveNet: A generative model for raw audio [J]. SSW, 2016, 125: 2.
25. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.
26. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and

translate[J]. arXiv preprint arXiv:1409.0473, 2014.
27. Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[C]//International Conference on Machine Learning. PMLR, 2017: 1243-1252.
28. Gu J, Bradbury J, Xiong C, et al. Non-autoregressive neural machine translation [J]. arXiv preprint arXiv:1711.02281, 2017.
29. Guo J , Tan X , He D , et al. Non-Autoregressive Neural Machine Translation with Enhanced Decoder Input[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33:3723-3730.
30. Ren Y, Ruan Y, Tan X, et al. Fastspeech: Fast, robust and controllable text to speech[J]. Advances in Neural Information Processing Systems, 2019, 32.
31. Ren Y, Hu C, Tan X, et al. Fastspeech 2: Fast and high-quality end-to-end text to speech[J]. arXiv preprint arXiv:2006.04558, 2020.
32. Gibiansky A, Arik S, Diamos G, et al. Deep voice 2: Multi-speaker neural text-to-speech [J]. Advances in neural information processing systems, 2017, 30.
33. Hirose K, Tao J. Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis[J]. Prosody Phonology & Phonetics, 2015.
34. Liu Zhaoxiong. A Brief History of Dongxiang Dialects [M]. Ethnic Publishing House, 1981.
35. Khuselt. A database-based study of vowel acoustics in Dongxiang dialect [D]. Lanzhou: Northwest University for Nationalities, 2015.
36. Ma Guoliang, Liu Zhaoxiong. Research on Dongxiang Language [J]. Northwest Ethnic Studies, 1986: 167-184.
37. Qu Shiwei. Research on Dongxiang Phonetics and Acoustics [D]. Lanzhou: Northwest University for Nationalities, 2012.
38. Qiqigma. Research on Chinese loan words in Dongxiang dialect [D]. Hohhot: Inner Mongolia University, 2017.
39. Ma Yan. A Comparative Study of Word Formation in Chinese and Dongxiang [D]. Yili: Yili Teachers College, 2013.

40. Bao Saren. Research on the Contact between Dongxiang Language and Chinese in Mongolian Language Group [D]. Beijing: Peking University, 2006.
41. Zhang Jiayuan. The phonetic transcription of Mandarin Chinese phonetics SAMPA-SC[J]. Chinese Journal of Acoustics. 2009, 34(1): 81-86.
42. Wali A, Alamgir Z, Karim S, et al. Generative adversarial networks for speech processing: A review - ScienceDirect[J]. 2022, 72: 101308.
43. Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks[J]. Advances in Neural Information Processing Systems, 2014, 3: 2672-2680.
44. Pascual S, Bonafonte A, Serra J. SEGAN: Speech enhancement generative adversarial network[J]. arXiv preprint arXiv:1703.09452, 2017.
45. Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125-1134.
46. Mao X, Li Q, Xie H, et al. Least squares generative adversarial networks [C] // Proceedings of the IEEE international conference on computer vision. 2017: 2794-2802.
47. He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification [C] // Proceedings of the IEEE international conference on computer vision. 2015: 1026-1034.
48. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
49. Recommendation ITUT. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs[J]. Rec. ITU-T P. 862, 2001.
50. Kingma D, Ba J. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.
51. Boersma P. Praat, a system for doing phonetics by computer[J]. Glot. Int., 2001, 5(9): 341-345.
52. Sun Zhi. PHS Speech Quality Assessment (MOS) Analysis [J]. The 2nd Wuhan City Academic Annual Conference, Proceedings of the 2006 Academic Annual Conference of

the Society of Communications, 2006.
53. Murphy A H. General decompositions of MSE-based skill scores: Measures of some basic aspects of forecast quality [J]. Monthly Weather Review, 1996, 124(10): 2353-2369.