Dongxiang speech synthesis based on statistical parameter method

Man Wang[†], Fangkun Qi[†], Hongwu Yang[*†#] and Jingwen Sun[†]
[*]School of Educational Technology, Northwest Normal University, Lanzhou, China
[†]College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, China
[#]National and Provincial Joint Engineering Laboratory of Learning Analysis Technology in Online Education, Lanzhou, China
E-mail: yanghw@nwnu.edu.cn Tel: +86-18193125028

Abstract—**Dongxiang language is a kind of text-free Chinese dialect. Therefore, it is difficult to employ traditional text-to-speech(TTS) technology to realize a Dongxiang dialect TTS system. The paper realized a Dongxiang dialect speech synthesis using Hidden Markov Model-based statistical parametric (HMM), Deep Neural Network (DNN)-based method and speaker adaptive training method by analyzing the linguistic features and acoustic characteristics of Dongxiang language. The experimental results show that, in the case of a certain corpus, the DNN-based speaker adaptive speech synthesis method can achieve better performance than the other two methods and can synthesize more natural speech.**

Keywords: **Dongxiang speech synthesis; Text-free Dialect speech synthesis; Hidden Markov model; Deep Neural Network model; speaker adaption**

I. Introduction

In recent years, most researches have focused on multiple languages speech synthesis[1-2]. However, the languages for speech synthesis research are based on the premise of textual representation. The speech synthesis in the current stage mainly refers to text-to-speech (TTS), which converts the information expressed by text into the form of speech, and has a wide range of uses in the fields of communication and smart home. The main methods of speech synthesis can be divided into formant synthesis, waveform mosaic synthesis and statistical parameter speech synthesis.

Early formant synthesis[3] is difficult to obtain high naturalness synthesized speech, but in the practical application of speech synthesis, people want to synthesize more natural and fluent speech. Later, researchers have proposed waveform stitching technology[4], which needs to be store a large number of synthetic speech primitives first, and it also requires a lot of manpower and resources to build a large-scale speech database, and the synthesized speech sound quality and naturalness are greatly affected by the environment, the synthesized speech is relatively unitary. A method of speech synthesis based on statistical parameters[5] was proposed. Among them, the synthetic method widely used in the early stage is based on HMM statistical parameter speech synthesis[6] method, the advantage is that the required corpus is small, and no manual intervention is required to automatically build the system. The whole process is completed by the program automation, and the synthesized sound quality smooth, with high robustness. However, the speech synthesized by this method is not high in definition and the emotional rhythm is not rich enough. It seems too dull in listening. In the later period, with the continuous improvement of computer performance, the method of deep learning[7] has made significant progress in speech synthesis. The advantage of deep learning lies in the independent learning with supervised or unsupervised features. The hierarchical feature extraction algorithm is more advantageous than the manual feature extraction, and the extracted features have stronger generalization performance than the artificially set features. it can learn shallow semantic information, and can also learn deep semantic information, so that the features can be automatically acquired and the synthesized speech has a great improvement in naturalness and emotional rhythm. At the same time, deep learning also has shortcomings. Because of the large amount of network parameters, the training speed is very slow and storage is very inconvenient.

This paper takes the Dongxiang dialect as the research object. Dongxiang language has no textual expression to make it face the challenge in the speech synthesis system compared with most other languages. We use HMM-based statistical parameter speech synthesis method, DNN-based speech synthesis method and DNN-based speaker adaptive method to realize Dongxiang language speech synthesis. It is of great significance to protect the language without written expressions.

II. Construction of Dongxiang language corpus

Dongxiang language is the language used by Dongxiang people. It belongs to the Mongolian language family of the Altaic language. Compared with the language of the same family, with the change of history, Dongxiang language has merged Persian, Turkic, Arabic, Mongolian and Chinese[8-9]. So there are many Chinese loanwords in Dongxiang language vocabulary, as well as many Turkic loanwords, Arabic and Persian loanwords. Dongxiang language has the characteristics of integration and colloquialism. With the development of the times, social progress, national integration, the modern Dongxiang dialect has absorbed many Chinese vocabulary reflecting new things in the new era. Although there is no written expression in Dongxiang dialect, it has the same linguistic and acoustic characteristics as other languages. Analysis of the characteristics of the Dongxiang dialect is the basis of the design of the Dongxiang language corpus and the scheme of mechanical pronunciation.

Phoneme Characteristics of Dongxiang Language

Vowel characteristics of Dongxiang dialect: Based on the viewpoints of scholars Ma Guoliang and Liu Zhaoxiong, this paper holds that there are seven vowels in Dongxiang dialect **i, ə, a, o, u, ɯ, ɚ**[9]. The vowels of Dongxiang dialect have the following characteristics:

(1) Vowels ə Read [e] at the back of “A Consonant + [i]” . For example: **məiliə** should be [məilie], meaning "front".

(2) Vowels ɯ are combined with the two consonants will undergo sound changes. In one case, they are pronounced [ɿ] after consonants **dz, ts** and **s**. For example: **ʦɯdao** reads [ʦ ʻ ɿ tao], meaning "bayonet"; In another case, it is pronounced [ʅ] after consonants **dʐ, tʂ, ʂ, ʐ** and without consonants n. For example: **gəidʐɯ** read as [kəitʂʅ] means “ring” , if there is a consonant **n**, it is still pronounced as [ɯ] .

(3) Vowels **i** and **u** are in words containing two or more syllables. When the first consonant is aspirated stopper, affricate or clear affricate, the **i** and **u** pronunciation will be cleared up. For example: **fudu** reading [fu̥tu] means "long"; **ɕidu** reading [ɕi̥tu] means "sharp".

(4) In closed syllables ending in **n**, the main vowels have nasalized sounds. For example: **xon** reads [xuɐ̈́ ŋ] meaning "year". Dongxiang dialect has eleven vowels of **ia, iə, iao, iu, ai, əi, ao, ou, ua, uai, ui**. These vowels are included in the design of the corpus.

Dongxiang language consonant features:There are 28 consonants in Dongxiang language **b, p, m, f, d, t, n, l, r, ʣ, ʦ, s, ʥ, ʨ, ɕ, dʐ, tʂ, ʂ, ʐ, ɡ, k, ɣ, ɢ, q, h, j, w**[9], the consonants of Dongxiang dialect have the following characteristics:

(1) When the consonant n appears at the end of the syllable, the beginning n is the tip nasal sound, such as: **indʑi-** pronounced [ĩntɕi], meaning "printing"; if the syllable is the following syllable or the last syllable, n is the root of the tongue nasal sound, such as: **antan** read [ãŋtãŋ] means "gold".

(2) The consonant **ɣ** only appears in the middle of the inherent word, the consonant **X** only appears in the first part of the inherent word, and the consonant x in the Chinese loan word can appear in the middle of the word, for example, **miənxua** means “cotton” ; **dʐixua** means “ plan".

Dongxiang dialect stress characteristics: The stress of Dongxiang dialect is very obvious when it is pronounced. Whether it is a disyllabic word or a trisyllabic word, the stress is usually increased on the last syllable of the word. For example:

In disyllabic words, **a'na** means "mother"; **u'su** means "water"; **na'ran** means "sun" and so on.

The ana’ ni in the trisyllabic word means "mother."

Dongxiang dialect also uses stress position to distinguish the meaning of words, such as:

ʂən’dzɯ, with the stress moving forward, [’ʂəndzɯ] means “fan”.

bao’dzɯ, with the stress moving forward, [’baodzɯ] means “leopard”.

Phonetic Characteristics of Dongxiang Language

Dongxiang dialect does not have its own expressive characters. Therefore, studying the acoustic characteristics of the Dongxiang dialect and summarizing the acoustic characteristics of the Dongxiang dialect plays an important role in the subsequent establishment of the Dongxiang dialect corpus and the training of the Dongxiang dialect's phonetic primitive model. According to the above, the syllables of Dongxiang dialect are different from the Chinese syllables. The Chinese syllables are composed of vowels, while the syllables of Dongxiang dialects are composed of vowels and consonants. Based on the vocabulary appendix after the book “Brief Records of Dongxiang Dialect”, this paper selects the monosyllabic words and disyllabic words of Dongxiang dialect, and extracts its fundamental frequency for five-degree analysis. The fundamental frequency is extracted by Praat software, and each word is extracted from about 70 to 100 fundamental frequency points. Each fundamental frequency value is calculated according to the five-degree toner model, and the five-degree curve of the word, that is, the fundamental frequency curve can be obtained. The corresponding values between 0-1 and 1,1-2 correspond to 2,2-3 and 3,3-4 and 4,4-5 correspond to 5.

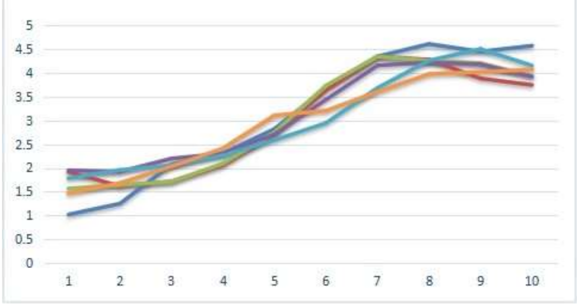

Fig. 1 Five-degree Curve of Monosyllabic Words in Dongxiang Language.

The five-degree curve of the monosyllabic words in Dongxiang dialect is shown in Figure 1. From the graph, we can see that the starting value of five-degree of monosyllabic words in Dongxiang dialect is between 1 and 2, corresponding to 2 of the five-degree value; the final value is mostly between 5 and 4, corresponding to 5 of the five-degree value. Considering only the authentic Dongxiang dialect and excluding borrowings from other languages, the fifth tone of monosyllabic words in Dongxiang dialect is 25, which is roughly Yangping tone.

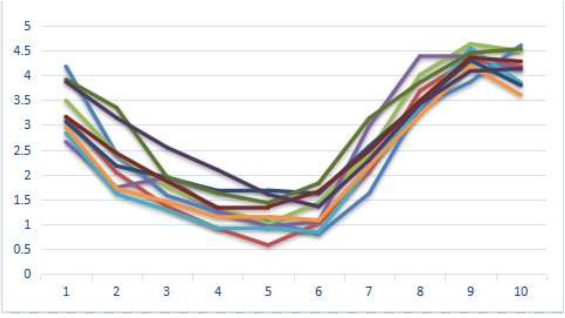

Fig. 2 Five-degree Curve of Disyllabic Words in Dongxiang Language.

Five-degree curves of disyllabic words in Dongxiang dialect are shown in Fig. 2. From the figures, we can see that the starting values of the five-degree values of disyllabic words are mostly between 2.5 and 3.5; the median values are mostly between 0.5 and 1.2; the maximum values are close to the end of pronunciation, between 4 and 5, and the final values are mostly between 3.5 and 4.5. Considering only the authentic Dongxiang dialect and excluding borrowings from other languages, the fifth tone of the disyllabic words in Dongxiang dialect is roughly a combination of upper tone and Yinping tone.

With the development of the times and social progress, Dongxiang dialect has absorbed a lot of new Chinese vocabulary, which is called Chinese loanword. The characteristics of Chinese loanwords in Dongxiang dialect are that they have the same pronunciation as Putonghua, but there are changes in tone. With the help of Dongxiang students, the following tone changes of Chinese loanwords are summarized. It is worth noting that these tone changes appear in the last two words:

(1) Yinping + Yangping →Shangsheng + Yangping

(2) Yinping + Qusheng →Shangsheng + Yinping

(3) Yangping + Qusheng →Shangsheng + Yinping

(4) Shangsheng + Shangsheng →Shangsheng+ Yinping

(5) Shangsheng + Qusheng →Shangsheng+ Yinping

(6) Qusheng + Shangsheng →Shangsheng + Yang Ping

Although Dongxiang dialect is a syllable composed of vowels and consonants, the spelling of Dongxiang dialect is almost the same as that of Chinese, which is a consonant and vowel structure. Below are four typical syllable structures of Dongxiang dialect, namely:y

(1) Only vowels (including diphthong) can form their own syllables. For example: **a-na** means “mother” and **ui-ra** means “near”.

(2) The monosyllable plus N also forms a syllable. For example: **un-du** means “high”, **əngiə** means “clothing”

(3) Consonant + unit sound, consonant + diphthong also constitutes syllables. For example: **bu** means "no", **qa-** means "closed", and **kao** means "bitter."

(4) Consonant plus vowels include complex vowels plus n. For example: **xon** means "year" and **kiən** means "who".

(5) If the beginning syllable of the word has no vowel, the consonant between the two vowels belongs to the later syllable when spelling.

Dongxiang Language Corpus

Through the above research and analysis of Dongxiang language, we can conclude that there are 7 vowels and 28 consonants in Dongxiang language. There are four kinds of tones: Yinping, Yangping, Shangsheng and Qusheng. The tuning classes are the same, but the tuning values are different.

In terms of vocabulary, the vocabulary of Dongxiang dialect consists of two parts: the intrinsic word and loanwords. The inherent words in Dongxiang dialect mainly refer to the words that are homologous to the relative languages. We define them as authentic Dongxiang dialect in the following corpus design.

According to the difference between the syllables of Dongxiang dialect and those of Chinese, we choose all vowels and consonants of Dongxiang dialect as the basic unit of synthesis[10-12], including silence and pause. A total of 2400 sentences of Dongxiang dialect corpus is designed. Considering the characteristics of a large number of Chinese loanwords in modern Dongxiang dialect, a set of bus stop corpus is designed. Give an example sentence:hello passengers, welcome to the 131 bus, next stop at Lanzhou station passengers getting off the bus, please go to the back door. It can not only be applied to real life, but also cover all phonemes, tone changes, vowels and consonants of Dongxiang dialect. We screened six students of Dongxiang nationality, each of whom recorded 400 Dongxiang dialect. Speech signal sampling rate is 16KHz. Finally, from the recorded 2400 Dongxiang corpus, 100 sentences are randomly selected as the test sentences 2300 sentences as the training set. Acoustic models are built based on HMM, DNN model and DNN adaptive algorithm, respectively.

Dongxiang Corpus Processing and Labeling

At this stage of speech synthesis, speech is synthesized by inputting text, that is, text to speech (TTS).Because Dongxiang dialect has no characters, this paper designs a set of machine-readable phonetic symbol scheme SAMPA-DX for Dongxiang dialect to mark the Dongxiang dialect corpus, which is based on SAMPA of Chinese.

Firstly, the international phonetic symbols of Dongxiang dialect and Mandarin Chinese are compared. If they are the same, they will be marked with the Chinese phonetic alphabet corresponding to the international phonetic symbols of Mandarin. If they are different, take the international phonetic symbols of Dongxiang dialect as reference, and mark the vowels and consonants of Dongxiang dialect with custom symbols. Since Dongxiang dialect itself has no written language,We need to label the Dongxiang dialect corpus represented by Chinese characters with SAMPA-DX, the machine-readable phonetic symbol of Dongxiang dialect. Because the pronunciation of Dongxiang dialect is almost the same as that of Chinese phonetic alphabet, in order to facilitate the subsequent tagging of corpus, we use capital letters or lower-case letters plus h to express the international phonetic alphabet of Dongxiang dialect, which is different from Putonghua international phonetic alphabet.

The basic unit of synthesis we selected is all vowels and consonants in Dongxiang dialect, including silence and pause. This paper designs a set of four-level context-related annotations for Dongxiang dialect, which are vowel-consonant level, syllable level, word level and sentence level, to obtain the context-related features of synthetic basic units.

The content of contextual information needs to be represented by symbols. Contextual information annotation in this paper is based on context-related annotation formats. Context includes the pronunciation information of the current syllable, the pronunciation information of the contextual environment, the duration information and the prosodic information. Through the design of context annotation format, annotation files are generated by program analysis. According to the annotation format of the above design, through program analysis, the mono-phone annotation files and context-related annotation files of Dongxiang dialect are generated. Mono-phone annotation contains all phoneme information, and context-related annotation files contain context information.

In the HMM-based statistical parameter speech synthesis system, in order to improve the accuracy of the model, a decision tree algorithm is needed to establish a context-dependent hidden Markov model. Therefore, this paper designs a set of statistical parameter-oriented speech synthesis problems for Dongxiang dialect. The question set covers all the features of context-related annotations, including more than 3000 context-related issues with Dongxiang vowels and consonants as the basic unit of synthesis.

III. DONGXIANG SPEECH SYNTHESIS FRAMEWORK

This paper mainly studies the speech synthesis of Dongxiang dialect. Based on the statistical parameters of the hidden Markov model, the speech synthesis method based on the deep neural network [13] and the speaker adaptive method based on DNN, the speech synthesis of the Dongxiang dialect is realized and the experimental results are obtained. Conduct analysis and evaluation.

A. HMM-based statistical parameter Dongxiang speech synthesis framework

The HMM-based statistical parameters Dongxiang speech synthesis mainly includes two stages of training and synthesis. The basic block diagram of the synthesis system is shown in Figure 3

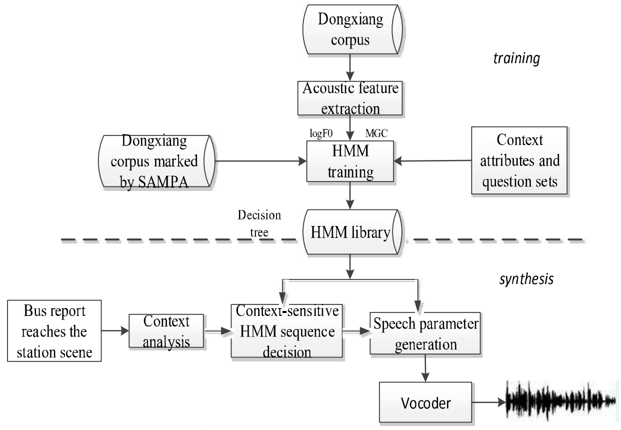
Dongxiang corpus
training
Acoustic feature extraction
logF0
MGC
Dongxiang corpus marked by SAMPA
HMM training
Context attributes and question sets
Decision tree
HMM library
synthesis
Bus report reaches the station scene
Context analysis
Context-sensitive HMM sequence decision
Speech parameter generation
Vocoder

Fig. 3 Framework of Dongxiang dialect speech synthesis base on HMM.

The training phase includes information preprocessing and HMM training. In the pre-processing part, the Dongxiang language corpus is first given, and then the acoustic parameters that need to be modeled are configured. In the speech synthesis based on HMM statistical parameters, the generalized Mel-generalized Cepstral (MGC) and log-fundamental frequency (LogF0) in speech are extracted as characteristic parameters. We use a single phoneme of Dongxiang dialect as a synthetic primitive. In the HMM training part, the single phoneme HMM training is carried out first. Under the guidance of the maximum likelihood criterion, the model parameters are re-evaluated by the EM [14] algorithm (Expectation Maximization Algorithm, EM). After training the Dongxiang dialect monophone model, the monophone model needs to be extended by the context attribute set and then re-evaluated. The decision tree is used to cluster the context-related models to obtain more abundant and balanced data. After the clustering is completed, the clustered model needs to be re-evaluated again to obtain the final Dongxiang dialect HMM library.

In the speech synthesis stage, this paper designs a Dongxiang dialect bus station station scene statement. Since the Dongxiang dialect has no text, the SAMPA-DX standard scheme is used to mark the corpus. After the context analysis program, the context-sensitive labeling of each pronunciation primitive is obtained. When the bus station name and the bus serial number are input, the context analysis program is used to generate the contextual information annotation sequence of the public transportation station; then the context-dependent HMM sequence of each pronunciation primitive is obtained by the decision tree, and the corresponding speech parameters are generated, and finally used. The STRAIGHT vocoder synthesizes Dongxiang speech.

B. Dongxiang Speech Synthesis Based on DNN

The DNN-based Dongxiang dialect speech synthesis framework is shown in Figure 4. The DNN experiment is divided into several DNN training phases and speech synthesis phases.

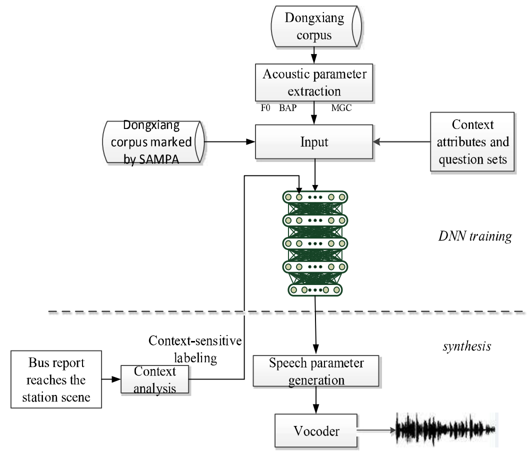
Dongxiang corpus
Acoustic parameter extraction
F0 BAP MGC
Dongxiang corpus marked by SAMPA
Input
Context attributes and question sets
DNN training
synthesis
Context-sensitive labeling
Bus report reaches the station scene
Context analysis
Speech parameter generation
Vocoder

Fig. 4 Framework of Dongxiang dialect speech synthesis base on DNN.

In the training phase, the Dongxiang speaker corpus is first given. In the DNN-based speech synthesis, the Dongxiang dialect is subjected to the acoustic parameter [15] extraction process to obtain the fundamental frequency (F0) and the generalized Mel cepstrum system required for the training model. (Mel-generalized Cepstral, MGC), band aperiodicity (BAP) acoustic parameters. The Dongxiang dialect corpus marked with SAMPA-DX is a composite primitive with a single phoneme. With the guidance of the dictionary and the grammar rule base, the monosyllabic labeling and context-dependent are obtained through grammar analysis, prosody analysis, and word-to-speech conversion. The context-sensitive annotation, single phoneme annotation and acoustic parameters are input as input to the DNN for DNN training to obtain a DNN acoustic model of Dongxiang language.

In the speech synthesis stage, enter the Dongxiang dialect bus station station scene statement, after context analysis, get the context-related annotation of each pronunciation primitive, and then send it to DNN training to get the acoustic parameters such as the fundamental frequency and spectral parameters required for training. Finally, the speech is synthesized by the WORLD vocoder by using the generated speech parameters.

C. Speaker-adaptive speech synthesis based on DNN

The Dongxiang speech synthesis framework based on DNN speaker adaptation[16-17] is shown in Figure 5. The experiment is divided into data preparation stage, training average sound model stage, speaker adaptation stage and speech synthesis stage.

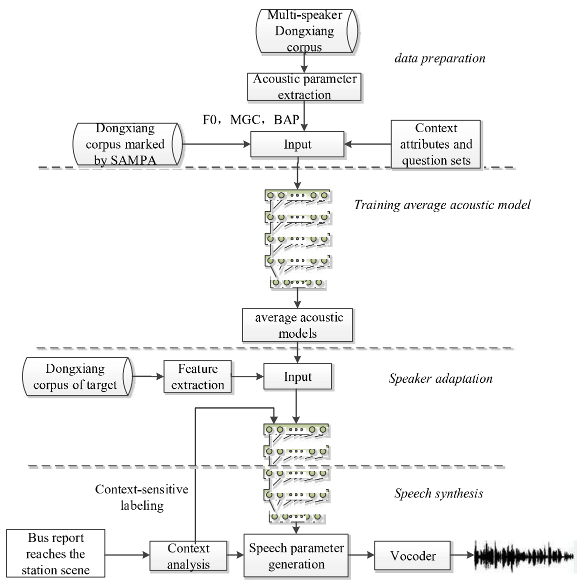
Multi-speaker Dongxiang corpus
data preparation
Acoustic parameter extraction
F0，MGC，BAP
Dongxiang corpus marked by SAMPA
Input
Context attributes and question sets
Training average acoustic model
average acoustic models
Dongxiang corpus of target
Feature extraction
Input
Speaker adaptation
Context-sensitive labeling
Speech synthesis
Bus report reaches the station scene
Context analysis
Speech parameter generation
Vocoder

Fig. 5 Framework of Dongxiang dialect speech synthesis base on DNN adaptive.

In the data preparation stage, the Dongxiang phonetic corpus of the five speakers is subjected to the acoustic parameter extraction process to obtain the acoustic parameters such as the fundamental frequency and spectral parameters required by the training model. The Dongxiang dialect corpus marked with SAMPA-DX is subjected to grammatical analysis, prosody analysis, word-to-speech conversion, etc., to obtain context-sensitive annotations containing phoneme information and context information.

In the training average sound model phase, the acoustic parameters of the five speakers and context-sensitive annotations are used as inputs to train the DNN model. During training, the DNN architecture shares a hidden layer between different speakers to model the language parameters. The acoustic feature is modeled by the stochastic gradient descent process of the backpropagation algorithm, and finally the acoustic average model of the multi-speaker is obtained.

In the adaptive training phase, the target speaker's corpus is given. The target speech is extracted by acoustic characteristic parameters to obtain acoustic parameters (F0, MGC, BAP). Firstly, the average speaker duration and acoustic model of the average sound model training stage are put into the target speaker DNN model for adaptive duration and acoustic feature training. Under the guidance of the adaptive model, the adaptive time length and acoustic model are trained and the corresponding speech parameters are generated. In the synthesis stage, the speech text to be synthesized is first input, and the context-related annotation including the phoneme and the context-related information is obtained through the text analysis process, and then the target speech acoustic parameters (F0, MGC, BAP) generated by the adaptive training are used, and the WORLD vocoder is adopted. Finally, the speech of the target speaker is synthesized.

IV. Dongxiang Language Speech Synthesis Experiment Evaluation

In this paper, the speech synthesis experiments of Dongxiang dialect are carried out by using three different models of HMM, DNN and DNN adaptive model. Using 100 randomly selected Dongxiang dialects as test sets, 2300 sentences are collected for training and the experiment is performed. The results were compared and evaluated.

A. *Subjective Evaluation*

Subjective evaluation is divided into Mean Opinion Score (MOS) and DMOS (Degradation Mean Opinion Score, DMOS). The MOS evaluation scores the speech quality such as the naturalness, fluency and sound quality of the synthesized Dongxiang dialect. Select 10 reviewers, 5 of whom are Dongxiang students, and 5 are not Dongxiang students; then the original Dongxiang dialect and the synthesized Dongxiang dialect are grouped together and grouped into 10 reviewers. The synthesized Dongxiang dialect is played, and the Dongxiang dialect synthesized in each group is played twice. Finally, the reviewer scores the synthesized Dongxiang dialect voice quality according to the MOS evaluation standard of [18], and takes the average score of each group.

Figure 6 shows the MOS evaluation results of the Dongxiang dialect synthesized by the Dongxiang family reviewer and the non-Dongxiang

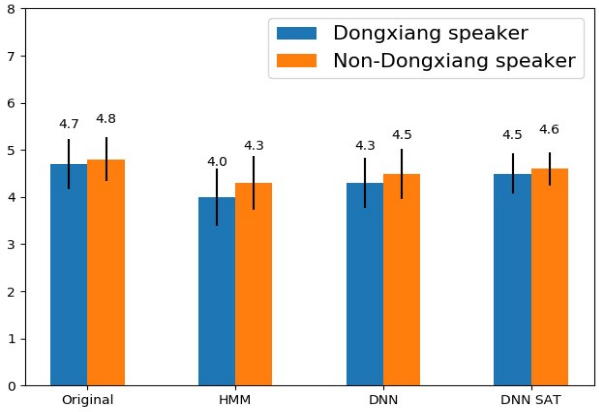

Fig. 6 MOS Evaluation of Dongxiang Dialect Under 95% Confidence Interval.

The results show that when the corpus data is 2400, the Dongxiang dialect synthesized by DNN is better than the DNN model and the speech quality is better than HMM, and the score is slightly higher; while the non-Dongxiang people who do not know Dongxiang language think that the synthesized speech quality is better, while the Dongxiang people think that the synthesized speech has some murmurs, and some words sound is not very clear.

The DMOS evaluation is used to evaluate the similarity between the synthesized Dongxiang dialect and the original Dongxiang dialect. As with the MOS evaluation, 10 reviewers were selected, 5 were Dongxiang nationality college students, and 5 were non-Dongxiang nationality college students; then we recorded the original Dongxiang dialect and the synthesized Dongxiang dialect as a group, and 10 evaluations. The people play the synthesized Dongxiang dialect group by group, and the Dongxiang dialect synthesized in each group plays twice; the reviewer is required to carefully compare the phonetic similarity between the original voice and the synthesized Dongxiang voice, and finally evaluate the DMOS evaluation standard according to [18]. Score, take the average score for each group.

Figure 6 below shows the DMOS evaluation results of the Dongxiang dialect and the non-Dongxiang tester on the synthesis of Dongxiang dialect

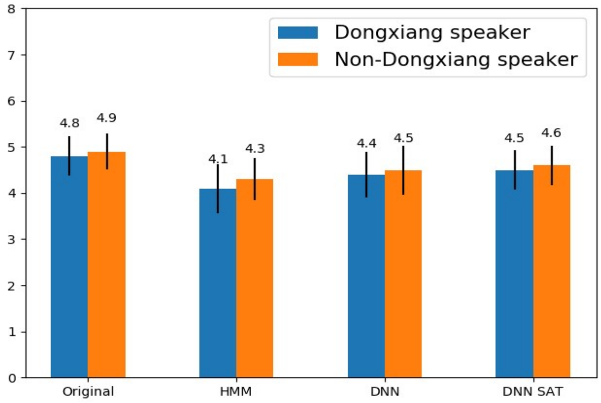

Fig. 7 DMOS Evaluation of Dongxiang Dialect Under 95% Confidence Interval.

It can be seen from the graph that the Dongxiang dialect synthesized by DNN model is obviously better than HMM, and the Dongxiang dialect synthesized by DNN self-adaptively is slightly better than the DNN model under the corpus data of 2400 sentences. The speech similarity between DNN adaptive model and DNN model is higher,and the score is slightly higher; the non-Dongxiang people who do not understand Dongxiang language think that the synthesized speech similarity is better, while the Dongxiang people think that the synthesized speech is slightly different, and some words have slightly changed in tone with the original speech.

B. *Objective Evaluation*

The objective evaluation mainly uses the F0 and dur distortion in the calculation of the root mean square error (RMSE), and calculates the quality of the synthesized speech by the Spectral Centroid (SC) in the DNN experiment. The results of subjective evaluation may be biased because of the different perceptions of the human body, and the objective evaluation may not be affected by the will of the reviewer. Firstly, the original and generated phonetic parameters such as the fundamental frequency, duration and spectral centroid of Dongxiang dialect are extracted, and the root mean square error analysis is performed. Table 1 is a comparison of root mean square errors of fundamental frequency F0/Hz, Dur/s and spectral centroid SC/Hz of HMM, DNN and DNN adaptive algorithms when the corpus is 2400 sentences, respectively.

Tab.1 objective evaluation of Dongxiang dialect

Training model	F0/Hz	Dur/s	SC/Hz
HMM	8.74	0.118	61.08
DNN	4.81	0.112	33.27
DNN Speaker Adaptation	3.72	0.087	21.18

V. Conclusions

Taking Dongxiang dialect as the research object, this paper completes the speech synthesis of Dongxiang dialect using HMM-based statistical parameter speech synthesis method, deep neural network-based speech synthesis method and DNN-based speaker adaptation speech synthesis method. The evaluation data show that in the case of 2400 sentence corpus, Dongxiang dialect synthesized by DNN speaker adaptive algorithm is better than Dongxiang dialect synthesized by deep neural network. There is a big gap between Dongxiang dialect phonetic synthesized by HMM statistical parameter speech synthesis method and Dongxiang dialect phonetic synthesized by DNN speaker adaptive algorithm. This method can generate speech without text language. The generated Dongxiang dialect has high naturalness, similarity and quality of speech. Further work, we will optimize the structure of neural network, and use other in-depth learning methods and end-to-end speech synthesis methods to carry out experiments to improve the quality of synthetic speech.

Acknowledgment

The research leading to these results was partly funded by the National Natural Science Foundation of China (Grant No.11664036), High School Science and Technology Innovation Team Project of Gansu (2017C-03), Natural Science Foundation of Gansu (Grant No. 1506RJYA126).

References

[1] P. Wu, H. Yang, and Z. Gan, "Towards realizing mandarin-tibetan bilingual emotional speech synthesis with mandarin emotional training corpus," in *Data Science*, B. Zou, Q. Han, G. Sun, W. Jing, X. Peng, and Z. Lu, Eds. Singapore: Springer Singapore, 2017, pp. 126–137.

[2] Z. Wu, G. Cao, H. Meng, and L. Cai, "A unified framework for multilingual text-to-speech synthesis with ssml specification as interface,"*Tsinghua Science Technology*, vol. 14, pp. 623–630, 10 2009.

[3] S. Hertz, "Integration of rule-based formant synthesis and waveform concatenation: A hybrid approach to text-to-speech synthesis," 10 2002, pp. 87 – 90.

[4] Z. Wu, G. Cao, H. Meng, and L. Cai, "A unified framework for multilingual text-to-speech synthesis with ssml specification as interface," *Tsinghua Science Technology*, vol. 14, pp. 623–630, 10 2009.

[5] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, 01 2007.

[6] K. Tokuda, H. Zen, and A. Black, "An hmm-based speech synthesis system applied to english," pp. 227 – 230,10 2002.

[7] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 12 2014.

[8] Z. Liu, "Brief Notes on Dongxiang Language," *The Ethnic Publishing House*, 1981.

[9] G. Ma and Z. Liu, "A study on dongxiang language," *A Study of the Northwest Ethnic Groups*, pp. 167–184, 1986.

[10] R. Fu, Y. Li, Z. Wen, and J. Tao, "Automatic prosodic boundaries labeling based on fusing the duration of silence and the lexical features,"in *NCMMSC2017*, 2017.

[11] H. Yang and Z. Ling, "Prediction chinese prosodic boundary based on syntactic features," *Journal of Northwest Normal University (Natural Science)*, vol. 49, no. 1, pp. 41–45, 2013.

[12] R. Barra-Chicote, J. Yamagishi, S. King, J. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Communication*,vol. 52, pp. 394–404, 05 2010.

[13] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks,"*2013 ieee international conference on acoustics, speech and signal processing*, pp. 7962–7966, 10 2013.

[14] R. I. Damper, Y. Marchand, J. D. S. Marsters, and A. I. Bazin, "Aligning text and phonemes for speech technology applications using an em-like algorithm." *International Journal of Speech Technology*, vol. 8, no. 2, pp. 147–160, 2005.

[15] Z. H. Ling, S. Y. Kang, H. Zen, A. Senior, M. Schuster, X. J. Qian, H. M. Meng, and D. Li, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, 2015.

[16] P. Zhi, H. Yang, N. Song, "DNN-based emotional speech synthesis by speaker adaptation," *Journal of Chongqing University of Posts and Telecommunications(Natural Science)*, vol. 30, no. 5, pp. 673–679, 2018.

[17] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm," *IEEE Transactions on Audio Speech Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[18] P. C. Loizou, Speech Quality Assessment, 2011.