Synthesis of Tibetan Amdo based on VITS

Chao Wei[a], Guanyu Li*, Dongliang Chen[b]

Key Laboratory of Linguistic and Cultural Computing Ministry of Education（Northwest Minzu University), Lanzhou China

[a]y232430386@stu.xbmu.edu.cn, *xxlgy@xbmu.edu.cn,
[b]905285386@qq.com

Abstract. Tibetan Amdo, a significant dialect of the Tibetan language, currently lacks large-scale, high-quality speech databases. It faces challenges such as a limited number of researchers, incomplete and inaccurate coverage of the Tibetan phoneme lexicon, and subpar quality of synthesized speech. This paper employs the VITS framework for Tibetan Amdo speech synthesis, exploring the conversion of Tibetan characters into Latin letters for speech synthesis. The experimental results indicate that synthesizing natural and fluent Tibetan Amdo speech based on Latin alphabet conversion yields better outcomes, with a Mean Opinion Score (MOS) of 4.13, providing an effective approach for Tibetan Amdo speech synthesis.

Keywords: Tibetan Amdo, VITS, Latin alphabet, speech synthesis.

1 INTRODUCTION

As one of China's minority languages, Tibetan has a multitude of dialects, among which the three most prominent are the Weizang, Amdo, and Khampa dialects[1]. Amdo Tibetan, as an important dialect of the Tibetan language, possesses unique phonetic characteristics and linguistic structure[2]. The primary users of Amdo Tibetan are mainly located in the Tibet Autonomous Region of China, as well as in Tibetan-inhabited areas of provinces such as Qinghai, Gansu, and Sichuan[3]. The development of Amdo Tibetan speech synthesis technology has a profound impact on promoting local informatization and improving the quality of education.

Despite the rapid advancements in speech synthesis technology in recent years, there remain several shortcomings in the research of Amdo Tibetan. Firstly, existing speech synthesis systems predominantly focus on languages with abundant resources, and there is a relative scarcity of research dedicated to Amdo Tibetan. Secondly, due to the unique grammatical structure of Amdo Tibetan, the constructed phoneme lexicons for Tibetan are not comprehensive or accurate, leading to a deficiency in the naturalness and precision of synthesized speech[4].

To address the above problems, the study aims to explore how to better learn and synthesize Tibetan Amdo speech using VITS model.

A. Haldorai et al. (eds.), *Proceedings of the 2024 3rd International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID 2024)*, Atlantis Highlights in Intelligent Systems 11,
https://doi.org/10.2991/978-94-6463-490-7_43

2 MODEL ARCHITECTURE

VITS is a fully end-to-end text-to-speech (TTS) synthesis model that incorporates Variational Inference and Adversarial Learning concepts to achieve end-to-end speech synthesis from text directly to waveforms[5]5. The VITS model architecture is shown in Fig. 1.

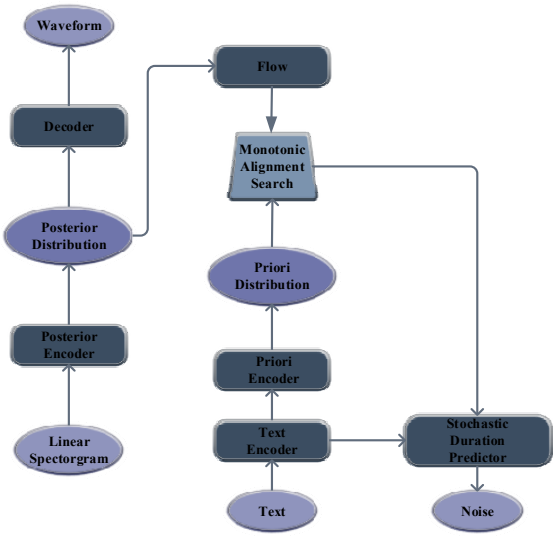

Fig. 1. VITS model architecture

VITS mainly consists of components such as a posteriori encoder, a priori encoder, random duration predictor, decoder, discriminator, etc[6]. Among them, the main task of the a posteriori encoder is to use the linear spectrogram of the target speechx_{lin} as input to learn the posterior distribution of potential representations of z from the given speech waveform data$q_{\varphi}(z|x_{lin})$[7] .The a priori encoder consists of a text encoder The a priori encoder consists of a text encoder and a normalized stream, and the a priori encoder builds a text-driven a priori distribution of potential representations z by learning the mapping between textual information and potential speech features $p_{\theta}(z|c_{text}, A)$ wherec_{text} is the phoneme extracted from the text and A is the alignment between the phoneme and the latent variable[8]. Stochastic Duration Predictor is a stream-based generative model, where the stochastic duration predictor generates the duration of each phoneme or word based on the input text features, and from the conditional inputsh_{text} from which the phoneme duration distribution is estimated. The decoder's generates speech waveforms based primarily on the potential representation z obtained from the a priori and a posteriori encoders[9]. The discriminator receives two types of input: a real speech waveform and a synthesized speech waveform generated by the decoder$\hat{y}$.

VITS as a whole is composed of VAE and GAN. Where the goal of VAE is to maximize the variational lower bound. In VITS, variational inference is used to bridge the gap between the generative model and the real data distribution in the text-to-speech

synthesis process. Variational inference is usually implemented by optimizing the reconstruction loss and KL Divergence.

The reconstruction loss is used to measure the similarity between the generated speech and the real speech. In VITS, the Mel spectrogram is used as the target data point, denoted as x_{mel}. the latent variable z is upsampled to the waveform domain$\hat{y}$ and $\hat{y}$is transformed to the Mel spectrogram domain$\hat{x}_{mel}$. the L1 loss between the predicted Mel spectrogram and the target Mel spectrogram is then used as the reconstruction loss:

$$L_{recon} = \|x_{mel} - \hat{x}_{mel}\| \quad (1)$$

The KL Divergence is used to measure the difference between the potential representation z posterior distribution of the posterior encoder output and the potential representation z prior distribution of the prior encoder output.The KL Divergence is:

$$L_{kl} = log\ q_{\varphi}(z|x_{lin}) - log\ p_{\theta}(z|c_{text}, A) \quad (2)$$

$$z{\sim}q_{\varphi}(z|x_{lin}) = N\big(z; \mu_{\varphi}(x_{lin}), \sigma_{\varphi}(x_{lin})\big) \quad (3)$$

3 TIBETAN ALPHABET TRANSCRIPTION AND CORPUS CONSTRUCTION

3.1 Transcription of the Tibetan alphabet into the Latin alphabet sequence

Tibetan transcription of Latin alphabet sequences is a method of converting Tibetan text into a representation using the Latin alphabet[10]. According to the Tibetan Latin Alphabet Transcription Program (Draft), the Tibetan Amdo text was transcribed through the Tibetan transcription Latin alphabet specification, e.g., "ཨ་བཟོ་འབྲོག་ཕ་དཔེ་མོ་ཅན་ཅོ་པོ་དགོ་ཡ།" was transformed into the Latin alphabetical sequence "'u-bzo-'brog-p-d-'di-mo-cn-co-po-dgo-y /"

3.2 The Construction of the Tibetan Amdo Corpus

A large amount of audio data and corresponding textual data were collected for the construction of the Tibetan Amdo corpus. For all audio, the sampling rate was set to 22kHz, the sampling precision was set to 16bit, the save format was WAV. The speech dataset is approximately 105 hours long and was recorded by 84 speakers containing 40 males and 44 females. The speech dataset is categorized into training set, validation set and development set. The training set contains 56,549 sentences recorded from 70 speakers. The validation set contains 4152 sentences recorded from 7 speakers; the test set contains 4149 sentences recorded from 7 speakers. Each speaker had about 800 sentences, and a small number of people had less than 700 sentences. Brief information about all the subsets in the speech dataset is shown in Table 1 below.

Table 1. Structure of the Tibetan Amdo corpus

subsets	Audio duration (hours)	male	female
training set	90.32	35	35
validation set	7.58	4	3
test set	7.54	3	4

The Tibetan text data are transformed into the corresponding Latin alphabet sequences after manual proofreading and sentence cutting to be saved in txt format. The transformed text corpus of Tibetan Amdo language is shown in Fig. 2.

```
/data2/user/wc/vits/data_Ando/resample_a_1_ymcr/a_1_ymcr-A54_23850.wav|1|'u-bzo-'brog-p-d-'di-mo-cn-po-dgo-y /
/data2/user/wc/vits/data_Ando/resample_a_1_ymcr/a_1_ymcr-A54_23851.wav|1|ls-khungs-thms-cd-dng-dpon-gns-kyi-ming-ni-yon-rgyl-rbs-kyis-gtn-'bebs-byed-dgos /
/data2/user/wc/vits/data_Ando/resample_a_1_ymcr/a_1_ymcr-A54_23852.wav|1|m'o-tse-tung-gi-dgongs-p'i-rlbs-chen-dr-ch-mthon-por-sgreng-b-de-red /
/data2/user/wc/vits/data_Ando/resample_a_1_ymcr/a_1_ymcr-A54_23854.wav|1|lo-rgyus-dng-dngos-yod-gng-zhig-gi-thog-ns-bshd-kyng- /
/data2/user/wc/vits/data_Ando/resample_a_1_ymcr/a_1_ymcr-A54_23855.wav|1|des-mth'-'khor-dm'-s'i-yul-du-sdod-mkhn-rnms-kyi-'ds-p'i-dus-kyi-rnm-p-mtshon-thub-bo / /
/data2/user/wc/vits/data_Ando/resample_a_1_ymcr/a_1_ymcr-A54_23856.wav|1|nub-phyogs-kyi-shes-rig-l-bg-chgs-gting-zb-bzhg-yod /
/data2/user/wc/vits/data_Ando/resample_a_1_ymcr/a_1_ymcr-A54_23857.wav|1|rgyng-bsgrgs-dng-brnyn-'phrin-grong-sde-tshng-mr-khyb-p /
/data2/user/wc/vits/data_Ando/resample_a_1_ymcr/a_1_ymcr-A54_23858.wav|1|mng-tshogs-kyi-khb-skud-tsm-yng-khyer-mi-chog-ces-pr-bsgyur /
```

Fig. 2. Segments of Tibetan Amdo transcribed text corpus results

4 DESIGN AND ANALYSIS OF THE EXPERIMENT

4.1 Experimental data

The experimental data selected for the experiment is a constructed corpus of Tibetan Amdo words. The corpus covers a wide range of speaker samples, including Tibetan Amdo speakers of different ages, genders and social backgrounds. All recording samples were pre-processed, including denoising, segmentation, text labeling and text transcription.

4.2 Experimental setup

The window size was set to 1024, the jump size was set to 256, the batch size used for the model was 8, and the training steps were5×10^5 , . The network was trained using the Adam W optimizer withβ_1 = 0.8 andβ_2 = 0.99 and weight decay λ = 0.01. The initial learning rate is2×10^{-4} .

4.3 Experimental results and analysis

In order to clarify the performance difference of Latin alphabet speech modeling unit in synthesizing Tibetan Amdo. During the training process, several loss metrics were recorded as shown in Fig. 3, including discriminator loss, generator loss, duration prediction loss and KL Divergence.

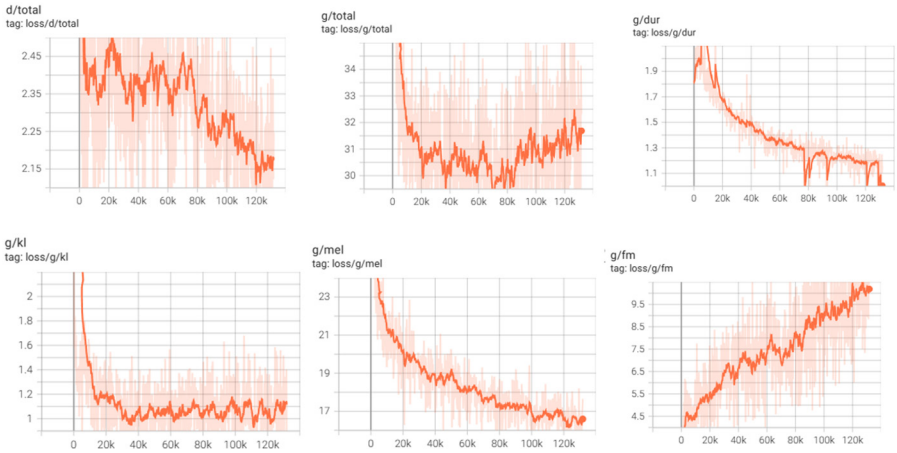

Fig. 3. VITS loss based on Latin alphabet modeling

During the training process, the encoder and decoder of the model were aligned and analyzed to assess the synergy of the internal components of the model, as shown in Fig. 4, and the results show that the model encoder and decoder are more accurate on the alignment.

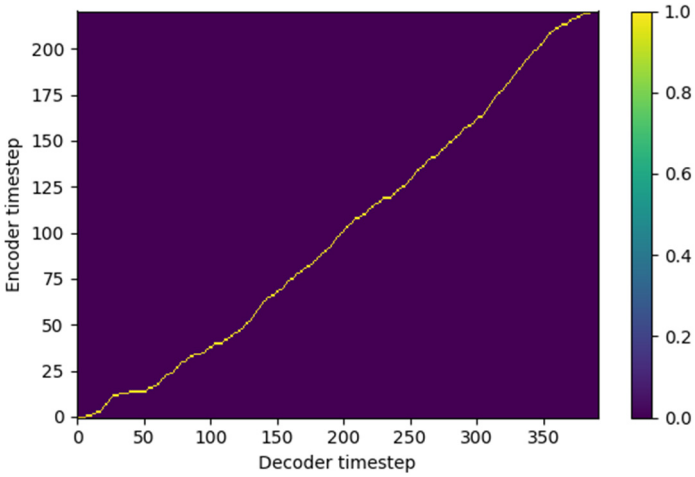

Fig. 4. VITS model alignment

Mel spectrogram analysis, as shown in Figure 5. It is further confirmed that the modeling model based on the Latin alphabet then performs well in terms of overall speech fluency.

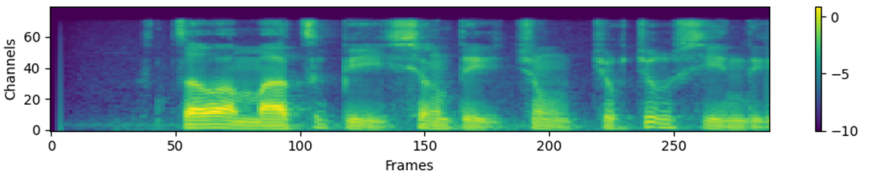

Fig. 5. Modeling VITS spectrogram based on Latin alphabet

The experiment invited seven evaluators with different majors and proficiency in Tibetan Amdo to rate the synthesized speech in MOS and calculate the average score. The results of comparing the MOS scores of the phoneme-based and the VITS model based on the Latin alphabet as the speech modeling unit are shown in Table 2.

Table 2. Synthesized speech MOS scores for different speech modeling units

system	MOS
VITS (phoneme sequence)	4.05
VITS (Latin alphabetical sequence)	4.13
Ground truth	4.36

According to the score results, the results of the synthesis based on the Latin alphabet modeling unit are better than the results of the synthesis based on the phoneme modeling unit.

5 CONCLUSIONS

In this paper, the research and practice of speech synthesis technology for Tibetan Amdo is discussed in depth, and the VITS model based on the Latin alphabet as the speech modeling unit is adopted for the experiments, which illustrates that the quality and naturalness of the synthesized speech can be effectively improved by transcribing the Latin alphabet through the Tibetan language.

ACKNOWLEDGMENT

This research was financially supported by a Basic scientific research business project of central universities (31920220010).

REFERENCES

1. Xu, X., Yang, L., Zhao, Y., & Wang, H. (2021). End-to-end speech synthesis for tibetan multidialect. Complexity, 2021, 1-8.
2. Sun, Jingwen. (2020). Deep Learning-based Speech Recognition of Tibetan Amdo Dialect Master's Degree (Dissertation, Northwest Normal University). Master https://link.cnki.net/doi/10.27410/d.cnki.gxbfu.2020.000830 doi:10.27410/d.cnki.gxbfu.2020.000830.

3. Luo, L., Li, G., Gong, C., & Ding, H. (2019, April). End-to-end speech synthesis for Tibetan Lhasa dialect. In Journal of Physics: Conference Series (Vol. 1187, No. 5, p. 052061). IOP Publishing.
4. Li, G., Li, G., & Song, Z. (2023). Tibetan Lhasa Dialect Speech Synthesis Method Based on End-to-End Model. Journal of Artificial Intelligence Practice, 6(1), 59-65.
5. Kim, J., Kong, J., & Son, J. (2021, July). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In International Conference on Machine Learning (pp. 5530-5540). PMLR.
6. Kim, M., Jeong, M., Choi, B. J., Ahn, S., Lee, J. Y., & Kim, N. S. (2022). Transfer learning framework for low-resource text-to-speech using a large-scale unlabeled speech corpus. arXiv preprint arXiv:2203.15447.
7. Mitsui, K., Zhao, T., Sawada, K., Hono, Y., Nankaku, Y., & Tokuda, K. (2022). End-to-end text-to-speech based on latent representation of speaking styles using spontaneous dialogue. arXiv preprint arXiv:2206.12040.
8. Shirahata, Y., Yamamoto, R., Song, E., Terashima, R., Kim, J. M., & Tachibana, K. (2023, June). Period VITS: variational inference with explicit pitch modeling for end-to-end emotional speech synthesis. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.
9. Li, J., & Zhang, L. (2023). Zse-vits: A zero-shot expressive voice cloning method based on vits. Electronics, 12(4), 820.
10. Lhagpa Dhondup, Chuje, Ojuk & Nyima. (2023). An end-to-end approach to Tibetan speech synthesis. Applied Acoustics (02), 324-332.