Unlocking the Potential: an evaluation of Text-to-Speech Models for the Bahnar Language

Giang Dinh Lu
University of Social Sciences and Humanities, Vietnam
Vietnam National University Ho Chi Minh City,
Ho Chi Minh City, Vietnam
giangdl@hcmussh.edu.vn

Tho Quan Thanh
Ho Chi Minh City University of Technology (HCMUT),
Vietnam
Vietnam National University Ho Chi Minh City,
Ho Chi Minh City, Vietnam
qttho@hcmut.edu.vn

Duc Nguyen Quang
Ho Chi Minh City University of Technology (HCMUT),
Vietnam
Vietnam National University Ho Chi Minh City,
Ho Chi Minh City, Vietnam
nqduc@hcmut.edu.vn

Hai Vu Hoang
Ho Chi Minh City University of Technology (HCMUT),
Vietnam
Vietnam National University Ho Chi Minh City,
Ho Chi Minh City, Vietnam
hai.vu.tharios19@hcmut.edu.vn

Quy Nguyen Tran
University of Social Sciences and Humanities, Vietnam
Vietnam National University Ho Chi Minh City,
Ho Chi Minh City, Vietnam
tranquynguyen@hcmussh.edu.vn

Abstract

In this study, we have implemented a sequential combination of two models, commencing with the utilization of the Grad-TTS model and subsequently employing the Hifi-GAN model. Grad-TTS is employed to enable the system to pronounce words in the Bahnar language without being constrained by the dataset. The strengths of Hifi-GAN have been fine-tuned for Bahnaric to enhance the quality of synthesized audio, aiming to produce a voice closely resembling the native Bahnar accent. The artificially generated sound from our model achieves a high level of naturalness.

Keywords

Bahnar language, speech synthesis, text-to-speech conversion, Mean Opinion Score (MOS), modified rhyme test

I. INTRODUCTION

Language is a crucial medium for communication and information transmission in a multi-ethnic country like Vietnam. The diversity of languages has generated an increasingly growing demand for effective machine translation systems, especially in the context of globalization and rising information exchange. Within this landscape, machine translation stands as an important tool to assist users in expressing and comprehending information in the languages of various ethnic communities.

The conversion of text into spoken Bahnar language is among the crucial tasks since Bahnar is widely used in fields like commerce, education, and culture. The Bahnar community consists of approximately 287,000 people, according to the government's statistics from 2019. They reside across three provinces: Kon Tum, Gia Lai, and Binh Dinh. However, creating accurate and natural Bahnar language audio remains a significant challenge due to the complexity in its syllabic structure, phonetics, and intonations.

Despite numerous research efforts and developments in the field of Vietnamese Text-to-Speech (TTS), Bahnar TTS is an entirely new task. Conventional TTS models based on rules and statistics have limitations in accuracy and the naturalness of synthetic audio. The advancement of deep learning models, especially Deep Neural Networks (DNN) and Recurrent Neural Networks (RNN), has created opportunities to enhance the efficiency of machine translation.

Therefore, in this study, we raise the question: Can a deep neural network-based text-to-speech model for the Bahnar language improve efficiency compared to traditional methods? In other words, the research focuses on evaluating and comparing the effectiveness of modern deep learning-based machine translation models in creating audio for this language. The aim is to achieve better, more natural, and practical sound generation for various real-world applications.

In Vietnam, since 2010, researchers have been applying Text-to-Speech synthesis technology to the Vietnamese language. For instance, Võ Quang Diệu Hà and colleagues developed a speech synthesis system based on concatenative techniques [1]. They used syllables as the basic unit to create speech in their TTS system. A primary limitation of this method is the complex selection of speech database units by the system. Furthermore, the final product often doesn't express intonation and rhythm well.

Following that, Nguyễn Trang constructed a Vietnamese TTS system, developed on the Mary TTS platform, using a statistical-based speech synthesis method [2] with an architecture based on the Hidden Markov model [3]. The research results indicate a specific need for further enhancement in processing tonal aspects to generate better automated speech for Vietnamese.

In 2019, Lâm Phùng Việt and colleagues utilized a Deep Learning model to improve their TTS system [4]. Their system is based on the Tacotron 2 model and the WaveGlow neural vocoder. Comparing it to the previous SPSS approach, the results show significantly higher performance.

In 2020, author Chenfeng Miao utilized the Tacotron 2 model and the Hifi-gan vocoder to synthesize speech. The research results showed that the synthesized sound achieved up to 89.3% similarity with the natural speaker's voice [5]. However, there still exists a significant challenge when processing long sentences, especially in expressing intonation and accurately pronouncing borrowed words.

Author Vadim Popov suggests that the Grad-TTS model, when applied in TTS systems, can address the limitations of previous models, such as the need for extensive data, unnatural sound, lack of intonation, and difficulty handling long passages. Currently, the Grad-TTS model has shown positive results in English. As the Bahnar language lacks the

intonation found in English and has a simpler syllabic structure, there is hope that it will yield positive results in synthesizing Bahnar speech.
Classical methods used to build Text-to-Speech systems include articulatory synthesis.However, there still exists a significant challenge when processing long sentences, especially in expressing intonation and accurately pronouncing borrowed words.

Author Vadim Popov suggests that the Grad-TTS model, when applied in TTS systems, can address the limitations of previous models, such as the need for extensive data, unnatural sound, lack of intonation, and difficulty handling long passages. Currently, the Grad-TTS model has shown positive results in English. As the Bahnar language lacks the intonation found in English and has a simpler syllabic structure, there is hope that it will yield positive results in synthesizing Bahnar speech.

Classical methods used to build Text-to-Speech systems include articulatory synthesis [6], formants synthesis [7], concatenative synthesis [8] and statistical parametric speech synthesis [9].

Applying advanced techniques and achieving high performance in languages with limited data, such as Bahnar, is a challenge. In this report, we'll describe the process of selecting and building a suitable artificial intelligence model, as well as constructing a suitable phoneme set for the Bahnar language. Therefore, this report also serves as a reference document and is one of the first research works in developing an artificial speech synthesis system for Bahnar.

The main goal of the text-to-speech system is to convert arbitrary text into spoken language in the form of audio. Text processing and speech synthesis are the two main components of the text-to-speech system. The aim of text processing is to analyze the input text and generate phoneme sequences. These phonemes are marked by the speech synthesis component or are synthesized from parameters collected from a sufficiently large audio data source. To produce natural speech synthesis, the text processing is structured in an appropriate sequence, matching the phonemes to the arbitrary input text.

Currently, several models have been applied for text-to-speech conversion. However, selecting an appropriate model for an isolated tone language like Vietnamese requires specific attention. This becomes particularly crucial for Bahnar, a language lacking tonality. In terms of grammatical structure and vocabulary, Vietnamese and Bahnar share many similarities.

II. METHOD

One of the prominent methods applying Artificial Intelligence to this issue is Tacotron 2 [10], utilizing a combined structure of Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), Tacotron 2 has become popular in this field. However, Tacotron 2 has not met the requirements for natural speech synthesis, especially for languages like Vietnamese and Bahnar, as it has only been tested on English. Therefore, instead of the conventional approach of Tacotron 2, we have developed an end-to-end process using the Grad-TTS architecture [11], a method involves using a neural network utilizing a denoising diffusion probabilistic model. This approach is also suitable for related studies in languages close to Bahnar, such as Vietnamese [5]. Then, the results are transferred for processing by the Hifi-GAN model.

The process involves two main stages. In the first stage, audio data is prepared along with its corresponding written text. The TTS model is applied to each social feature set of the contributors, with each voice created corresponding to gender, age group, and dialect.

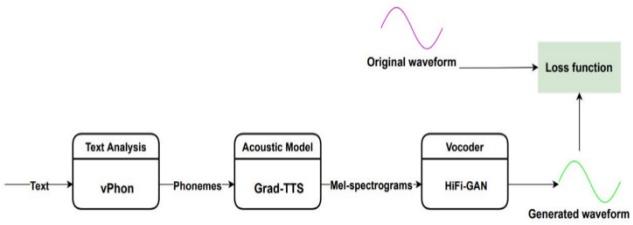

Figure 1: TTS structure: Text analysis using vPhon, Grad-TTS acoustic model, Vocoder using Hifi-GAN

The recording equipment requirements include using microphones with a maximum recording distance of 10 cm. Each word must be recorded at least 20 times in different contexts. The minimum recording time for a contributor is 60 hours.

The language data is categorized into two types: natural language data and pre-arranged data. Natural data comprises unprepared natural conversations. Pre-arranged data involves a list of Bahnar language phonetic types, aiming to cover all syllabic forms, including various consonants and vowels. This is to ensure the computer can store numerous phonetic contexts while requiring minimal phonetic inputs.

The second stage of the implementation process involves three primary processing steps. The first step is Text Analysis, where the input is the text needing a voice, and the output is the corresponding speech for that sentence.

Our Text-to-Speech system consists of three main components. Firstly, the Input Text Analysis module processes the text into a phoneme sequence corresponding to the sound sequence of Bahnar language. This phoneme sequence is suitable for processing by the neural network. Since the system involves converting from Vietnamese to Bahnar, the system integrates both the Vietnamese and Bahnar writing systems. The computer will be trained to recognize which characters are used in Vietnamese and which characters are used for Bahnar. Characters used in Bahnar writing are as follows:

- Vowel characters: i, ĭ, iê, iễ, ê, ễ, e, ĕ, u, ŭ, ươ, ưở, ơ, ở, a, ă, u, ŭ, uô, uỗ, ô, ỗ, o, ŏ.
- Consonant characters: p, t, th, ph, ch, c, q, h, m, n, nh, ng, b, d, b', d', g, h, k, tr, s, l, r, y.

We can see that each language has its distinct characteristics, and it's not feasible to efficiently use input analysis modules from different languages. Thus, we have built a phoneme set specifically designed for both languages.

Each word in the input text will be mapped to a corresponding phoneme sequence based on the alphabet for both languages. The input text (INPUT) after processing by the analysis module results in a corresponding phoneme sequence (PROCESSED). This character string also serves as input for the training and utilization of the Grad-TTS model.

INPUT: Inh kăt 'ba kơpưng adơi.

PROCESSED: I- ɲ-k-ă-t- ɓ-a-k- ə-p- ɨ- ŋ-a-d- ə-j.

(English: I'm harvesting rice in the field.)

The input can be a sentence or a paragraph. The computer will analyze each word in the sentence into phonemes. Based on the sound database for phonemes, which includes consonants, vowels, and semi-vowels, the computer will synthesize these phonemes to reproduce the sound. The sound produced by the model is not simply a replay of previously recorded sound but a synthesis process from phonemes in the Bahnar language.

Sound quality is maintained by recording speech in a naturally noisy environment. The recorded audio data is then converted to digital data using software like Wavsurfer. The current system uses 16-bit resolution data along with a sampling rate of 22.05 KHz. (A sampling rate of around 20 KHz is considered sufficient to maintain sound quality). Recording of each phoneme is done at a 22.05 kHz sampling rate. The phoneme sound database includes all consonants and vowels.

We utilize the conversion method of vPhon [12] The rule of phoneme analysis for the Vietnamese and Bahnar transcription. The computer learns from improved phonemes because they represent sound more closely. Moreover, vPhon is specifically designed for Vietnamese. For the most part, the phonetic learning structure corresponds between Vietnamese and Bahnar. Bahnar has the advantage of resource-saving without intonation.

The second module in the process is the acoustic model based on Grad-TTS. The input to this model is the pseudo-phoneme set, and through the training process, it generates a mel-spectrogram. A mel-spectrogram is a spectral representation of sound waves, encompassing frequency and time. It can reproduce detailed information about frequency bands predominating at each moment in the sound wave. From the mel-spectrogram, you can extract the original sound wave through the inverse problem.

The final step is the vocoder processing, where you use the HiFi-GAN network to generate Bahnar sound. What's unique about a vocoder is that it directly calculates on the mel-spectrogram, thus not depending on the input language. Therefore, HiFi-GAN can be directly applied to Bahnar without retraining. The results show that HiFi-GAN still achieves high-quality sound compatible with the Bahnar language, making the synthesis of Bahnar speech efficient and high-quality.

III. EXPERIMENTS

A. Dataset interpretation

Dataset information

A dataset of 10,000 speech segments from a single speaker, each 3-7 seconds long with corresponding phonemes, was collected and labeled with unique IDs that correspond to the names of the corresponding .wav files. The contributor's voice is clear, loud, expressive, and low-noise, with minimal variation in delivery or tone.

Audio segmentation

In the beginning, audiobooks that last approximately 20 hours are first selected. Then, they are exported to .wav format using the pydub library. On later version of pydub, we use the split on silence to segment the audio based on pauses. It is worth noting the two most important parameters of evaluating the freshness of data, which are *minimum_silence_length* (in milliseconds), which is the minimum duration of silence used to split the audio, and *silence_thresh* (in dBFS), which is the threshold below which sound is considered silence (the default is -16 dBFS). Choosing incorrect values for these parameters can lead to noisy data. Therefore, the best value can be determined by examining the waveform of an audio clip to identify the speaker's segmentation pattern. After splitting the audio, we collect all speech segments that are between 3 - 7 seconds long and use the cognitive service from Azure to generate text transcripts for each speech segment.

Preprocessing

Outliers are removed from the dataset by plotting a scatter graph of the duration of each audio record. Linear regression is then applied using two variables: duration and the number of words in each file, since the number of words in a phrase is proportional to the audio file's duration. Outliers on the graph are eliminated.

Next, the text is normalized by converting numbers, ordinals, and currency units into complete words (UTF-8). All punctuation is then removed from the text to enable the model to learn segmentation independently.

Finally, a phonemic analysis is generated for each audio record. This is the second attribute of the dataset used for model training. The transformation rules of vPhon [18] are applied to analyze phonemes for each audio record. The machine learns better from phonemes than from the original text because phonemes are representations closer to sound.

B. Bahnar synthesis model

The model is developed on the PyTorch platform. Overall, the data is throughput for over 2500 iterations on a server equipped with an NVIDIA RTX 3090 GPU. The accompanied audio recordings, with a total duration of approximately 57 hours, were divided into smaller segments to better serve the training and evaluation process. Before being fed into the TTS system, all audio files underwent filtering and preprocessing process to ensure no noise residues nor distortions. In addition, all input text was analyzed and converted into phoneme forms to generate suitable data for the model.

DCA-Tacotron 2 model is the forerunner on publicly available data (InfoRe dataset), using the same parameters and configurations as in the original paper. Subsequently, we evaluated the effectiveness of our proposed data enrichment method and framework on both the InfoRe dataset and our ViSpeech dataset, comparing it with the baseline model. Conducting an exploratory experiment on the Bahnar language, a low-resource language, was intended to demonstrate the utility of our proposed approach. For each dataset, 90% of the samples are used for training and the remaining 10% are reserved for validation. Previously, we selected and set aside 40 sentences for testing.

IV. RESULTS AND DISCUSSION

Once the Bahnar audio synthesis process was completed, the natural voices of Bahnar collaborators from the Gia Lai region were directly compared with the voices generated by the model. To evaluate the reliability of the artificial Bahnar sound, the Comparative Mean Opinion Score (CMOS) test was utilized. With such test, a total of 60 pairs of sentences with

corresponding audio were randomly selected from the dataset. The input text was then fed into the model to generate the synthetic Bahnar voice. This artificial Bahnar voice was then directly compared to the natural Bahnar voice..

To evaluate the quality of the artificial Bahnar voice, 30 Bahnar informants from three areas: Gia Lai, Kon Tum, and Binh Dinh, were invited to participate in the assessment. For each dialect (Gia Lai, Kon Tum, and Binh Dinh), three sets of assessment samples were created. Each collaborator listened to both versions of the voice (from a real person and the model-generated one) and then assessed the similarity level on a scale ranging from -3, meaning the model-generated voice is much worse than the real voice, to +3, meaning vice-versa. The obtained scores were then used to evaluate the quality of the artificial Bahnar voice.

Finally, the average scores were calculated and displayed in Table 1. The results indicate that the artificial intelligence model is only slightly inferior to the human real voice. Based on this outcome, it can be concluded that the selected model has achieved high performance and quality for the proposed artificial voice generation system.

Table 1: CMOS results comparing
real voice and neural network-generated voice."

System	**CMOS**
Ground truth	0.000
OURS	-0.3276

Furthermore, we have developed two additional versions, one based on a female reading and another based on both male and female voices. Both models have yielded promising results and are intelligible to the Bahnar people. This implies that we have expanded the model's application to accommodate both male and female voices, enabling the creation of high-quality sound for various purposes and user groups within the Bahnar community.

V. CONCLUSIONS

In this study, we established the initial foundation for developing a natural speech synthesis technology for the Bahnar language in Vietnam. Despite the language's unique characteristics and limited dataset, our approach yielded promising and effective results. The speech generated by our system closely mimicked human speech, exhibiting flexibility in handling input text. The speech quality from our system was evaluated as clear and easily understandable, able to simulate both male and female voices as required.

Evaluation results demonstrated that the system is capable of producing high-quality sound, and the voice conversion model trained on Grad-TTS source data combined with the Hifi-GAN vocoder showed superior performance. Future work may involve enhancing the quality of noise-disturbed sound.

ACKNOWLEDGMENT *(Heading 5)*

This research is funded by Ministry of Science and Technology (MOST) within the framework of the Program "Supporting research, development and technology application of Industry 4.0" KC-4.0/19-25 – Project "Development of a Vietnamese- Bahnaric machine translation and Bahnaric text-to-speech system (all dialects)" - KC-4.0-29/19-25.

REFERENCES

[1] V. Q. D. Ha, N. M. Tuan, C. X. Nam, P. M. Nhut, and V. H. Quan, "Vos: The corpus-based vietnamese text-to-speech system," 2010.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech communication,* vol. 51, no. 11, pp. 1039-1064, 2009.

[3] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *Eee assp magazine,* vol. 3, no. 1, pp. 4-16, 1986.

[4] V. L. Phung, H. K. Phan, A. T. Dinh, K. D. Trieu, and Q. B. Nguyen, "Development of Zalo Vietnamese Text-to-Speech for VLSP 2019."

[5] Tung Tran *et al.*, "Naturalness improvement of Vietnamese Text-to-Speech System using Diffusion Probabilistic modelling and Unsupervised Data Enrichment," in *The First International Conference on Intelligence of Things (ICIT 2022)*, 2022.

[6] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*, 2016, pp. 1558-1566: PMLR.

[7] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, vol. 27, M. I. Jordan, Y. LeCun, and S. A. Solla, Eds., 2014.

[8] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794-2802.

[9] K. Kumar *et al.*, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in neural information processing systems*, vol. 32, 2019.

[10] J. Shen *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2018, pp. 4779-4783: IEEE.

[11] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning*, 2021, pp. 8599-8608: PMLR.

[12] J. Kirby, "VPhon: A Vietnamese Phonetizer," ed, 2008.