Science & Technology Development Journal – Engineering and Technology 2024, ():1-12

Open Access Full Text Article

Research Article

Voice conversion for natural-Sounding speech generation on low-Resource languages: A case study of bahnaric

Dang Tran Dat[1,2], Tang Quoc Thai[1,2], Duc Q. Nguyen[1,2], Vo Duy Hung[1,2], Quan Thanh Tho[1,2,*]

ABSTRACT

Bahnar is an ethnic minority group in Vietnam, prioritized by the government for the preservation of their cultural heritage, traditions, and language. In the current era of AI technology, there is substantial potential in synthesizing Bahnar voices to support these preservation endeavors. While voice conversion technology has made strides in enhancing the quality and naturalness of synthesized speech, its focus has predominantly been on widely spoken languages. Consequently, low-resource languages like the Bahnaric language family encounter numerous disadvantages in voice synthesis. This study addresses the formidable challenge of synthesizing natural-sounding speech in low-resource languages by exploring the application of voice conversion techniques to the Bahnaric language. We introduce the BN-TTS-VC system, a pioneering approach that integrates a text-to-speech system based on Grad-TTS with voice conversion techniques derived from StarGANv2-VC, both tailored specifically for the nuances of the Bahnaric language. Grad-TTS allows the system to articulate Bahnaric words without vocabulary limitations, while StarGANv2-VC enhances the naturalness of synthesized speech, particularly in the context of low-resource languages like Bahnaric. Moreover, we introduce the Bahnaric-fine-tuned HiFi-GAN model to further enhance voice quality with native accents, ensuring a more authentic representation of Bahnaric speech. To assess the effectiveness of our approach, we conducted experiments based on human evaluations from volunteers. The preliminary results are promising, indicating the potential of our methodology in synthesizing natural-sounding Bahnaric speech. Through this research, we aim to make significant contributions to the ongoing efforts to preserve and promote the linguistic and cultural heritage of the Bahnar ethnic minority group. By leveraging the power of AI technology, we aspire to bridge the gap in speech synthesis for low-resource languages and facilitate the preservation of their invaluable cultural heritage.

Key words: Bahnaric speech synthesis, text-to-speech, natural-sounding voice conversion

[1]*Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Vietnam*

[2]*Vietnam National University Ho Chi Minh City (VNU-HCM), Vietnam*

Correspondence

Quan Thanh Tho, Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Vietnam

Vietnam National University Ho Chi Minh City (VNU-HCM), Vietnam

Email: qttho@hcmut.edu.vn

History

- Received: 08-9-2023
- Accepted: 27-3-2024
- Published Online:

DOI :

INTRODUCTION

The Bahnar or Ba-Na (Vietnamese pronunciation: [□a□na□]) represents a distinct ethnic minority within the diverse tapestry of ethnic populations in Vietnam. Contemporary efforts spearheaded by the Vietnamese government aim to enhance their integration through advancements in socio-cultural and scientific literacy. A significant portion of this endeavor includes translating key documents into the Bahnaric language by governmental and community stakeholders. Concurrently, there is growing interest among domestic research groups to devise automatic translation systems for Vietnamese to Bahnaric ethnolects. Notwithstanding these advancements, the distinct characteristics of the Bahnar, given their status as a smaller ethnic faction, result in hesitations in engaging with the predominant Kinh (Vietnamese) population. This occasionally impedes their complete access to written information. Thus, conveying information with native-like Bahnaric speech could significantly enhance accessibility for this community.

Modern TTS (text-to-speech) systems [1] can assist in pronouncing words from text based on a trained dataset. However, these systems require a substantial amount of training data. For extremely low-resource languages like Bahnaric, gathering a high-quality training dataset becomes particularly arduous, resulting in suboptimal pronunciation outputs. For the small Bahnaric ethnic group, this also poses significant challenges to communication.

Another solution is to develop voice conversion systems [2] that convert the voice quality to match that of a genuine Bahnar individual. Due to the low-resource nature of Bahnaric, we have proposed an effective approach that combines the Grad-TTS model [3] and the StarGANv2-VC model [4]. The use of the Grad-TTS model enables the system to pronounce an unlimited vocabulary from available texts. Meanwhile, the StarGANv2-VC model assists in generating a converted voice from an existing Bahnaric

Cite this article : Dat D T, Thai T Q, Nguyen D Q, Hung V D, Tho Q T. **Voice conversion for natural-Sounding speech generation on low-Resource languages: A case study of bahnaric**. *Sci. Tech. Dev. J. – Engineering and Technology* 2024; ():1-12.

voice. Particularly, the combination of Grad-TTS and StarGANv2-VC aids in refining and cleaning words and phonemes that Grad-TTS has not generated well, especially when trained from low-resource and low-quality sources like direct recordings of Bahnaric people's speech. In addition, we also introduce the HiFi-GAN-BN model, a variant of HiFi-GAN [5] pre-trained by Bahnaric voice, to resemble the Bannaric accents better when transforming the mel-spectrogram output of StarGANv2-VC into human-listenable waveform.

We have experimented with our system, known as BN-TTS-VC, using real-world data collected from the Bahnar community in the provinces of Gia Lai, Kontum, and Binh Dinh. When evaluated by human assessments, we have obtained favorable results.

The remainder of the paper is organized as follows. Section 2 describes previous works which are related to our study. Section 3 gives details of the Bahnaric phonological system. Section 4 describes the methodology to develop the BN-TTS-VC system. Section 5 presents the experiment results. Section 6 provides a discussion of the results obtained from our experiment. Section 7 presents conclusions and future work.

RELATED WORKS

Text-to-speech techniques

Text-to-speech synthesis is a task that involves converting written text into spoken words. The goal is to generate synthetic speech that sounds natural and resembles human speech as closely as possible. Classical methods used to construct text-to-speech systems include articulatory synthesis [6], formant synthesis [7], concatenative synthesis [8], and statistical parametric speech synthesis [9]. These methods usually generate a voice with less of a natural or lack of emotion and the voice quality is low due to containing screeching and jerking sounds. Certain end-to-end models such as ClariNet [10], FastSpeech 2s [11], and EATS [12] that create audio directly from text have been proposed based on simplification of text analysis modules and directly taking character strings or phonemes as input, also as to simplify acoustic properties with timbre spectra. The advantages of neural network-based speech synthesis over previous Text-to-speech systems include high voice quality in terms of intelligibility and naturalness as well as less reliance on the construction of input properties. Concerning Vietnamese text-to-speech systems, the Tacotron 2 acoustic model [13] is considered a classical deep-learning method that is widely applied in these systems. The ZALO group developed a Text-to-speech system [14] based on Tacotron 2 [13] and WaveGlow [15] whose performance of their system is superior to the statistical parametric speech synthesis classical method.

Voice conversion techniques

Voice conversion (VC) is a technique for converting one speaker's voice identity into another while preserving linguistic content. Though most voice conversion methods that require parallel utterances achieve high-quality natural conversion results, it strongly limits the conditions to apply. Regarding non-parallel voice conversion methods, it can mainly be divided into three categories. *Auto-encoder approach* [16–19] requires carefully designed constraints to remove speaker-dependent information, and the converted speech quality depends on how much linguistic information can be retrieved from the latent space. By contrast, *GAN-based approaches*, such as CycleGAN-VC3 [20] use a discriminator that teaches the decoder to generate speech that sounds like the target speaker. Due to the lack of learning meaningful features from the real data in the discriminator, this approach often suffers from problems such as dissimilarity between converted and target speech, or distortions in voices of the generated speech. On the other hand, *TTS-based approaches* like Cotatron [21], AttS2S-VC [22], and VTN [23] extract aligned linguistic features from the input speech to give the converted speaker identity that is similar to the target speaker identity. However, the text labels for this approach are not often available at hand.

BAHNARIC HONOLOGICAL SYSTEM

To develop a speech synthesis system, it is essential to construct a phonological system for this particular language. Figure 1 illustrates an example of a Bahnaric language text. We can see that the language has its characteristics, and using the input parsing modules of other languages is impossible. Therefore, we analyze this language elaborately and build a set of pseudo-phonemes for the Bahnaric language, which is suitable input for the text-based speech generation model. The set of pseudo-phonemes is shown in Figure 2.

Each word in the input text will be compared to the corresponding phoneme sequence based on the above alphabet. An example is shown in Figure 3. From the text (INPUT) passed through the analyzer, the result is the corresponding phoneme sequence (PROCESSED). That sequence is also the input for training and using the TTS model.

```
adriêng nganh y teâ adriêng bet teêk weêk pôloêk phun bôgang bet sôhmeêch
minh suaât kua tri giaê 01 trieâu ñoâng
tôplih lôêm tôdrong tôme rong jaêng pran ñeh oei xa vinh kim
trö jeân pôm minh sônaêm kung thu yoêk ñei khoang 60 trieâu ñoâng
rim moâ hinh anu jôh pôjing thu yoêk tôpaê pônhoâm lö naê ma adriêng pôm
```

Figure 1: An example of text in Bahnaric language.

```
a b c d e f g h i j k l m n o p q r s t u v w x y z à á â ã ä å è ê ë ì í ï ñ ò ó ô õ ö
ø ù ú û ý ă ĩ đ ũ ơ ư ạ ả ấ ầ ẫ ậ ằ ẵ ẻ ế ề ể ễ ệ ỉ ị ọ ố ồ ổ ỗ ộ ớ ờ ở ỡ ợ ụ ủ ứ ừ ữ ự
```

a) Monophonic

```
ia iă ie ië iô iö ua uă ue uë uê
```

b) Diphthong vowels

```
bl br by ch dj dr gl gr gy hl hm hn hñ hr hy jr kh kl kr ky ly ml mr ny my ñr ng ph pl pr py sr th tr ty
```

c) Double consonants

```
hng ngl nhr
```

d) Triple consonants

Figure 2: A set of pseudo-phonemes for Bahnaric language.

```
INPUT:     adriêng nganh y teâ
PROCESSED: a-d-r-i-ê-ng ng-a-n-h y t-e-â
```

Figure 3: An example of an input text analyzer in Bahnaric language.

RESEARCH METHODOLOGY

Overview of the combined system of Text-to-speech and voice conversion for Bahnaric language

This system is constructed based on two main modules including Text-to-speech and Voice Conversion, as illustrated in Figure 4. The first module gets the Bahnaric language text as input to generate a native voice with the content of the input text. There are two sub-models in this module, which are the vocoder and acoustic model. While the acoustic model generates acoustic properties directly from input phonemes mentioned in Section 3, a vocoder transforms these features into sound waveforms. After that, the sound waveforms are passed to the Voice conversion module for generating the other types of voice of native based on the reference voice. This module is built from three main component models for the purpose of extracting the characteristics of voice, converting the voice, and transforming the mel-spectrogram into a human-listenable waveform.

Grad-TTS system for Bahnaric speech synthesis

According to our research, there so far has been no reported work on building an artificial voice generation system for the Bahnaric language. In this domain, there is an existence of certain different characteristics between the Bahnaric and other popular languages. Therefore, applying techniques with high efficiency in those languages to Bahnaric is a highly complex problem.

One of the typical methods of applying AI to solve this problem is Tacotron 2 [13], which uses the architecture of recurrent neural network (RNN) and convo-

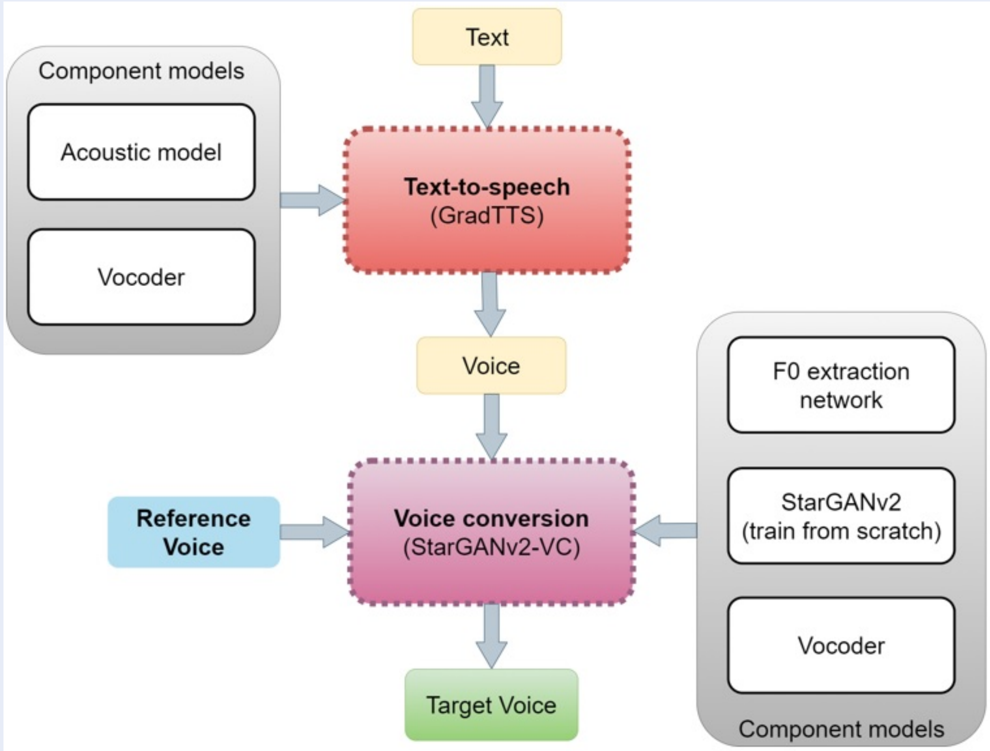

Figure 4: An overall architecture of the system.

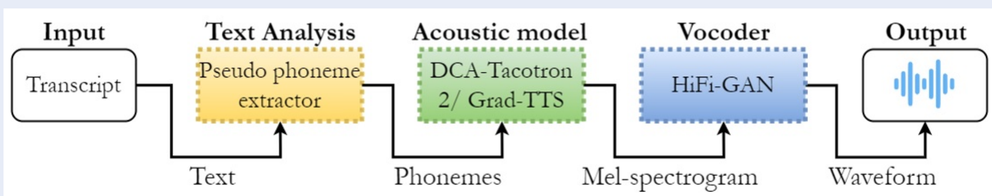

Figure 5: An artificial voice generation system. There are 3 main components in this system. The input to the system is the text that needs to be voice-generated, and the output will be the voice for the corresponding sentence.

lutional neural network (CNN). Tacotron 2 has been and is being used commonly in this field. However, this model has not yet met the requirements for the naturalness of the generated voice, especially for languages such as Vietnamese and Bahnaric, because the original Tacotron 2 has only been experimented with English. Therefore, instead of using the common approach of Tacotron 2, we develop an end-to-end process using the Grad-TTS architecture [3], a neural network using the denoising diffusion probabilistic model. This approach is also consistent with related studies in the group of languages closely related to Bahnaric, such as Vietnamese [24].

Our text-to-speech system consists of three main components, as shown in Figure 5. First, the *Text Analysis* module parses the text into a pseudo-phonetic representation, which is suitable for neural network processing.

The second module is an acoustic model based on Grad-TTS, from the input of which is a set of pseudo-phonemes, it goes through the training process to generate the mel-spectrogram representation. Mel-spectrogram is a representation in the form of a spectrum of sound waves, consisting of two dimensions, frequency and time. Mel-spectrograms can be extracted directly from the sound wave and contain more detailed information about the frequency bands that prevail at each moment in the sound wave. Conversely, it is also possible to extract sound waves from the mel-spectrogram through the inverse problem.

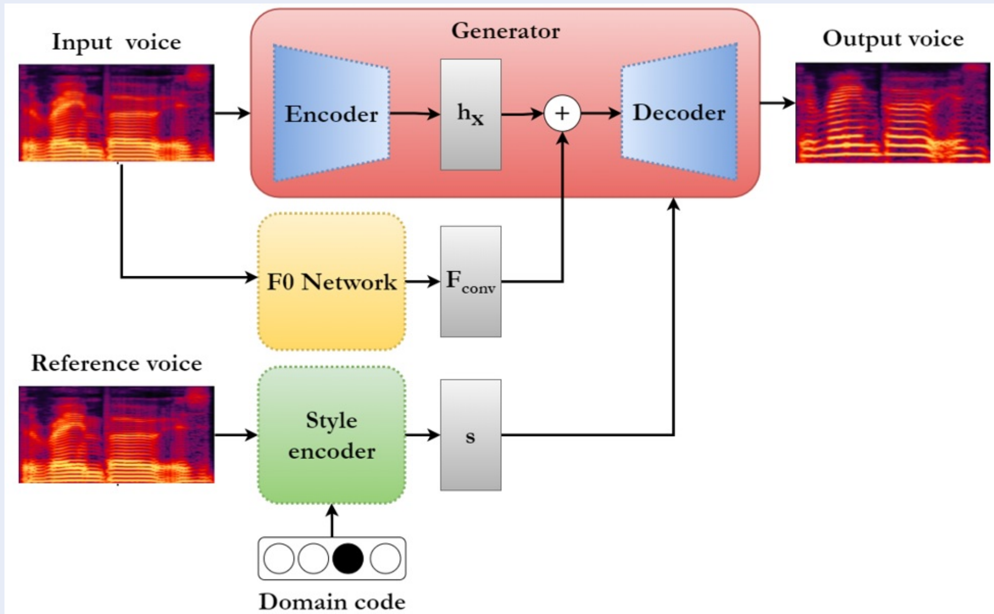

Figure 6: The overall framework of StarGANv2-VC.

The detailed architecture of this model can be consulted through related previous works, and in this publication, in order to be accessible to a wide audience and not be too technical, we do not go through the details of this model.

The final step is performed by the vocoder. We use the HiFi-GAN network[5] to convert the output from mel-spectrogram to waveform. More specifically, instead of using pre-trained HiFi-GAN for the English language, we retrained a pre-trained HiFi-GAN-BN system from Bahnaric to generate the final voice.

StarGANv2-VC model for Bahnaric voice conversion

The Grad-TTS model can pronounce without limitation the vocabulary of Bahnar texts, but due to the low resource characteristics of this language, the sound quality still lacks the naturalness of humans. To overcome this problem, we propose to use the StarGANv2-VC model to convert the voice synthesized by Grad-TTS into a sample voice of the native Bahnar. The proposed methodology has been developed based on the foundational principles of StarGANv2-VC[4], a pioneering framework that employs a solitary discriminator and generator to produce a diverse array of images across various domains. These domains are characterized by the utilization of domain-specific style vectors sourced either from the style encoder or the mapping network. In the domain of voice conversion, each speaker is treated as a discrete domain. To ensure the maintenance of consistent fundamental frequency (F0) conversion, the network architecture has been thoughtfully enhanced through the integration of a pre-trained joint detection and classification (JDC) F0 extraction network[25]. Figure 6, presented herein, offers an illustrative depiction of the StarGANv2-VC framework for elucidation. In StarGANv2-VC, a sample $X \in X_{y_{src}}$ from the source domain $y_{src} \in Y$ undergoes transformation to a corresponding sample $\widehat{X} \in X_{y_{trg}}$ in the target domain $y_{trg} \in Y$ via a mapping function, denoted as $G : X_{y_{src}} \to X_{y_{trg}}$. Crucially, this transformation is achieved independently of parallel data.

Throughout the training process, the selection of the target domain, $Y_{trg} \in Y$, is random, and its style code, $s \in S_{y_{trg}}$, is encoded through a style encoder. This encoder utilizes a reference input $X_{ref} \in X$ from the target domain to produce the style code, designated as $s = S\left(X_{ref}, y_{trg}\right)$. Using a mel-spectrogram $X \in X_{y_{src}}$ from the source domain $y_{src} \in Y$ and the target domain $y_{trg} \in Y$, our model is trained by minimizing the subsequent loss functions.

Adversarial loss. The generator is trained to produce a new mel-spectrogram, denoted as G(X,s), from an input mel-spectrogram X and a style vector s by utilizing the adversarial loss.

$$L_{adv} = E_{x,y_{src}}\left[\log D\left(X, y_{src}\right)\right] + E_{x,y_{trg},s}\left[\log\left(1 - D\left(G\left(X, s\right), y_{trg}\right)\right)\right] \quad (1)$$

where $D(\cdot, y)$ represents the output of the real/fake classifier of the domain $y \in Y$.

Adversarial source classifier loss. Another adversarial loss function, involving the source classifier C, is employed (refer to Figure 7).

$$I_{advcls} = E_{x,y_{trg},s}\left[CE\left(C\left(G\left(X,s\right),y_{trg}\right)\right)\right] \quad (2)$$

where CE($\cdot$) denotes the cross-entropy loss function.

Style reconstruction loss. To guarantee that the style code can be reconstructed from the generated samples, the style reconstruction loss is used.

$$L_{sty} = E_{x,y_{trg},s}\left[||s - S\left(G\left(X,s\right),y_{trg}\right)||_1\right] \quad (3)$$

Style diversification loss. The different samples must be generated with different style codes. We enforce the generator to learn this constraint by maximizing the style diversification loss. In addition to maximizing the mean absolute error (MAE) between generated samples, the MAE of the F0 features between samples generated with different style codes is also maximized.

$$L_{ds} = E_{X,s_1,s_2,y_{trg}}\left[||G\left(X,s_1\right) - G\left(X,s_2\right)||_1\right] + E_{X,s_1,s_2,y_{trg}}\left[||F_{conv}\left(G\left(X,s_1\right)\right) - F_{conv}\left(G\left(X,s_2\right)\right)||_1\right] \quad (4)$$

where $s_1, s_2 \in S_{y_{trg}}$ are two randomly sampled style codes from domain $y_{trg} \in Y$ and $F_{conv}(\cdot)$ is the output of convolutional layers of F0 network F.

F0 consistency loss. An F0-consistent loss is added to produce F0-consistent results with the normalized F0 curve provided by F0 network F. For a given input mel-spectrogram X, the function F(X) calculates the absolute fundamental frequency (F0) value in Hertz for each frame within X. Given that male and female speakers tend to exhibit distinct average F0 values, a normalization step is employed to standardize the absolute F0 values captured by F(X) This normalization process is represented as $\widehat{F}(X) = \frac{F(X)}{||F(X)||_1}$.

Consequently, the F0 consistency loss is formulated as follows

$$L_{f0} = E_{X,s}\left[||\widehat{F}(X) - \widehat{F}\left(G\left(X,s\right)\right)||_1\right] \quad (5)$$

Speech consistency loss. Ensuring the linguistic fidelity of the converted speech is paramount, achieved through the implementation of a speech consistency loss mechanism. This mechanism relies on convolutional features extracted from a pre-trained joint Connectionist Temporal Classification (CTC) - attention model, particularly the VGG-Bidirectional Long Short-Term Memory (BLSTM) network, detailed in reference [26] and accessible within the Espnet toolkit [27]. Adhering to the approach of previous research [28], we leverage the output from the intermediate layer preceding the Long Short-Term Memory (LSTM) layers, denoted as $h_{asr}(\cdot)$, to encapsulate the linguistic feature. Consequently, the formal definition of the speech consistency loss is as follows

$$L_{asr} = E_{X,s}\left[||hasr\left(X\right) - hasr\left(G\left(X,s\right)\right)||_1\right] \quad (6)$$

Norm consistency loss. In order to maintain the temporal integrity of generated samples, we employ a norm consistency loss. This loss mechanism is designed to ensure the preservation of speech and silence intervals in the generated output. To calculate the absolute column-sum norm for a mel-spectrogram X, which comprises N mel frequency bins and T frames at the tth frame, we define it as $||X_{\cdot,t}|| = \sum_{n=1}^{N} ||X_{n,t}||_1$, where t $\in$ {1, … T} represents the frame index. The norm consistency loss can be expressed as follows

$$L_{norm} = E_{X,s}\left[\frac{1}{T}\sum_{1}^{T} |||X_{\cdot,t}|| - ||G(X,s)_{\cdot,t}|||\right] \quad (7)$$

Cycle consistency loss. Finally, we introduce the cycle consistency loss, as outlined in reference [17], with the purpose of preserving all remaining features present in the input data.

$$L_{cyc} = E_{X,y_{src},y_{trg},S}\left[||X - G\left(G\left(X,s\right),\widetilde{s}\right)||_1\right] \quad (8)$$

where $\widetilde{s} = S\left(X, y_{src}\right)$ is the estimated style code of the input in the source domain $y_{src} \in Y$.

Full objective. The entirety of our generator's objective functions can be condensed as follows:

$$\begin{aligned} &\min_{G,S,M} L_{adv} + \lambda_{advcls}L_{advcls} + \lambda_{sty}L_{sty} \\ &-\lambda_{ds}L_{ds} + \lambda_{f0}L_{f0} + \lambda_{asr}L_{asr} \\ &+\lambda_{norm}L_{norm} + \lambda_{cyc}L_{cyc} \end{aligned} \quad (9)$$

where λ_{advcls}, λ_{sty}, λ_{ds}, λ_{f0}, λ_{asr}, λ_{norm} *and* λ_{cyc} are hyperparameters for each term.

The complete objective for our discriminator is as follows

$$\min_{C,D} -L_{adv} + \lambda_{cls}L_{cls} \quad (10)$$

where λ_{cls} is the hyperparameter for source classifier loss L_{cls}, which is given by

$$L_{cls} = E_{X,y_{src},s}\left[CE\left(C\left(G\left(X,s\right),y_{src}\right)\right)\right] \quad (11)$$

The pretrained HiFi-GAN-BN model from Bahnaric language for the vocoder of Grad-TTS model.

Vocoders serve as instruments employed for transforming a speech spectrogram into audible sound waves. They play a pivotal role in the voice conversion process, facilitating the creation of sound corresponding to the given spectrogram. As outlined in Section 4.2, when it comes to the Grad-TTS system, employing a pre-trained HiFi-GAN designed for the English language poses several challenges due to the distinct linguistic and acoustic characteristics inherent in the Bahnar language as opposed to English. Consequently, we took the approach of retraining a pre-existing HiFi-GAN system tailored to Bahnar voice, following the methodology illustrated in Figure 8.

Within this training pipeline, there are three key components: one generator and two discriminators. The generator, designed as a fully convolutional neural network, takes a mel-spectrogram as its input and employs transposed convolutions to up-sample it until the resulting sequence matches the temporal resolution of raw waveforms.

In terms of the discriminators, they consist of two distinct modules. Firstly, the multi-period discriminator (MPD) is composed of several sub-discriminators, each responsible for assessing specific segments of periodic signals within the input audio. Furthermore, to capture consecutive patterns and long-term dependencies, we incorporate the multi-scale discriminator (MSD) concept, which is inspired by the approach introduced in MelGAN[29]. This MSD evaluates audio samples at various levels to gain a comprehensive understanding of the data.

The training process involves adversarial training for both the generator and discriminators. Additionally, two supplementary loss functions are employed to enhance training stability and overall model performance.

GAN loss. The training objectives of this model adhere to the principles of LSGAN[30]. Specifically, they replace the binary cross-entropy terms from the original GAN objectives[31] with least squares loss functions to ensure non-vanishing gradient flows. In this setup, the discriminator's training goal is to classify ground truth samples as 1 and generated samples from the generator as 0. Conversely, the generator aims to deceive the discriminator by adjusting the quality of its generated samples to be classified as a value very close to 1.

The GAN losses for both the generator G and the discriminator D are defined as

$$L_{adv}(D;G) = E_{X,s}\left[(D(X)-1)^2 + (D(G(s)))^2\right] \quad (12)$$

$$L_{adv}(G;D) = E_s\left[(D(G(s))-1)^2\right] \quad (13)$$

where X denotes the ground truth audio anddenotes the mel-spectrogram of the ground truth audio.

Mel-Spectrogram loss. To enhance the training performance of the generator and ensure the synthesized audio's fidelity, we introduce a mel-spectrogram loss into the GAN objective. This addition is made with the expectation that the input condition should also play a role in improving the perceptual quality, taking into consideration the characteristics of the human auditory system.

The mel-spectrogram loss is calculated as the L1 distance between the mel-spectrogram of a waveform generated by the generator and that of a ground truth waveform. It is defined as

$$L_{Mel}(G) = E_{X,s}\left[||\phi(X) - \phi(G(s))||_1\right] \quad (14)$$

where ϕ represents the transform function used to derive the mel-spectrogram from the corresponding waveform.

Feature matching loss. The model can also undergo optimization based on a metric that quantifies the distinction in features extracted by the discriminator when comparing a ground truth sample to a generated sample[32]. This metric, known as the feature matching loss, is defined as follows

$$L_{FM}(G;D) = E_{X,s}\left[\sum_{i=1}^{T}\frac{1}{N}||D^i(X) - D^i(G(s))||_1\right] \quad (15)$$

Full objective. The ultimate loss functions for both the generator and discriminator are defined as

$$\min_{G,D} L_G = L_{adv}(G,D) + \lambda_{fm}L_{FM}(G,D) + \lambda_{mel}L_{Mel}(G) \quad (16)$$

$$\min_{G,D} L_D = L_{adv}(D,G) \quad (17)$$

EXPERIMENT RESULTS

There are two main models trained from scratch in this system including the StarGANv2-VC model for voice conversion and the HiFi-GAN for the vocoder of the Grad-TTS model. Both two these models are developed based on the Pytorch framework. Considering the StarGANv2-VC model, it is trained with 122 epochs using the GPU of NVIDIA RTX 3080. The dataset that we use to train this model is the recorded

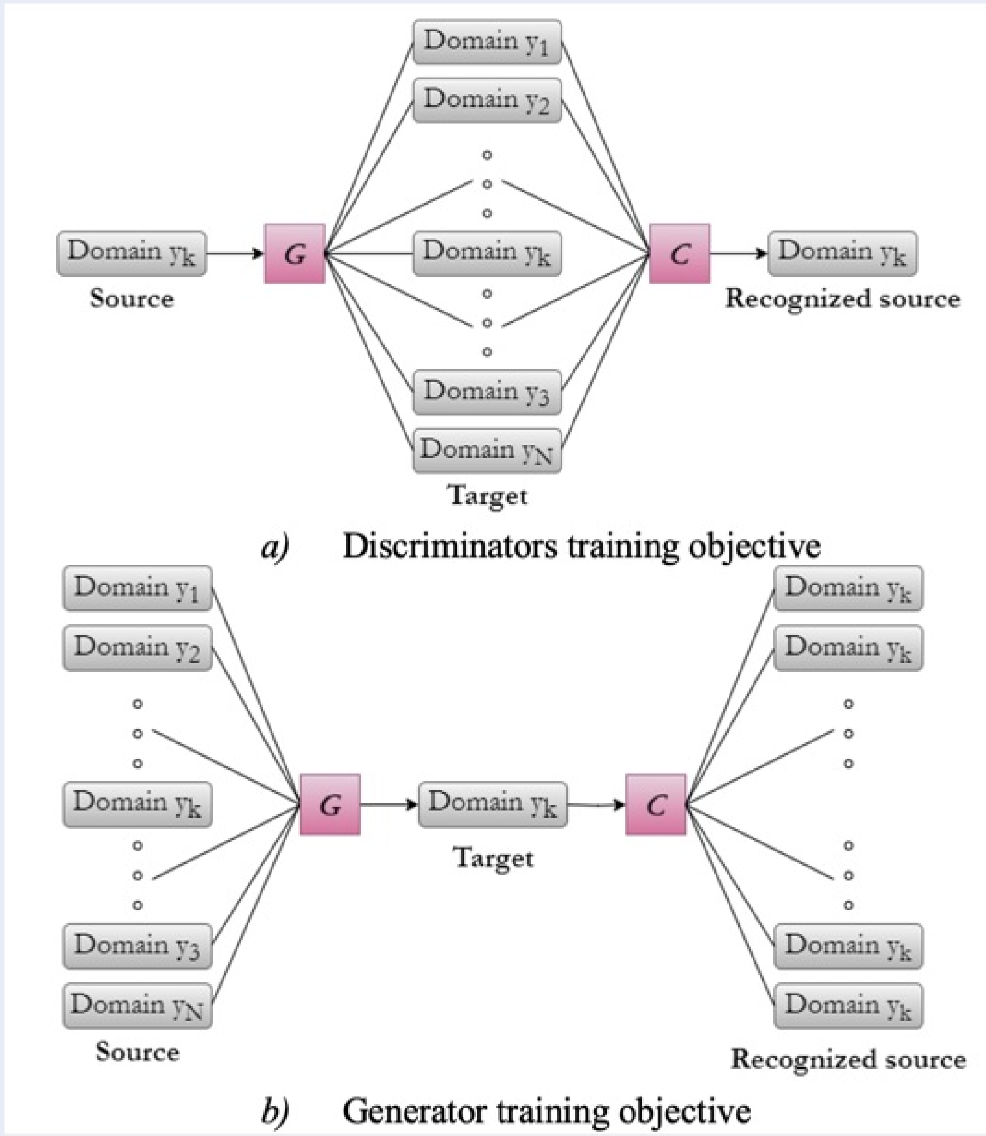

Figure 7: Training schemes of adversarial source classifier for a domain y_k.

voices gained manually by native Bahnaric from the provinces of Gia Lai, Kon Tum, and Binh Dinh in Vietnam, where exist considerable communities of Bahnaric people. They are used as training input for audio files that are generated from Grad-TTS. On the other hand, the HiFi-GAN model is trained up to 1 million steps with two A100 GPUs. In other to train this model, we collected the from the YouTube channel of VTV5, which consists of 300 hours of Bahnaric speech.

Regarding the evaluation methodology, we built the web application as shown in Figure 9. A user-friendly web-based interface was developed using Streamlit to facilitate the evaluation process. This interface presented users with 20 questions, each representing a unique evaluation instance. Each evaluation instance consisted of the following components:

Original Speech Audio: The interface played an original speech audio recording from a human speaker. This audio served as a reference point for users to compare the converted audio against.

Converted Speech Audios: Two converted speech audios were played for each evaluation instance. These audios were generated using our two best-performing StarGANv2-VC models. The intention here was to compare the quality of voice conversion between the models.

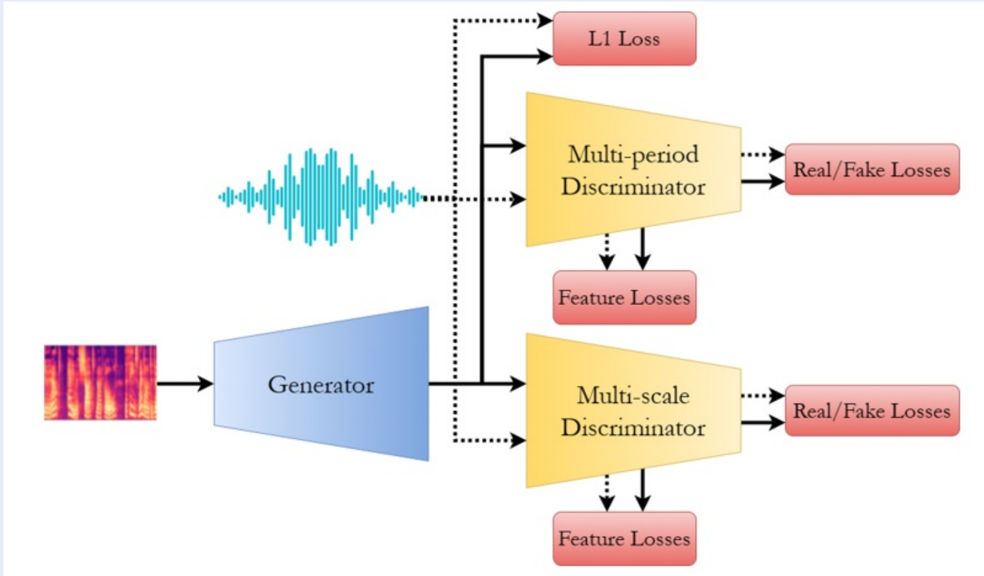

Figure 8: The training pipeline for HiFi-GAN from Bahnaric language.

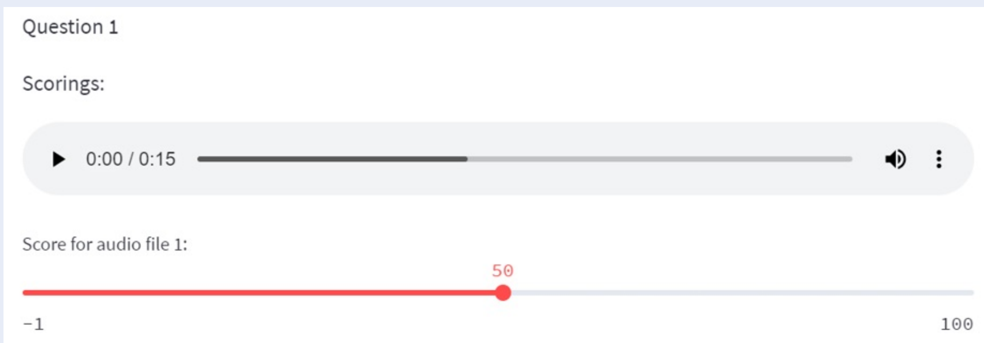

Figure 9: A web application for system evaluation.

With respect to the scoring mechanism, users were given a scoring scale ranging from -1 to 100 to rate the quality of the converted audio. This scoring scale was designed to capture a broad spectrum of quality perceptions. The interpretation of the scale was as follows:

- ***-1***: *Unrealistic Sound*. The converted audio needs to be more realistic and unconvincing to represent the target speaker.
- ***0-49***: *Poor to Fair*. The converted audio is poor to fair quality, with significant discrepancies from the original speaker's voice.
- ***50-69***: *Moderate*. The converted audio resembles the target speaker's voice, but improvements are needed.
- **70-89**: *Good*. The converted audio is of good quality and reasonably captures the target speaker's characteristics.
- **90-99**: *Very Good*. The converted audio is of outstanding quality, closely resembling the target speaker's voice with minor discrepancies.
- **100:** *Perfect*. The converted audio is indistinguishable from the audio of the actual target speaker; no improvements are necessary.

The scale ranging from -1 to 100 (comprising 6 levels) has been designed with specific intentions. At the lower end, -1 is assigned to instances where the AI-generated sound is exceptionally poor, to the extent that it is practically unbearable. Conversely, at the upper end, 100 signifies that within the provided audio, at least one sentence closely resembles an original human-generated recording. Essentially, the scale is employed to convey to the evaluator that there can be a wide range of sound quality, spanning from severely subpar to human-level excellence. The evaluation result is collected from 46 voluntary participants, whose statistics are shown in Table 1.

As shown in Table 1, there is no evaluation result of bad quality in the samples of the original voice that recorded by native speakers. Concerning voice conversion models, the VC-original model is trained from original voice data and the VC-Grad-TTS is trained with a suitable amount of data in the source domain that is taken from the output of Grad-TTS.

It can be seen that the VC-original model generates sounds with acceptable quality. However, there is an existence of bad quality samples and it accounts for 4.24% of the evaluation set. The number of samples having very good quality is also quite low at 11.96%. Overall, the voice converted by this model is evaluated as having good quality with a mean score of 74.07.

On the other hand, the VC-Grad-TTS model gives better performance. The number of samples that have poor to fair quality is reduced significantly (accounting for 0.87%). In addition, most generated sample from this model is evaluated from good to perfect. The mean evaluation score is also high with 80.33, which belongs to the scale of good quality sound.

DICUSSION

This research addresses the challenge of generating natural-sounding speech in the Bahnaric language, which is often marginalized and lacks adequate resources. Our system shows promising results in synthesizing Bahnaric speech. Table 1 illustrates that models trained with Grad-TTS output as the domain source outperform those trained directly with native speaker data, with synthesized voice quality also rated as good. Moreover, the HiFi-GAN-BN model, pre-trained with Bahnaric voice data, enhances the authenticity of synthesized speech to resemble Bahnaric accents when converting mel-spectrogram output. On the other hand, further optimization and evaluation across diverse linguistic and cultural contexts are necessary. Collaboration with linguists and community stakeholders is vital to ensure the cultural relevance and acceptance of synthesized Bahnaric voices. Ultimately, our work contributes to the preservation and promotion of cultural diversity and linguistic heritage, not only within the Bahnaric community in Vietnam but also in similar contexts worldwide.

CONCLUSION

The Vietnamese government is endeavoring to enhance their integration through advancements in socio-cultural and scientific literacy. In order to contribute to conveying information with native-like Bahnaric speech, we have proposed an effective approach called BN-TTS-VC system. Most of the text-to-speech systems require a substantial amount of training data. It is particularly arduous to gather a high-quality training dataset of extremely low-resource languages like Bahnaric. Therefore, our system combined Grad-TTS model [3] and the StarGANv2-VC model [4] to solve this problem. In addition, we also introduce the HiFi-GAN-BN model, a variant of HiFi-GAN [5] pre-trained by Bahnaric voice, to resemble the Bannaric accents better when transforming the mel-spectrogram output of StarGANv2-VC into human-listenable waveform. The evaluation results have shown that the system is able to generate good-quality audio and the voice conversion model that is trained with the source domain data taken from the output of Grad-TTS gives better performance. Future work includes improving the quality of sound that is not clear or missing the vocabulary of the text.

ACKNOWLEDGMENT

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

LIST OF ABBREVIATIONS

TTS: Text-to-speech

VC: Voice conversion

Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech

StarGANv2-VC: A Diverse, Unsupervised, Non-parallel Framework for Natural-Sounding Voice Conversion.

HiFi-GAN: A GAN-based model capable of generating high fidelity speech efficiently.

BN-TTS-VC: The combined system of text-to-speech and voice conversion for Bahnaric language.

HiFi-GAN-BN: A GAN-based model from Bahnaric language for the vocoder of Grad-TTS model.

CONFLICTS OF INTEREST

All authors declare that they have no conflicts of interest.

Table 1: The evaluation result of StarGANv2-VC models.

Type of sample	Quality (%)						Mean score
	-1 ↓	0-49 ↓	50-69 ↑	70-89 ↑	90-99 ↑	100 ↑	
Original	0.0	0.0	2.06	56.31	39.02	2.61	87.12
VC-original	0.0	4.24	30.22	52.39	11.96	1.19	74.07
VC-Grad-TTS	0.0	0.87	18.26	55.54	23.59	1.74	80.33

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Dang Tran Dat: Methodology, Model development, Evaluation, Writing – Original Draft.
Tang Quoc Thai: Methodology, Model development, Evaluation, Writing.
Nguyen Quang Duc: Methodology, System Deployment, Resources, Data Collection, Data Curation, Writing.
Vo Duy Hung: Methodology.
Quan Thanh Tho: Supervision, Project Administration, Methodology, Writing - Review & Editing.

REFERENCES

1. Tan X, Chen J, Liu H, Cong J, Zhang C, Liu Y, Wang X, Leng Y, Yi Y, He L, et al. Naturalspeech: End-to-end text to speech synthesis with human-level quality. arXiv preprint arXiv:2205.04421; 2022;.
2. Sisman B, Yamagishi J, King S, Li H. An overview of voice conversion and its challenges: From statistical modeling to deep learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020;29:132-157;Available from: https://doi.org/10.1109/TASLP.2020.3038524.
3. Popov V, Vovk I, Gogoryan V, Sadekova T, Kudinov M. Grad-tts: A diffusion probabilistic model for text-to-speech. In: International Conference on Machine Learning; 2021;.
4. Choi Y, Uh Y, Yoo J, Ha J-W. Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020;Available from: https://doi.org/10.1109/CVPR42600.2020.00821.
5. Kong J, Kim J, Bae J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In: Advances in Neural Information Processing Systems. 2020;33:17022-17033;.
6. Scully C. Articulatory synthesis. In: Speech production and speech modelling. Springer; 1990. p. 151-186;Available from: https://doi.org/10.1007/978-94-009-2037-8_7.
7. Lukose S, Upadhya SS. Text to speech synthesizer-formant synthesis. In: 2017 International Conference on Nascent Technologies in Engineering (ICNTE); 2017;PMID: 29031741. Available from: https://doi.org/10.1109/ICNTE.2017.7947945.
8. Mao X, Li Q, Xie H, Lau RYK, Wang Z, Smolley SP. Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision; 2017;Available from: https://doi.org/10.1109/ICCV.2017.304.
9. Kumar K, Kumar R, De Boissiere T, Gestin L, Teoh WZ, Sotelo J, De Brebisson A, Bengio Y, Courville AC. Melgan: Generative adversarial networks for conditional waveform synthesis. In: Advances in neural information processing systems. 2019;32;.
10. Park J, Zhao K, Peng K, Ping W. Multi-speaker end-to-end speech synthesis. arXiv preprint arXiv:1907.04462; 2019;.
11. Polyak A, Wolf L, Adi Y, Taigman Y. Unsupervised cross-domain singing voice conversion. arXiv preprint arXiv:2008.02830; 2020;Available from: https://doi.org/10.21437/Interspeech.2020-1862.
12. Watanabe S, Hori T, Karita S, Hayashi T, Nishitoba J, Unno Y, Soplin NEY, Heymann J, Wiesner M, Chen N, et al. Espnet: End-to-end speech processing toolkit. arXiv preprint arXiv:1804.00015; 2018;PMID: 29730221. Available from: https://doi.org/10.21437/Interspeech.2018-1456.
13. Kim S, Hori T, Watanabe S. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP); 2017;Available from: https://doi.org/10.1109/ICASSP.2017.7953075.
14. Kum S, Nam J. Joint detection and classification of singing voice melody using convolutional recurrent neural networks. Applied Sciences. 2019;9:1324;Available from: https://doi.org/10.3390/app9071324.
15. Tran T, Nguyen T, Bui H, Nguyen K, Vo NG, Pham TV, Quan T. Naturalness Improvement of Vietnamese Text-to-Speech System Using Diffusion Probabilistic Modelling and Unsupervised Data Enrichment. In: International Conference on Intelligence of Things; 2022;Available from: https://doi.org/10.1007/978-3-031-15063-0_36.
16. Huang W-C, Hayashi T, Wu Y-C, Kameoka H, Toda T. Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining. arXiv preprint arXiv:1912.06813; 2019;Available from: https://doi.org/10.21437/Interspeech.2020-1066.
17. Tanaka K, Kameoka H, Kaneko T, Hojo N. AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019;Available from: https://doi.org/10.1109/ICASSP.2019.8683282.
18. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision; 2017;Available from: https://doi.org/10.1109/ICCV.2017.244.
19. Park S-w, Kim D-y, Joe M-c. Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data. arXiv preprint arXiv:2005.03295; 2020;Available from: https://doi.org/10.21437/Interspeech.2020-1542.
20. Kaneko T, Kameoka H, Tanaka K, Hojo N. Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion. arXiv preprint arXiv:2010.11672; 2020;Available from: https://doi.org/10.21437/Interspeech.2020-2280.
21. Huang W-C, Luo H, Hwang H-T, Lo C-C, Peng Y-H, Tsao Y, Wang H-M. Unsupervised representation disentanglement using cross domain features and adversarial learning in variational autoencoder based voice conversion. IEEE Transactions on Emerging Topics in Computational Intelligence. 2020;4:468-479;Available from: https://doi.org/10.1109/TETCI.2020.2977678.
22. Ding S, Gutierrez-Osuna R. Group Latent Embedding for Vector Quantized Variational Autoencoder in Non-Parallel Voice Conversion. In: Interspeech; 2019;PMID: 31791587. Available from: https://doi.org/10.21437/Interspeech.2019-1198.
23. Qian K, Zhang Y, Chang S, Yang X, Hasegawa-Johnson M. Au-

tovc: Zero-shot voice style transfer with only autoencoder loss. In: International Conference on Machine Learning; 2019;.

24. Prenger R, Valle R, Catanzaro B. Waveglow: A flow-based generative network for speech synthesis. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019;Available from: https://doi.org/10.1109/ICASSP.2019.8683143.
25. Lam QT, Do DH, Vo TH, Nguyen DD. Alternative vietnamese speech synthesis system with phoneme structure. In: 2019 19th International Symposium on Communications and Information Technologies (ISCIT); 2019;Available from: https://doi.org/10.1109/ISCIT.2019.8905142.
26. Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, Chen Z, Zhang Y, Wang Y, Skerrv-Ryan R, 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;Available from: https://doi.org/10.1109/ICASSP.2018.8461368.
27. Donahue J, Dieleman S, Bińkowski M, Elsen E, Simonyan K. End-to-end adversarial text-to-speech. arXiv preprint arXiv:2006.03575; 2020;.
28. Ren Y, Hu C, Tan X, Qin T, Zhao S, Zhao Z, Liu T-Y. Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558; 2020;.
29. Zen H, Tokuda K, Black AW. Statistical parametric speech synthesis. Speech Communication. 2009;51:1039-1064;Available from: https://doi.org/10.1016/j.specom.2009.04.004.
30. Schwarz D. Corpus-based concatenative synthesis. IEEE signal processing magazine. 2007;24:92-104;Available from: https://doi.org/10.1109/MSP.2007.323274.
31. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in neural information processing systems. 2014;27;.
32. Larsen ABL, Sønderby SK, Larochelle H, Winther O. Autoencoding beyond pixels using a learned similarity metric. In: International conference on machine learning; 2016;.

Tạp chí Phát triển Khoa học và Công nghệ – Engineering and Technology 2024, ():1-1

Open Access Full Text Article

Bài nghiên cứu

Phương pháp thay đổi giọng tăng cường tính tự nhiên cho quá trình sinh giọng nói ở ngôn ngữ ít tài nguyên: Thí nghiệm với ngôn ngữ Ba Na

Đặng Trần Đạt[1,2], Tăng Quốc Thái[1,2], Nguyễn Quang Đức[1,2], Võ Duy Hùng[1,2], Quản Thành Thơ[1,2,*]

TÓM TẮT

Ba Na là một nhóm dân tộc thiểu số ở Việt Nam, được chính phủ ưu tiên bảo tồn di sản văn hóa, truyền thống và ngôn ngữ. Trong kỷ nguyên của công nghệ AI hiện nay, việc tổng hợp giọng nói tiếng Ba Na để hỗ trợ những nỗ lực bảo tồn này chứa đựng tiềm năng đáng kể. Mặc dù công nghệ chuyển đổi giọng nói đã có những bước tiến trong việc nâng cao chất lượng và tính tự nhiên của giọng nói được tổng hợp nhưng nó chỉ được chú trọng phát triển chủ yếu đối với các ngôn ngữ được sử dụng rộng rãi. Do đó, các ngôn ngữ có nguồn tài nguyên hạn chế như ngôn ngữ thuộc họ tiếng Ba Na gặp nhiều khó khăn trong việc tổng hợp giọng nói. Nghiên cứu này giải quyết thách thức lớn trong việc tổng hợp giọng nói có tính tự nhiên ở các ngôn ngữ có nguồn tài nguyên thấp bằng cách khám phá các ứng dụng của kỹ thuật chuyển đổi giọng nói cho tiếng Ba Na. Chúng tôi giới thiệu hệ thống BN-TTS-VC, một phương pháp tiên phong tích hợp hệ thống chuyển văn bản thành giọng nói dựa trên Grad-TTS, với các kỹ thuật chuyển đổi giọng nói dựa trên StarGANv2-VC, và cả hai đều được thiết kế riêng cho các sắc thái của tiếng Ba Na. Grad-TTS cho phép hệ thống phát âm các từ trong ngôn ngữ Ba Na mà không bị giới hạn từ vựng, trong khi StarGANv2-VC nâng cao tính tự nhiên của giọng nói được tổng hợp, đặc biệt là trong bối cảnh các ngôn ngữ có nguồn tài nguyên thấp như tiếng Ba Na. Ngoài ra, chúng tôi còn giới thiệu mô hình HiFi-GAN được tinh chỉnh bằng tiếng Ba Na để nâng cao chất lượng giọng nói so với giọng bản địa, đảm bảo thể hiện giọng nói tiếng Ba Na chân thực hơn. Để đánh giá hiệu quả của phương pháp tiếp cận, chúng tôi đã tiến hành thử nghiệm dựa trên đánh giá của con người từ các tình nguyện viên. Các kết quả sơ bộ đầy hứa hẹn, cho thấy phương pháp của chúng tôi chứa nhiều tiềm năng trong việc tổng hợp giọng nói mang tính tự nhiên tiếng Ba Na. Qua nghiên cứu này, mục tiêu của chúng tôi là đóng góp vào các nỗ lực để bảo tồn và thúc đẩy di sản ngôn ngữ và văn hóa của nhóm dân tộc thiểu số Bahnar. Bằng cách tận dụng sức mạnh của công nghệ AI, chúng tôi mong muốn thu hẹp khoảng cách trong tổng hợp giọng nói cho các ngôn ngữ nguồn tài nguyên thấp và tạo điều kiện thuận lợi cho việc bảo tồn di sản văn hóa quý báu của họ.

Từ khoá: Tổng hợp giọng nói tiếng Ba Na, chuyển văn bản thành giọng nói, chuyển đổi giọng nói tự nhiên

[1]*Khoa Khoa học và Kỹ thuật Máy tính, Trường Đại học Bách khoa – ĐHQG-HCM, Việt Nam*

[2]*Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam*

Liên hệ

Quản Thành Thơ, Khoa Khoa học và Kỹ thuật Máy tính, Trường Đại học Bách khoa – ĐHQG-HCM, Việt Nam

Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam

Email: qttho@hcmut.edu.vn

Lịch sử

- Ngày nhận: 08-9-2023
- Ngày chấp nhận: 27-3-2024
- Ngày đăng:

DOI :

Trích dẫn bài báo này: Đạt D T, Thái T Q, Đức N Q, Hùng V D, Thơ Q T. **Phương pháp thay đổi giọng tăng cường tính tự nhiên cho quá trình sinh giọng nói ở ngôn ngữ ít tài nguyên: Thí nghiệm với ngôn ngữ Ba Na** . *Sci. Tech. Dev. J. - Eng. Tech.* 2024; ():1-1.