BaNaVA: A cross-platform AI mobile application for preserving the Bahnaric languages

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

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

Khoa Tran Ngoc Dang
Ho Chi Minh City University of Technology (HCMUT), Vietnam
Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam
khoa.tranngocdang03@hcmut.edu.vn

Minh Tran Duy
Ho Chi Minh City University of Technology (HCMUT), Vietnam
Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam
minh.tranduy2209@hcmut.edu.vn

***Abstract*—AI-powered translation is a promising solution to the language barrier faced by the Bahnar people. However, developing low-resource text-to-speech translation systems is challenging. The authors propose a mobile application called BaNaVA to address these challenges. BaNaVA uses a combination of natural machine translation and linguistic analysis to translate between Vietnamese and Bahnaric with high accuracy, while also using a voice conversion system to convert the voice quality to match that of a genuine Bahnar individual. They incur as two connected and important services residing inside the logical framework of the application. BaNaVA is designed using microservices and React Native software framework, which allows the application to be developed cross-platform. This mobile application utilizes specific neural machine translation (NMT) and text-to-speech (TTS) technologies to efficiently operate within the edge computing environment of popular mobile devices.**

***Index Terms*—natural machine translation, low-resource language, mobile application, text-to-speech, voice conversion**

I. Introduction

The community of Bahnaric ethnic minority, pronounced as [ɓaːnaː] in Vietnamese dialect, constitutes a significant proportion of rich customs and heritage to the breadth of Vietnamese culture. Their spoken and written Bahnaric language belongs to the Mon-Khmer branch in the Central Highlands, with a current population figure estimated at over 280,000. Despite the egalitarian community spirit, the Bahnar folks face significant challenges in accessing information and engaging with mainstream society due to the language barrier. However, recent advancements in artificial intelligence (AI), particularly natural machine translation (NMT), offer a glimmer of hope.

AI-powered translation offers a promising solution to break the language barrier faced by the Bahnar people. NMT systems have made remarkable advancements in recent years, achieving unprecedented levels of accuracy and fluency. This has made it possible to translate a wide range of document types, including government documents, educational materials, and news articles, from general Vietnamese to native Bahnaric dialect. As a result, the continued efforts are shown to be beneficial for the community themselves. Firstly, it can enhance their access to information and essential services. For example, Bahnar people can utilize readily available translation systems to access government websites, read news articles, and translate educational materials with minimal assistance. Secondly, AI-powered translation can help to break down language barriers and promote inclusivity. It allows Bahnar people to engage with mainstream society on an equal basis and participate in civic affairs. Finally, AI-powered translation helps to preserve the Bahnaric language and culture of the people in the digital medium, making Bahnaric contents more accessible in the palm of their hand. In turns, these translation tools encourage individuals to use their languages more frequently and transcend it to future generations.

Besides, contemporary modus operandi for AI-powered translation relies on text-to-speech (TTS) systems to perform language interpretation at the expense of large datasets. For rather low-resource languages like the Bahnaric family, collecting such pool of data is a formidable task without certain diminishing factors that affect quality and accuracy. This

results in poorly-pronouncing TTS systems that hinder general communication. In addition, since the Bahnaric people and other communities live in remote areas where it is difficult to deploy modern information systems, mobile applications are undoubtedly the most effective modes of exchanging information. However, since NMT and TTS techniques often require high computing resources with large-scale data processing infrastructure, we need to find appropriate approaches to develop mobile applications based on lightweight AI models that can translate and synthesize speech effectively for the Bahnaric people. To address these issues, we introduce BaNaVA, a cross-platform AI-powered mobile application that aims to strengthen and sustain the Bahnaric languages with convenience of communication in mind. On a basic level, the application offers the following features for well-delivered translation services:

- BaNaVA uses a translation mechanism that combines the strengths of NMT and the analysis of linguistic similarities between Vietnamese and Bahnaric. This allows fluent conversion between the two dialects with high accuracy while maintaining coherence and cohesion in a small-scale infrastructure. This makes it possible to deploy on a mobile environment where resources are scarce.
- BaNaVA implements a voice conversion system that not only performs TTS but also converts 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 and the StarGANv2-VC model to achieve this objective.
- BaNaVA is designed using micro-services and React Native software framework, which allows the application to be developed cross-platform. It also enables fluid and responsive interaction for users with minimum stuttering during the utilization of translation models. Currently, the application is available on two platforms, Android and iOS, to provide users hands-on experience of our services.

II. Related Works

A. Machine translation system on relatively-sized models

Early machine translation models followed a standard structure with two main components: an encoder to process input data and a decoder to generate translated sentences. These models used either two RNNs or two LSTMs, a more sophisticated type of RNN. These predecessor designs allowed for a seamless process of reading, understanding, and translating texts that are systematically nuanced However, RNN models encountered challenges in retaining information from extensive input sequences. The advent of attention mechanisms like Bahdanau Attention [3] and Luong Attention [4] established a foundation for subsequent research, not only in machine translation but also in diverse fields such as question-answering systems, information retrieval, and information extraction.

To address the challenge of limited data availability, some researchers have explored the use of multilingual models, some coupled with encoders and decoders for specific language pairs [9]. For example, the BARTpho model has pre-trained weights specifically optimized for Vietnamese vocabularies [10].

B. Text-to-speech (TTS) system

Text-to-speech synthesis involves the transformation of written text into spoken words with the aim of creating artificial speech that closely mirrors human speech. Recently, booster models to the current TTS system such as those incorporating Tacotron 2 model and the WaveGlow neural encoder was announced [12]. In 2020, Tacotron 2 model and the Hifi-gan vo-coder was the latest pairing attempt to speech synthesis [13]. The highest standard TTS, the Grad-TTS model [14], addresses the limitations of previous models such as the need for large data sets, unnatural sound, lack of intonation, and difficulty in processing long passages. Currently, it has achieved positive results in English. Because Bahnaric dialects does not have the same intonation as English and has a simpler syllable structure, it is expected to produce positive results in synthesizing Bahnaric speeches that are of low-resource.

C. Voice conversion techniques

Voice conversion (VC) refers to a method used to change the identity of one speaker's voice into another while retaining the linguistic content. The first approach broadcasts the use of an auto-encoder [15] to eliminate speaker-dependent information. However, the converted speech quality relies on the extent to which linguistic information can be extracted from the latent space. To alliviate this issue, GAN-based methods like CycleGAN-VC3 [16] employ a discriminator to guide the decoder in generating speech resembling the target speaker. Conversely, the last category in the batch takes advantages of TTS systems to extract aligned linguistic features from the various unconditioned input speech. Most notably, Cotatron [17], AttS2S-VC [18], and VTN [19] models impart the converted speaker's identity to closely resemble that of the target speaker.

III. Bahnaric Phonological System

For any speech synthesis paradigm, understanding the phonological system of the target language is essential. The passage in Figure 1 depicts an example of Bahnaric language. Because the language has unique sets of characters and stress symbols, input parsing modules from other languages cannot be used. Therefore, we conducted a detailed analysis of the language and developed a set of pseudo-phonemes specifically for Bahnaric expressions.

```
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
```

Fig. 1. Sample of a newsletters paragraph written in Bahnar language

An alphabet standardizes sets of written symbols and letters that represent the sounds of a language. For the Bahnaric language, its alphabet consists of four categories: monophonic

characters, diphthong vowels, double consonants and triple consonants, which are all illustrated in Figures III(a), III(b), III(c), and III(d) respectively. In addition to the basic consonants and vowels in common alphabets, many letters in the language have diacritics to indicate stress. Diphthongs are also common, where two vowels are paired together to form long sounds. On the other hand, several consonants can be combined to form consonant digraphs or trigraphs, which are unvoiced part of speech.

```
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 ẩ ẫ ậ ắ ằ ẳ ẵ ặ ẻ é è ẽ ẹ ỉ ị ọ ố ồ ổ ỗ ộ ớ ờ ở ỡ ợ ụ ủ ứ ừ ữ ự
```

(a) Monophonic characters in Bahnar language

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

(b) Diphthong vowels using Bahnaric phonemes

```
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 using Bahnaric phonemes

```
hng ngl nhr
```

(d) Triple consonants using Bahnaric phonemes

Fig. 2. Different examples in using Bahnaric phonemes.

Speech synthesis for two languages works by mapping each word in the input text to a corresponding phoneme sequence based on the alphabets of both languages. The analysis module in Figure 3 processes the input text and produces a phoneme sequence, which is then used to train the Grad-TTS in the BaNaVA architecture in order to output a meaningful sequence.

Input sequence:	Inh kăt 'ba kơpưng adơi,
Processed sequence:	I-nh-k-ă-t-'b-a-k-ơ-p-ư-ng-a-d-ơ-i,
Vietnamese equivalence:	Tôi cắt lúa trên đồng.

Fig. 3. An example of Bahnaric speech synthesis procedure

IV. BaNaVA system architecture

We introduce the foundation for BaNaVA mobile application in Figure 4. Its architectural design is a three-layered approach that ensures low-resource deployment, tight application integration, and cross-platform compatibility.

The business layer is the core language synthesis component of BaNaVA. It consists of two main services: one that translates Vietnamese text into Bahnaric dialects using a combination of linguistic engineering techniques, and another that converts the output into synthesized speech. Two checkpoints are marked in black in the pipeline, corresponding to halting points where the process can stop based on the options selected in the presentation layer.

The bottom layer of the mobile application caches data for use in the synthesis model, including chunks of Vietnamese words fragmented to map to Bahnaric vocabulary, and fragments of voice recordings or constituents downloaded from the server hosting all the speeches of various dialects. Lastly, a local model acts as a gateway to communicate with the service pipeline, sending and receiving requests to and from the language hosting server.

V. Vietnamese-Bahnaric translation service

For an overview, Figure 5 illustrates the comprehensive translation process from Vietnamese to Bahnaric.

Bahnaric languages are classified as low-resource due their limited linguistic data for natural language processing tasks. This makes conventional machine translation methods impractical and inefficient, as they rely on large parallel datasets for training. However, Vietnamese and Bahnaric languages share some structural features in their grammar, such as word order, morphology, and syntax. Leveraging these similarities can facilitate the machine translation process and improve the quality of the output. Our proposed chunking translation approach combines word mapping and fine-tuning of BNBARTpho, which is specifically for machine translation.

The chunking translation approach is less computationally intricate compared to traditional methods that necessitate complete sentence alignment and parsing. The proposed method encompasses an end-to-end pipeline with two primary phases:

- **Segmentation Phase**: This initial stage of the machine translation process takes a Vietnamese sentence as input and uses methods such as word segmentation and named entity recognition (NER) to identify anchors and chunks within the sentence. Anchors are words or phrases that can be directly translated to Bahnaric, while chunks are words or phrases that require further processing. The output of this stage is a list of anchors and chunks that collectively form the input sentence.
- **Mapping Phase**: In this subsequent phase, the list of anchors and chunks serves as input and undergoes various techniques for their translation into Bahnaric. For anchors, the specific mapping method employed depends on whether they are words listed in the dictionary or entities. For dictionary words, a word mapping method is utilized, referencing a bilingual dictionary to find the

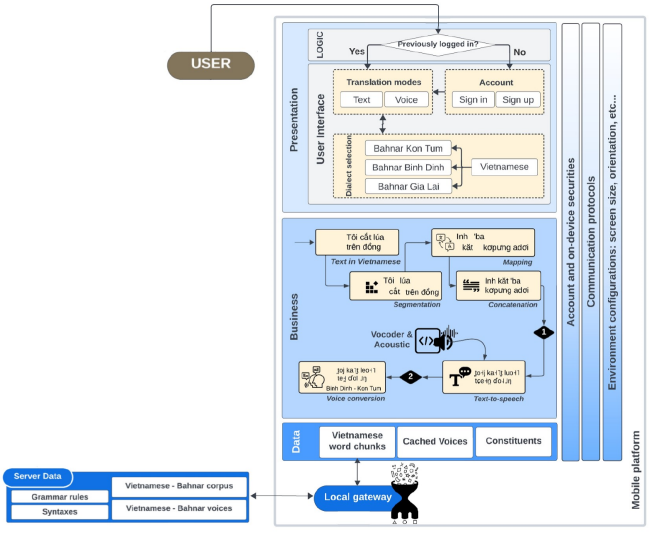


Fig. 4. Complete mobile infrastructure design for BaNaVA

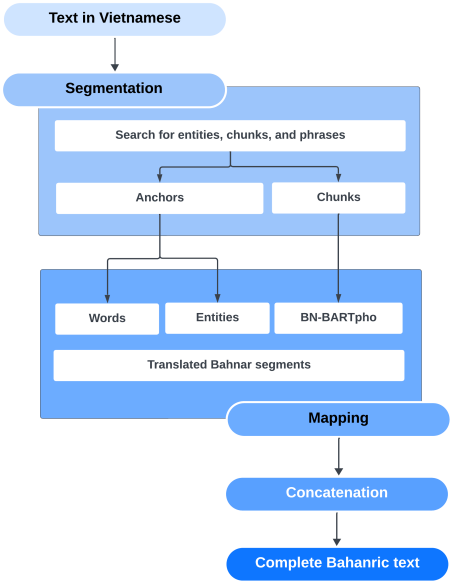


Fig. 5. Proposed pipeline for Bahnaric translation service

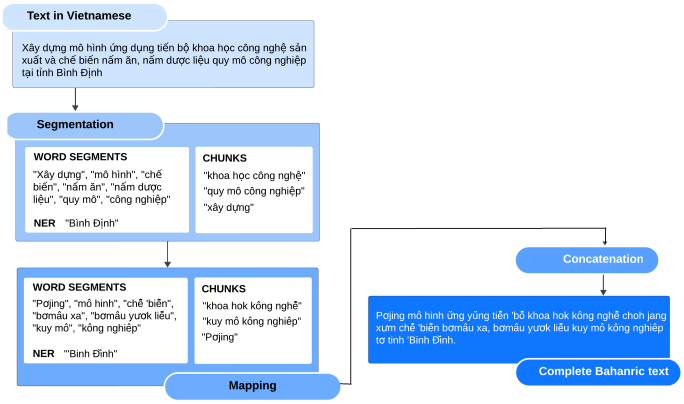


Fig. 6. An execution example for the Bahnaric translation service

corresponding Bahnaric word. Entities, on the other hand, undergo an entity mapping method employing phonetic rules to convert the Vietnamese entity into its Bahnaric counterpart. Chunks, on the other hand, are translated into Bahnaric phrases using a fine-tuned BARTpho model [9]. The output of this phase is a list of translated segments derived from both anchors and chunks.

After two phases, the list of translated segments is concatenated together to form a complete sentence. Figure 6 depicts an example run of the service from a Vietnamese sentence to a Bahnaric equivalence based on our proposed pipeline.

VI. Bahnaric Voice Conversion Service

This essential secondary service is composed of two primary modules: Text-to-Speech and Voice Conversion, which are depicted in Figure 7. The Bahnaric text-to-speech module generates a natural-sounding voice from the input text using a vocoder and an acoustic model. The vocoder converts the acoustic properties generated by the acoustic model into sound

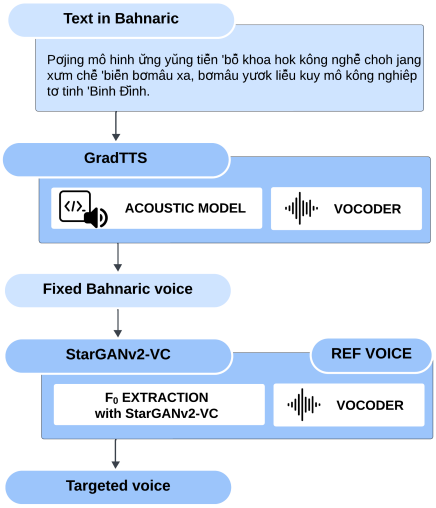


Fig. 7. Proposed pipeline for Bahnaric voice conversion service

waveforms. The acoustic model takes into account the phonetic context of each phoneme in the input text to generate accurate acoustic properties.

Following this, the resulting sound waveforms are directed to the Voice Conversion module to generate alternate types of native voices based on a reference voice. This module is constructed from three main component models designed to extract voice characteristics, transform the voice, and convert the mel-spectrogram into a human-audible waveform. Our text-to-speech system comprises three principal elements, outlined in Figure 8. Initially, the Text Analysis module dissects the text into a pseudo-phonetic representation tailored for neural network processing.

The second module involves an acoustic model rooted in Grad-TTS. Given a set of pseudo-phonemes as input, this model undergoes a training process to generate the mel-spectrogram representation. A mel-spectrogram constitutes a spectrum representation of sound waves, featuring two dimensions: frequency and time. These representations offer detailed insights into prevalent frequency bands at each moment within the sound wave. Both the extraction of mel-spectrograms from sound waves and the conversion of sound waves from mel-spectrograms are feasible through the inverse problem. For specific technical architectural details, these can be referred to in prior related works; in this publication, we refrain from delving into intricate technical specifics to ensure accessibility to a broader audience.

The final stage is executed by the vocoder, employing the HiFi-GAN network [24] to convert the output from the mel-spectrogram into a waveform. Notably, instead of utilizing a pre-trained HiFi-GAN designed for the English language, we retrained a pre-existing HiFi-GAN-BN system specifically adapted for Bahnaric to produce the ultimate voice output.

The Grad-TTS model has the capacity to articulate Bahnar vocabulary without limitations. However, due to the limited linguistic resources available for this language, the sound

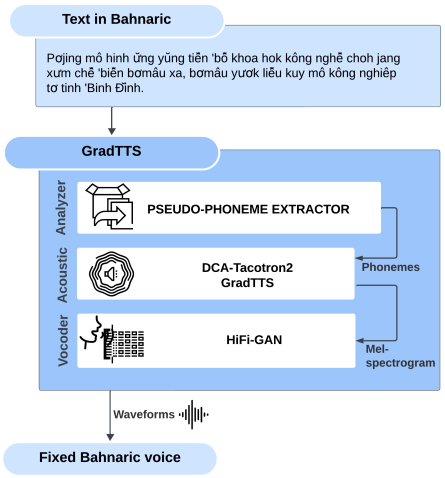


Fig. 8. Expanded processing pipeline for GradTTS module

quality still lacks the natural nuances of human speech. To address this issue, we propose implementing the StarGANv2-VC model to transform the voice synthesized by Grad-TTS into a sample voice that resembles a native Bahnar voice.

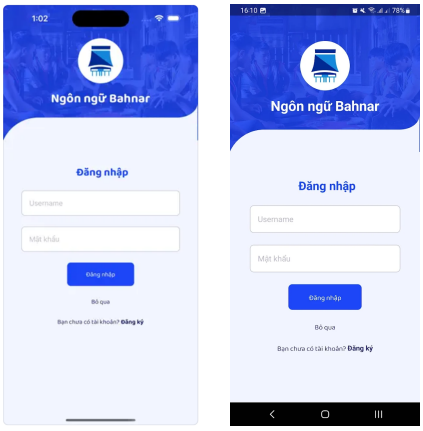
VII. Experimental Results

A. The cross-platform AI mobile BaNaVA appliction

We have developed a cross-platform AI as discussed above. This app is available on both iOS and Android environment. Interested users can download app at:

- **iOS App:** https://apps.apple.com/vn/app/d%E1%BB%8Bch-thu%E1%BA%ADt-ba-na/id6462193072?l=vi
- **Android App:** https://play.google.com/store/apps/details?id=com.hcmut.bahnar&pcampaignid=web_share.

Figure 12 illustrates the interfaces of these two apps.



(a) iOS version (b) Android version

Fig. 9. Application BaNaVA in iOS and Android environment.

B. Performance of Vietnamese-Bahnaric translation service

We conducted experiments using a Vietnamese-Bahnaric dataset. This dataset consists of formal greetings, formal and informal conversations, narrative stories, and folktales composed in Bahnar Kriem. The dataset was segmented into three subsets: a training set, a test set, and a validation set, allocated for training, testing, and validation, respectively. Each subset comprises two text files. The first file stores Vietnamese sentences (saved as .vi files), while the other file contains Bahnaric sentences (saved as .ba files). In total, the training set consists of 16,105 sentence pairs, the test set comprises 1,988 pairs, and the validation set encompasses 1,987 pairs of sentences.

Following thorough research and multiple experiments conducted on the dataset, the authors calculated the respective BLEU scores for each model, presented in Table I. The obtained results indicate that the BN-BARTpho model secured the highest BLEU score, showcasing its outstanding efficiency within our methodology. These findings underscore the remarkable performance of the BN-BARTpho model in our research pursuits.

C. Performance of Bahnaric voice conversion service

In order to evaluate the voice conversion quality in Bahnaric, we created a user evaluation interface. This interface presented users with a set of 20 questions, each representing a unique evaluation scenario.

Regarding the scoring system, users were provided with a 6-level scale ranging from -1 to 100. This scale was designed to encompass a wide spectrum of perceived quality. It aims to convey to the evaluator the broad spectrum of sound quality, ranging from significantly subpar to human-level excellence. The evaluation results were gathered from 46 voluntary participants, and their statistics are depicted in Table II.

As indicated in Table II, there were no instances of poor quality in the original voice samples recorded by native speakers. Regarding the voice conversion models, the VC-original model is trained using data from the original voice, while the VC-GradTTS model is trained with an appropriate volume of data from the source domain derived from the Grad-TTS output. In particular, the VC-Grad-TTS model exhibits superior performance. The number of samples with poor to fair quality is notably reduced (representing 0.87%). Additionally, most samples generated by this model are evaluated as ranging from good to perfect. The mean evaluation score is also high at 80.33, falling within the range of good-quality sound.

TABLE I
BLEU score comparison
between BN-BARTpho and other Bahnar infused model

Model	BLEU
Transformer	28.78
PhoBERT-fused NMT	38.49
BARTphoEncoderPGN	47.91
Loanformer	42.63
PE-PD-PGN	49.00
BN-BARTpho	**49.20**

TABLE II
STARGANV2-VC SCORE COMPARISON WITH VC-ORIGINAL AND VC-GRADTTS

Sample type	Quality of recordings (%)						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-GradTTS	0.0	**0.87**	18.26	**55.54**	**23.59**	**1.74**	**80.33**

VIII. CONCLUSION

In this paper, we introduce application BaNaVA that aims to preserve the Ba Na language. To achieve this goal, the application supports translation and pronunciation of the Ba Na language on mobile phones. We have developed specialized NMT and TTS techniques to efficiently execute these AI models on commonly used smartphones by the Ba Na people. This application has been developed as a cross-platform solution and is available on both iOS and Android environments. Experimental results demonstrate that our AI models have achieved promising outcomes and are practical for real-world use.

ACKNOWLEDGMENTS

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. We also would like to extend our sincere thanks to Fessior Community (formerly Google Developer Student Club - HCMUT), particularly president Ly Gioi An, for their support in the development of the application. Our gratitude also goes to Pham Cong Thien, Dang Quang Vinh, Nguyen Phan Hoang Phuc, Hoang Duc Nguyen, and Tran Nguyen Thai Binh for their assistance in the final formatting and editing of this paper.

REFERENCES

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient Estimation of Word Representations in Vector Space. 2013.

[2] I. Sutskever, O. Vinyals, and Q. V. Le, Sequence to Sequence Learning with Neural Networks. 2014.

[3] D. Bahdanau, K. Cho, and Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate. 2016.

[4] T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Sep. 2015, pp. 1412–1421. doi: 10.18653/v1/D15-1166.

[5] S. Takase and S. Kiyono, "Lessons on Parameter Sharing across Layers in Transformers," in Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), Jul. 2023, pp. 78–90. doi: 10.18653/v1/2023.sustainlp-1.5.

[6] J. Zhu et al., "Incorporating BERT into Neural Machine Translation," 2020. [Online]. Available: https://openreview.net/forum?id=Hyl7ygStwB

[7] S. Takase and S. Kiyono, "Rethinking Perturbations in Encoder-Decoders for Fast Training," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Jun. 2021, pp. 5767–5780. doi: 10.18653/v1/2021.naacl-main.460.

[8] A. Currey, A. V. Miceli Barone, and K. Heafield, "Copied Monolingual Data Improves Low-Resource Neural Machine Translation," in Proceedings of the Second Conference on Machine Translation, Sep. 2017, pp. 148–156. doi: 10.18653/v1/W17-4715.

[9] R. Vázquez, A. Raganato, J. Tiedemann, and M. Creutz, "Multilingual NMT with a Language-Independent Attention Bridge," in Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), Aug. 2019, pp. 33–39. doi: 10.18653/v1/W19-4305.

[10] N. L. Tran, D. M. Le, and D. Q. Nguyen, "BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese," 2022.

[11] 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.

[12] 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."

[13] 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.

[14] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech," in Proceedings of the 38th International Conference on Machine Learning, Jul. 2021, vol. 139, pp. 8599–8608. [Online]. Available: https://proceedings.mlr.press/v139/popov21a.html

[15] S.-w. Park, D.-y. Kim and M.-c. Joe, "Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data," arXiv preprint arXiv:2005.03295, 2020.

[16] T. Kaneko, H. Kameoka, K. Tanaka and N. Hojo, "Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion," arXiv preprint arXiv:2010.11672, 2020.

[17] W.-C. Huang, H. Luo, H.-T. Hwang, C.-C. Lo, Y.-H. Peng, Y. Tsao and H.-M. Wang, "Unsupervised representation disentanglement using cross domain features and adversarial learning in variational autoencoder based voice conversion," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 4, p. 468–479, 2020.

[18] S. Ding and R. Gutierrez-Osuna, "Group Latent Embedding for Vector Quantized Variational Autoencoder in Non-Parallel Voice Conversion.," in Interspeech, 2019.

[19] K. Qian, Y. Zhang, S. Chang, X. Yang and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in International Conference on Machine Learning, 2019.

[20] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "VnCoreNLP: A Vietnamese Natural Language Processing Toolkit," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, Jun. 2018, pp. 56–60. doi: 10.18653/v1/N18-5012.

[21] D. Q. Nguyen, D. Q. Nguyen, T. Vu, M. Dras, and M. Johnson, "A Fast and Accurate Vietnamese Word Segmenter," in Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018, 2018. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2018/summaries/55.html

[22] N. L. Tran, D. M. Le, and D. Q. Nguyen, "BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese," 2022.

[23] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. doi: 10.18653/v1/2020.acl-main.703.

[24] J. Kong, J. Kim and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," Advances in Neural Information Processing Systems, vol. 33, p. 17022–17033, 2020.

[25] Y. Choi, Y. Uh, J. Yoo and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020.

[26] Q. T. Lam, D. H. Do, T. H. Vo and D. D. Nguyen, "Alternative vietnamese speech synthesis system with phoneme structure," in 2019 19th International Symposium on Communications and Information Technologies (ISCIT), 2019.