VIEACT-TTS-VC: A Vietnamese End-to-End Text-to-Speech and Voice Conversion Framework Using Low-Resource Adaptable Style Transfer

Hung D. Vo[1,2], Long S. T. Nguyen[1,2], Tri H. Trinh[1,2], Khang H. N. Vo[1,2], and Tho T. Quan[1,2(✉)]

[1] URA Research Group, Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
[2] Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam
qttho@hcmut.edu.vn

Abstract Recent advancements in end-to-end Text-to-Speech (TTS) systems have significantly improved speech naturalness. However, existing models often struggle with speaker adaptation in low-resource settings, leading to degraded speaker characteristics, unnatural prosody, and pronunciation errors, particularly in languages with complex phonetic structures. Vietnamese, a tonal language with six distinct tones, intricate pronunciation rules, and diverse regional accents, presents unique challenges for TTS synthesis. In this study, we introduce VIEACT-TTS-VC, a neural TTS and voice conversion system based on the VITS framework, designed to synthesize natural and intelligible Vietnamese speech while effectively adapting to speaker styles with limited data. The proposed system offers three key functionalities: 1) generating high-quality speech with style adaptation across diverse speakers, 2) enabling robust adaptation to new speech domains using small datasets, and 3) functioning as a hybrid system capable of both text-to-speech synthesis and voice conversion while preserving linguistic content. To evaluate the performance of VIEACT-TTS-VC, we employ both subjective and objective assessment metrics. Subjective evaluations, conducted through listener-based Mean Opinion Score, yield 4.31 ± 0.158 for comprehensibility and 3.68 ± 0.135 for naturalness with only 5 min of additional training data. Objective evaluations further demonstrate that VIEACT-TTS-VC surpasses state-of-the-art models in speaker style adaptation and exhibits remarkable robustness in low-resource scenarios.

Keywords: Vietnamese text-to-speech · Voice conversion · Low-resource style adaptation

1 Introduction

Natural Language Processing, a vital branch of Artificial Intelligence, enables machines to understand, interpret, and interact with human language. Among

K. Arai (Ed.): IntelliSys 2025, LNNS 1554, pp. 437–461, 2025.
https://doi.org/10.1007/978-3-031-99965-9_27

its subfields, speech synthesis plays a critical role in bridging the gap between machines and humans by generating natural sounding spoken language. *Text-to-Speech* (TTS) technology, in particular, converts digital text into speech, replicating linguistic content, tone, prosody, and rhythm [21]. Recent advancements in deep learning have significantly improved the quality of modern TTS systems, making them more natural and versatile [11]. Globally, over 7,150 languages are spoken, yet research and development resources are disproportionately focused on a few dominant languages like English, Mandarin Chinese, and Japanese. These languages benefit from extensive datasets and annotated corpora, which are crucial for training TTS systems. In contrast, low-resource languages, such as Vietnamese, face significant challenges. As the national language of Vietnam, Vietnamese is characterized by a complex system of tones, accents, and linguistic rules, but it lacks sufficient annotated corpora and linguistic tools tailored to its unique structure [16].

The linguistic complexity of Vietnamese demands robust synthesis methods capable of capturing its tonal and phonological attributes. Traditional approaches, such as concatenative and parametric TTS, often produce unnatural results for tonal languages. Similarly, statistical methods like Hidden Markov Models [22,25] require large datasets to generalize effectively. *Deep Neural Networks* (DNNs) [11], known for their success in other AI domains, offer a promising alternative. They can efficiently capture Vietnamese linguistic patterns and generate speech resembling human voice. Moreover, DNNs excel at transferring knowledge from resource-rich domains to resource-constrained ones, making them ideal for Vietnamese TTS [24]. In addition to TTS, *Voice Conversion* (VC) is gaining prominence as a complementary solution for modifying vocal characteristics, such as tone, prosody, and rhythm, while preserving the linguistic content of speech. Together, these technologies address the growing demand for customizable and adaptive voice applications in various industries.

We propose *VIEACT-TTS-VC*, an end-to-end Vietnamese TTS and VC system. Designed to overcome the challenges of low-resource data, VIEACT-TTS-VC integrates *state-of-the-art* (SOTA) speech synthesis models optimized for Vietnamese. By leveraging architectural modifications and transfer learning, the system models the intricate features of Vietnamese tones, accents, and rhythms. Furthermore, VIEACT-TTS-VC supports many-to-many VC, enabling transformations like gender or accent adaptation. Our systems aim to advance Vietnamese speech synthesis technology, address the challenges of low-resource languages, and provide an open-source framework to foster further innovation in this field.

Our main contributions to this study are summarized as follows.

- We generated natural-sounding speech audio directly from Vietnamese text, accurately reflecting native pronunciation despite limited training data.
- We transformed acoustic features of a source speaker's voice to match a target speaker while retaining prosody, content, and naturalness, enabling applications such as gender and accent conversion.

– We designed the system to require minimal data and computational resources, making VIEACT-TTS-VC accessible to individuals and small businesses. The lightweight pretrained model, achieved through quantization techniques, is also suitable for industrial deployment.

The remainder of this study is organized as follows: Sect. 2 reviews related work in TTS and VC modeling. Section 3 introduces the architecture of VIEACT-TTS-VC, along with its training and quantization strategies. Section 4 presents the experimentations, including datasets, evaluation metrics, results, and analysis. Section 5 concludes the study with a discussion of limitations and future directions. Supporting materials are included in the appendices, located after the References section.

2 Related Works

2.1 Overview of Traditional Text-to-Speech Methods

Traditional TTS technologies laid the foundation for modern speech synthesis by employing either pre-recorded speech segments or models of human speech production. Concatenative systems generate speech by combining stored audio units, with notable techniques such as diphone synthesis [3], which relies on small databases of phoneme transitions but struggles with rare linguistic contexts, and unit selection synthesis [6], which selects from larger, diverse databases to produce more natural speech, albeit at significant resource cost. In contrast, source-filter systems simulate the physical process of human speech production, as seen in articulatory synthesis [20], which replicates articulator movements but demands extensive data, and formant synthesis [1], which models resonant frequencies for efficient, rule-based synthesis, though constrained by manually defined rules. These traditional methods, while innovative at the time, faced challenges in handling diverse linguistic contexts and producing highly natural-sounding speech, especially for tonal languages like Vietnamese. These limitations led to the need for more data-driven approaches like statistical parametric synthesis.

2.2 Evolution to Statistical Parametric Synthesis

Statistical parametric synthesis, commonly known as HMM-based synthesis [22,25], represents a significant shift in TTS technology by adopting a data-driven approach. Unlike concatenative systems that directly concatenate speech segments, this method parameterizes speech characteristics and optimizes them during training. It consists of three components: a text analysis module for extracting linguistic features, an acoustic model such as HMMs that maps linguistic to acoustic features using probabilistic models, and a vocoder that generates speech from these features. Grapheme-to-phoneme conversion and linguistic feature extraction are critical sub-modules in this stage, bridging the gap between textual and acoustic representation. While HMM-based synthesis

reduced reliance on large datasets, it often produced robotic and noisy speech artifacts, limiting its ability to handle complex prosodic and tonal patterns found in languages like Vietnamese. This highlighted the necessity for neural-based methods to better capture these linguistic nuances.

2.3 Advances with Deep Learning in Speech Synthesis

Deep learning has transformed TTS systems by introducing end-to-end models capable of automatically learning features from raw data, eliminating the manual preprocessing required in traditional methods. Neural TTS models like WaveNet [23] revolutionized the field by directly generating waveforms, achieving unprecedented naturalness and intelligibility. Building on this, FastSpeech2 [19] improved phoneme prediction accuracy and reduced latency through linguistic feature integration and a simplified training pipeline, offering synthesis quality comparable to Tacotron2 while being significantly faster. Grad-TTS [17], a diffusion-based model, further advanced the field by removing dependencies on external aligners, achieving robust TTS alignment and generating mel-spectrograms with competitive quality at faster speeds. However, these neural models often require extensive datasets and computational resources, making them less suitable for low-resource tonal languages like Vietnamese, which demands tailored solutions for its complex prosodic and phonological characteristics.

2.4 Voice Conversion as a Complementary Technology

VC enables the transformation of a speaker's voice into another while preserving linguistic content, serving as a valuable complement to TTS. Recent advancements in deep learning have made VC pipelines fully end-to-end, simplifying acoustic features using mel-spectrograms, implementing mapping functions with neural networks, and employing neural vocoders for speech reconstruction. Non-parallel VC, which does not rely on paired data, has gained prominence due to its flexibility. StarGANv2-VC [13], a *Generative Adversarial Network* (GAN)-based approach, uses adversarial loss and cycle-consistency to convert speaker identity while discarding the source speaker's characteristics. While effective, these approaches often fail to address the needs of tonal languages like Vietnamese, where linguistic features such as pitch and tone must be preserved. Moreover, the lack of seamless integration between VC and TTS systems limits their applicability in creating versatile speech synthesis solutions. This gap underscores the importance of developing a unified framework tailored to tonal languages.

3 VIEACT-TTS-VC

3.1 Baseline Model

Kim et al. [8] proposed the *Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech* (VITS) architecture, which serves

as a strong foundation for end-to-end TTS synthesis. As with modern TTS systems, the VITS architecture comprises three core components, as illustrated in Fig. 1. The following descriptions are based on [8] and our insights during the reproduction process.

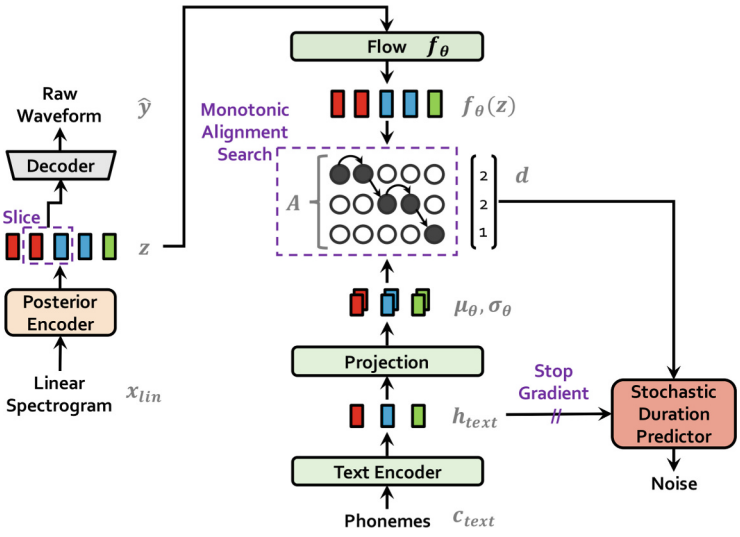

Fig. 1. Overview of the VITS architecture [8]

Text Encoder. A dense neural network transforms the input phonemes into a latent vector representation that captures essential linguistic and contextual features. In addition to linguistic content, the phoneme embeddings encode prosodic attributes such as stress, duration, and intonation that are critical elements for generating natural-sounding speech.

Prior Encoder. This component consists of an end-to-end text encoder and a normalizing flow module. The normalizing flow employs a stack of affine coupling layers [5], implemented using WaveNet residual blocks. Each block includes layers of dilated convolutions with gated activation units and skip connections. The normalizing flow function f_θ transforms the encoded acoustic features into a latent prior distribution.

Posterior Encoder. Adapted from the non-causal WaveNet architecture [16,18], this module generates the posterior distribution of latent variables. A linear projection layer outputs the mean and variance of a normal distribution. In multi-speaker scenarios, a speaker embedding is derived from the speaker ID using a dense neural network. The posterior encoder maps the input mel-spectrogram x_{lin} to latent variables Z.

Decoder. The decoder is a HiFi-GAN V1 generator [9], comprising multiple transposed convolutional layers, each followed by a Multi-Receptive Field Fusion module. This component synthesizes the final waveform from the latent representation.

Stochastic Duration Predictor. This module uses residual blocks with dilated and depthwise separable convolutions. It stochastically predicts phoneme durations to enhance speech variability and realism.

3.2 Our Proposed Framework

One of the main challenges in adaptive TTS and VC is the requirement for large-scale, high-quality datasets to ensure naturalness, speaker diversity, and cost-effectiveness, especially for low-resource languages like Vietnamese. However, there are only a few public datasets available for Vietnamese TTS due to the high costs associated with collecting and maintaining such resources. Creating a comprehensive dataset typically requires recordings of various speaking styles, accents, and demographics. On the other hand, there exists a significant amount of publicly available data for speech recognition tasks. By leveraging not only open-source data designed for TTS, but also datasets intended for automatic speech recognition, we can access a large pool of speakers with diverse styles and accents. This enables the construction of a high-quality and adaptive dataset suitable for both TTS and VC tasks. The data preprocessing pipeline is illustrated in Fig. 2.

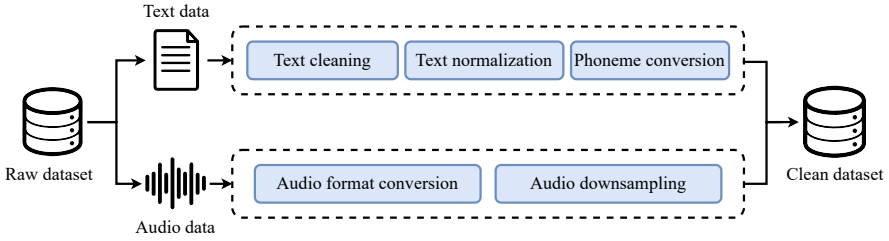

Fig. 2. Process of building a comprehensive and adaptive database

The preprocessing pipeline consists of two parallel branches: *Text Preprocessing* and *Audio Preprocessing*. In the text branch, three main steps are applied: *text cleaning*, *text normalization*, and *phoneme conversion*. Text cleaning involves removing punctuation, converting text to lowercase, and eliminating unnecessary characters. The cleaned text is then passed through a normalization module to expand abbreviations into their full forms. Finally, the normalized text undergoes phoneme conversion. In the audio branch, the input audio files are first converted to the WAV format and then downsampled to a standard sampling

rate of 16 kHz. This ensures consistency and compatibility with most TTS and VC models. The output is a clean, standardized dataset suitable for training adaptive speech models.

Grapheme-to-Phoneme Conversion. The baseline model's text processing pipeline was originally designed for English datasets. To adapt it for the Vietnamese language, several modifications were necessary to address language-specific characteristics.

First, Vietnamese-specific vowels (e.g., 'ă', 'ê', 'ơ') were added to the grapheme inventory. Second, because Vietnamese often uses compound words to convey complete meanings such as "lạnh lẽo" (where "lẽo" alone is semantically meaningless) - word segmentation becomes essential. In the TTS context, these compound words are typically not separated acoustically, so a special character ('_') is used to indicate segmentation within the grapheme dictionary. Additionally, to represent the six lexical tones of Vietnamese, numerical characters from 1 to 6 are also included. The process of mapping a text string to a sequence of symbol *Identifiers* (IDs), and the reverse conversion from ID sequences to text, has also been adapted accordingly. This transformation process is illustrated in Fig. 3.

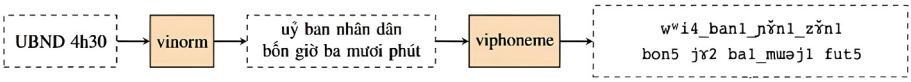

Fig. 3. Workflow of the vietnamese G2P module [15]

The *Grapheme-to-Phoneme* (G2P) module integrates two key components: the *Vinorm* and *Viphoneme* libraries [15]. The *Vinorm* module serves as a robust Vietnamese text normalization tool. It standardizes characters, removes redundant white spaces, corrects common spelling errors, and converts non-standard representations such as dates, abbreviations, and numbers into consistent formats. The *Viphoneme* module is responsible for converting Vietnamese text into phonetic representations. It first converts text into a grapheme format and then maps it to the *International Phonetic Alphabet* (IPA). Specifically, it supports conversion to and from a predefined set of 138 IPA symbols, ensuring accurate coverage of Vietnamese phonemes. Furthermore, the text cleaner is fine-tuned to handle punctuation such as commas and periods. These marks are removed appropriately while preserving natural spacing, thus contributing to smoother prosody in the generated speech.

Vietnamese Phoneme-BERT. Pretrained *Large Language Models* (LLMs) have demonstrated strong capabilities in capturing rich textual information. In TTS synthesis, phoneme representations play a vital role in determining speech naturalness. Recent studies, such as Phoneme-Level BERT [12], PnG BERT [7],

Mixed-Phoneme BERT [26], and XPhoneBERT [14], have shown that pretrained LLMs can effectively provide contextual information for phoneme-level representations, even when trained on out-of-distribution text data. These models significantly enhance the naturalness and prosody of TTS systems by injecting contextual knowledge into phoneme embeddings.

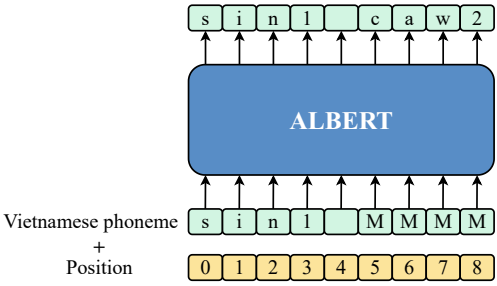

Fig. 4. Vietnamese Phoneme-BERT predicts the phrase "xin chào": `/sin1 caw2/`

However, the aforementioned works are primarily trained on English datasets. To adapt this approach for Vietnamese, we propose a customized *Vietnamese Phoneme-BERT*, as illustrated in Fig. 4, which adopts the same architecture as ALBERT-base [10]. The key modification lies in the vocabulary: it now consists of 138 phoneme symbols based on the IPA, as defined in the Vietnamese text-cleaner module. The number of transformer layers, hidden size, and embedding dimension are kept consistent with ALBERT-base: 12 layers, 768 hidden units, and 128-dimensional embeddings, respectively. The training procedure follows the standard masked language modeling approach, adapted for phoneme tokens. Specifically, the model is trained to predict masked phoneme tokens based on their surrounding context. In our framework, the original text encoder of the baseline VITS model [8] is replaced with our Vietnamese Phoneme-BERT to enhance contextual understanding at the phoneme level.

Training Objective. The training objective is based on *Masked Phoneme-Token Prediction* (MPP). Given a sequence of masked phoneme tokens, the model learns to predict the original tokens using a linear projection followed by a softmax layer applied to the hidden states from the final transformer layer. The training loss is computed using the cross-entropy function, which is commonly used for multi-class classification. We employ the `AdamW` optimizer with a learning rate of 2×10^{-5}. The objective is formalized in Eq. 1.

$$l = \mathbb{E}_{N,x} \left(\sum_{i=1}^{N} \mathrm{CE} \left(\mathrm{P}_{\mathrm{mpp}} \left(\mathrm{E}(x) \right)_{i,y_i} \right) \right), \tag{1}$$

where,

- x is the input sequence containing masked phoneme tokens,
- N is the total number of masked tokens in the sequence,
- i is the index of a masked token,
- $\mathrm{E}(x)$ denotes the phoneme-level encoder outputs for sequence x,
- $\mathrm{P}_{\mathrm{mpp}}$ is the linear projection layer used for predicting masked tokens,
- $\mathrm{CE}(\cdot)$ is the cross-entropy loss function,
- y_i is the ground-truth (unmasked) phoneme label corresponding to token i,
- $\mathbb{E}_{N,x}$ denotes the expectation over the training batch, averaging across all masked positions and input sequences.

Masking Strategy. To train the model to learn meaningful semantic and phonetic representations, we adopt a phoneme-level masking strategy. Specifically, 15% of graphemes in each sequence are randomly selected. For each selected grapheme, its corresponding phoneme tokens are: 1) replaced with a `[MASK]` token 80% of the time, 2) replaced with random phoneme tokens 10% of the time, and 3) left unchanged 10% of the time. This strategy enables the model to generalize better and learn contextual dependencies among phonemes.

Vietnamese Style Embedding Encoder. The proposed Vietnamese Style Embedding Encoder improves upon previous implementations by enabling the VIEACT-TTS-VC model to learn speech characteristics from raw utterances without requiring explicit style or speaker annotations. This capability enhances the system's adaptability to a wide range of speaking styles and improves its ability to capture subtle speaker-specific and prosodic variations. The architecture of the style encoder consists of three main components: a *reference encoder*, a *style token layer*, and an *acoustic extractor*. Given a reference speech signal in the form of a log-mel spectrogram, the encoder aims to extract a latent vector that encapsulates expressive style attributes such as speaker identity and prosody. The overall architecture is illustrated in Fig. 5.

Reference Encoder. The reference encoder consists of a stack of convolutional layers followed by a recurrent unit. It takes a log-mel spectrogram as input, which is processed through six 2D convolutional layers, each with a 3×3 kernel and a stride of 2×2. Batch normalization and ReLU activations are applied after each layer. The number of output channels for the six layers is set to 32, 32, 64, 64, 128, and 128, respectively. The resulting tensor is reshaped to maintain the temporal dimension and passed through a unidirectional GRU with 128 units. The final hidden state of the GRU serves as the reference embedding, which is passed to the style token layer.

Style Token Layer. The style token layer contains a learnable bank of style embeddings along with an attention mechanism. Unless stated otherwise, we use 10 learnable style tokens, which are sufficient to capture diverse prosodic patterns observed in the training data. Each token has a dimensionality of 256 to match the output of the text encoder.

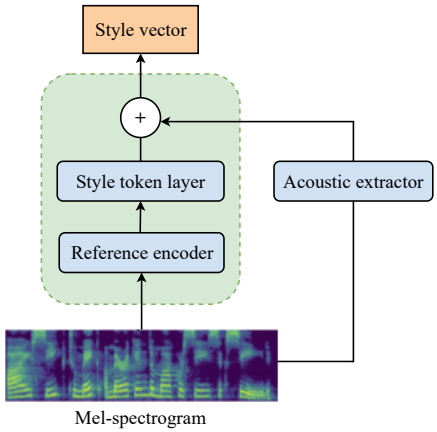

Fig. 5. Overview of the vietnamese style encoder architecture

Acoustic Extractor. The acoustic extractor further refines the style representation from the mel-spectrogram input. Its architecture is shown in Fig. 6.

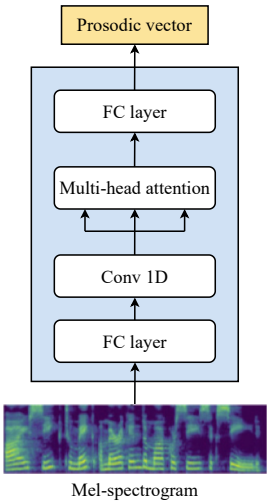

Fig. 6. Architecture of the acoustic extractor module

It consists of the following stages:

– *Spectral Processing:* The mel-spectrogram is first passed through fully connected layers to project each frame into a latent space, producing a sequence of hidden vectors.

– *Temporal Processing:* Gated convolutional neural networks [4] with residual connections capture temporal dependencies across frames.
– *Multi-Head Self-Attention:* To capture global style information, we apply a multi-head self-attention mechanism with residual connections. Unlike previous work [2], where self-attention is applied across entire audio samples, we apply it at the frame level to better capture style information from short utterances. The outputs are then averaged over time to yield a fixed-dimensional style vector.

Fine-Tuning Strategy for Vietnamese Low-Resource TTS. The architecture of VIEACT-TTS-VC is designed to learn speech characteristics directly from raw audio, without relying on speaker identifiers or manually annotated labels. This capability enables the model to take full advantage of large-scale unlabeled speech corpora during pretraining, allowing it to internalize a wide range of voice traits and prosodic patterns. As a result, VIEACT-TTS-VC provides a strong pretrained foundation that can be effectively adapted to downstream tasks. The fine-tuning process, illustrated in Fig. 7, aims to specialize the model for Vietnamese by optimizing its understanding of phoneme-level representations.

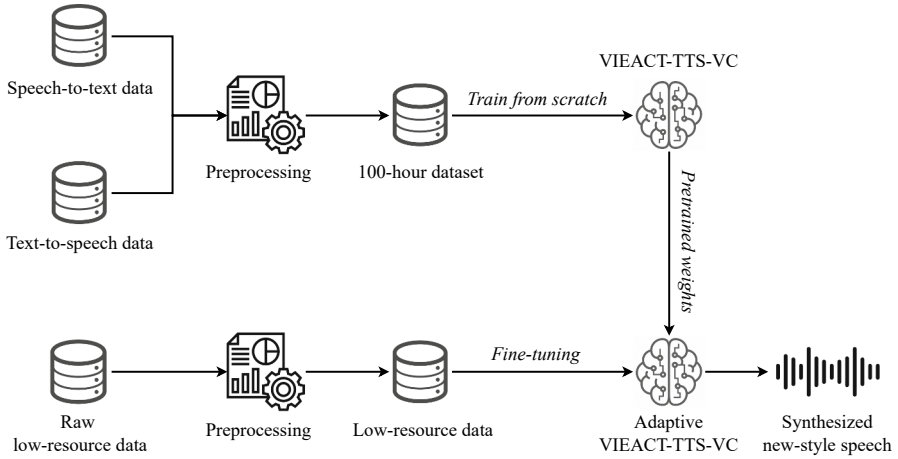

Fig. 7. Overview of the fine-tuning strategy

During fine-tuning, the model is further trained on phoneme sequences to improve its ability to represent the phonetic nuances of Vietnamese speech. Importantly, VIEACT-TTS-VC is optimized for low-resource scenarios, making it suitable for users with limited computational capacity. Furthermore, its strong transferability allows for rapid adaptation to new Vietnamese speech domains, even when training data is scarce, by leveraging the rich knowledge acquired during pretraining.

Hybrid Model for TTS and VC. VIEACT-TTS-VC is a hybrid architecture designed to address both TTS and VC tasks within a unified framework, as depicted in Fig. 8. It adopts an unsupervised and nonparallel training strategy, offering flexibility and robustness in low-resource scenarios (see Appendix A).

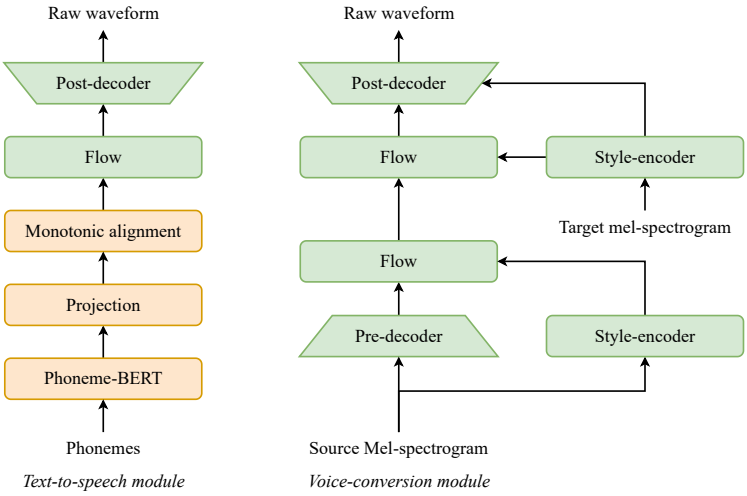

Fig. 8. The overall architecture of the VIEACT-TTS-VC framework

By leveraging the style encoder and pretrained model weights described in Sect. 3.2, the framework captures comprehensive prosodic features such as accent, speaking style, and gender, enabling expressive and speaker-adaptive synthesis. This design improves both synthesis quality and style transfer capability, allowing the system to generalize effectively across diverse speakers and application domains. Notably, VIEACT-TTS-VC supports nonparallel VC, removing the need for aligned training pairs and improving scalability in real-world usage.

Model Quantization with the ONNX Framework. Quantization in ONNX Runtime refers to the process of mapping high-precision floating-point values into a lower-precision 8-bit integer space. This technique reduces model size and inference latency while preserving acceptable accuracy. The linear relationship between a floating-point value x and its quantized counterpart x_q is defined as in Eq. 2.

$$x = S \cdot (x_q - Z), \tag{2}$$

where,

- $S \in \mathbb{R}^+$ is the scale, a positive real number used to map quantized values back to the floating-point domain.
- $Z \in \mathbb{Z}$ is the zero-point, an integer ensuring that the value zero is exactly representable in the quantized space.

– $x_q \in [q_{\min}, q_{\max}] \subset \mathbb{Z}$ is the quantized integer value.

The scale factor S is computed differently depending on the quantization scheme:

Asymmetric Quantization. In this scheme, both the scale and zero-point are derived from the data range. The scale is computed according to Eq. 3.

$$S = \frac{x_{\max} - x_{\min}}{q_{\max} - q_{\min}}, \tag{3}$$

where, $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of the original floating-point range, and $q_{\max}$, $q_{\min}$ define the bounds of the quantization range.

Symmetric Quantization. In this approach, the quantization range is centered around zero, and the zero-point Z is fixed at zero. The scale is computed using the maximum absolute value of the input range, as in Eq. 4.

$$S = \frac{2 \cdot \max(|x_{\max}|, |x_{\min}|)}{q_{\max} - q_{\min}}, \tag{4}$$

ensuring that the range is symmetric about zero and fully utilizes the available quantization space.

4 Experimentations

We conduct extensive experiments to evaluate the performance, adaptability, and efficiency of the proposed VIEACT-TTS-VC framework across both TTS and VC tasks. Our evaluation includes comparisons with several strong baselines. To gain deeper insights into the architecture, we also perform component-wise ablation studies. The overall assessment leverages both subjective human judgments and objective quantitative metrics, ensuring a comprehensive evaluation of speech quality, intelligibility, and computational efficiency.

4.1 Datasets

We conduct our experiments using three public Vietnamese speech datasets: *InFore*, *VLSP 2020*, and *VinBigData*. These datasets include both male and female voices, covering a wide range of accents from Northern and Southern Vietnam. Detailed descriptions and the specific roles of each dataset in our evaluation are provided in Appendix C.

4.2 Baselines

To benchmark the performance of our proposed system, we compare VIEACT-TTS-VC against several SOTA TTS architectures. All models were trained on the InFore dataset using varying amounts of training data: 5, 10, 30, and 45 min. The evaluated models include:

VITS. A high-quality end-to-end TTS system.

Grad-TTS and FastSpeech2. Two widely adopted models known for both synthesis quality and inference speed.

Pretrained VIEACT-TTS-VC. Fine-tuned for improved speaker adaptation and prosodic control.

8-Bit Quantized VIEACT-TTS-VC. A lightweight version designed to assess the trade-off between compression and quality.

We also perform an ablation study by selectively removing key components of VIEACT-TTS-VC, including the Phoneme-BERT module and the style encoder, to evaluate their individual contributions.

4.3 Evaluation Metrics

Objective Evaluation Metrics. To ensure consistent and reproducible comparisons across models, we apply the following automatic metrics:

- *Mel Cepstral Distortion with Dynamic Time Warping - Segment Level (MCD-DTW-SL)*: Measures spectral distance between synthesized and reference audio; lower is better.
- *Model Size and Inference Time*: Evaluates computational efficiency, especially for deployment in CPU-constrained environments.

Subjective Evaluation Metrics. We conduct human evaluations to assess two perceptual qualities: intelligibility and naturalness. Details of the sentence sets used in the experiments are available in Appendix D.

Intelligibility Assessment. Participants transcribed 15 utterances generated by the models. These utterances were designed to cover seven common syntactic structures:

1. Subject - Verb - Object - Adverbial
2. Verb - Subject - Adverbial
3. Question Word - Subject - Verb - Object - Adverbial
4. Subject - Verb - Object - Adverbial
5. Dependent Clause - Independent Clause
6. If - Clause - Main Clause
7. Subject - Relative Pronoun - Predicate

The evaluation interface is shown in Fig. 9.

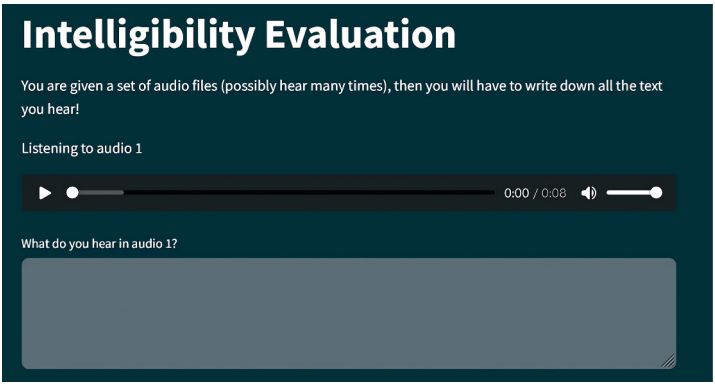

Fig. 9. Intelligibility evaluation session

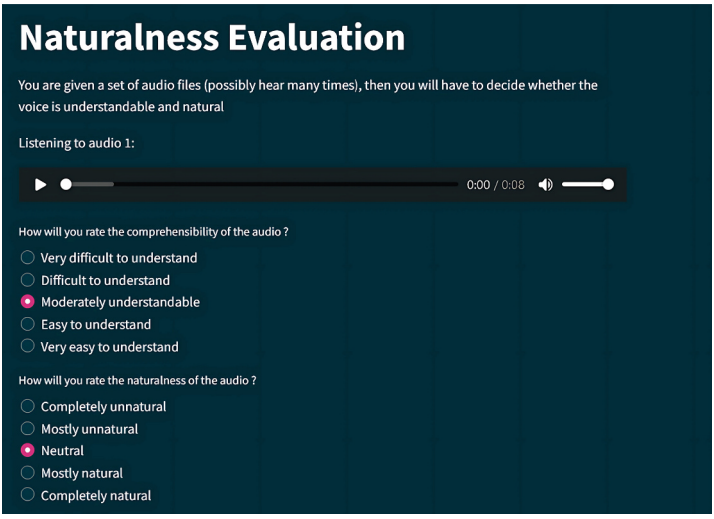

Fig. 10. Comprehensibility and naturalness evaluation session

Naturalness Assessment. Participants rated 15 sentences, phrases, or words of varying lengths based on perceived comprehensibility and naturalness, using a five-point *Likert scale*. The evaluation interface is depicted in Fig. 10.

The following subjective metrics were used:

- *MOS-Naturalness*: Rates the naturalness of synthesized speech.
- *MOS-Comprehensibility*: Rates how easy the speech is to understand.
- *Word Error Rate (WER)*: Calculated from transcription accuracy; lower values indicate higher intelligibility.

Definitions and formulas for these metrics are provided in Appendix B.

4.4 Results and Analysis

Evaluation Using MCD-DTW-SL Metric. We compare several SOTA approaches using the MCD-DTW-SL metric and conduct an ablation study to evaluate the contribution of individual components within our model. The MCD-DTW-SL metric serves as an objective measure of speech synthesis quality, where lower values indicate higher fidelity to the reference signal. Table 1 presents the results across different training durations. VIEACT-TTS-VC consistently achieves the lowest error across all data settings, confirming its effectiveness even under low-resource conditions.

Table 1. MCD-DTW-SL scores for different models across training durations

Model Name	5 mins	10 mins	30 mins	45 mins
VITS	14.66	13.33	14.12	13.71
Grad-TTS	22.00	20.08	19.16	18.12
FastSpeech2	23.31	21.96	19.21	18.57
VIEACT-TTS-VC	**8.64**	**8.75**	**8.40**	**8.33**
Quantized VIEACT-TTS-VC	8.91	**8.75**	8.49	8.67

Ablation Study on Core Components. To better understand the role of each component, we conduct an ablation study whose results are shown in Table 2. Key observations include:

- The full VIEACT-TTS-VC model, which incorporates both the Phoneme-BERT and style encoder, delivers the best performance across all training durations.
- The Quantized VIEACT-TTS-VC model shows only slight degradation in quality, suggesting that 8-bit quantization is a viable option for deployment on resource-constrained devices.
- Removing the Phoneme-BERT module leads to increased MCD-DTW-SL scores, especially in low-data settings, confirming its importance for phoneme-level precision.
- Excluding the style encoder results in further degradation, underscoring its role in maintaining prosody and speaker expressiveness.

Table 2. Ablation study results across training durations

Model Name	5 mins	10 mins	30 mins	45 mins
VIEACT-TTS-VC	**8.64**	**8.75**	**8.40**	**8.33**
Quantized VIEACT-TTS-VC	8.91	**8.75**	8.49	8.67
VIEACT-TTS-VC w/o Phoneme-BERT	8.99	8.86	8.51	8.45
VIEACT-TTS-VC w/o Style Encoder	9.30	9.33	8.45	8.41

Subjective Evaluation: MOS and WER. To further assess perceptual quality, we use human-annotated MOS for both naturalness and comprehensibility, as well as the WER for intelligibility. Table 3, Table 4, and Table 5 report these metrics for models trained with 5 and 45 min of data, along with 95% confidence intervals where applicable.

Table 3. MOS-naturalness comparison across different training durations

Model	5 mins	45 mins
VITS	N/A	N/A
Grad-TTS	N/A	N/A
FastSpeech2	N/A	N/A
VIEACT-TTS-VC	3.64 ± 0.216	4.31 ± 0.158
Quantized VIEACT-TTS-VC	3.68 ± 0.135	4.20 ± 0.122

Table 4. MOS-comprehensibility comparison across different training durations

Model	5 mins	45 mins
VITS	N/A	N/A
Grad-TTS	N/A	N/A
FastSpeech2	N/A	N/A
VIEACT-TTS-VC	4.48 ± 0.273	4.80 ± 0.104
Quantized VIEACT-TTS-VC	4.52 ± 0.249	4.75 ± 0.158

Table 5. WER comparison across different training durations

Model	5 mins	45 mins
VITS	N/A	N/A
Grad-TTS	N/A	N/A
FastSpeech2	N/A	N/A
VIEACT-TTS-VC	3.18	2.67
Quantized VIEACT-TTS-VC	3.18	2.67

Evaluation of Model Efficiency. We assess the model efficiency in terms of inference speed and memory footprint. Table 6 compares the model size, number of parameters, and inference time (CPU) across all baselines and our proposed system. The Quantized VIEACT-TTS-VC achieves a remarkable balance between performance and compactness, making it suitable for real-time and resource-constrained applications.

Table 6. Model comparison of size and inference time

Model Name	Model Size	No. Parameters	Inference Time
VITS	429 MB	34.3M	0.6 s/query
Grad-TTS	56.8 MB	**24M**	**0.1 s/query**
FastSpeech2	409.5 MB	78M	0.5 s/query
VIEACT-TTS-VC	540.5 MB	48.7M	0.9 s/query
Quantized VIEACT-TTS-VC	**50.4 MB**	48.7M	0.5 s/query

Evaluation of VC Capability. To evaluate the VC capability of the proposed framework, we compare VIEACT-TTS-VC with a SOTA baseline, StarGANv2-VC, using the MOS-naturalness metric. The assessment covers four conversion scenarios: *male-to-male* (M2M), *female-to-male* (F2M), *female-to-female* (F2F), and *male-to-female* (M2F). Table 7 reports the results along with 95% confidence intervals.

Table 7. MOS-naturalness comparison of VC models

Model Name	M2M	F2M	F2F	M2F
StarGANv2-VC	4.02 ± 0.22	4.06 ± 0.28	4.25 ± 0.26	4.03 ± 0.26
VIEACT-TTS-VC	3.57 ± 0.27	3.69 ± 0.216	3.75 ± 0.18	3.68 ± 0.159

Summary of Key Findings. Based on the experimental results across all tasks, several important observations are noted:

- *Adaptability to New Domains:* The model adapts well to new speakers and domains, even with limited data. With only 60 training samples, it successfully replicates the target speaker's style and prosody.
- *High Speech Naturalness:* The synthesized speech exhibits high naturalness, indicating that the model effectively captures speaker-specific acoustic characteristics. In contrast, conventional SOTA models trained from scratch often fail to generalize in low-resource scenarios (marked as N/A).
- *Effect of Utterance Length:* Naturalness and comprehensibility improve with longer utterances. A noticeable performance gap is observed for samples shorter than 3 s, which tend to be less fluent.
- *Model Compression Feasibility:* The quantized version performs comparably to the full model, suggesting that compression introduces minimal degradation. This confirms its practicality for deployment on resource-constrained devices.

– *VC Module Performance:* While the VC module produces intelligible and acceptable-quality speech, it still falls slightly behind dedicated SOTA VC systems. Nonetheless, it remains usable and suitable for integration into unified frameworks.

5 Discussion, Limitations, and Future Works

This study presents VIEACT-TTS-VC, a Vietnamese end-to-end framework designed for both TTS and VC tasks in low-resource and adaptive style transfer settings. By integrating Phoneme-BERT and a style encoder, the model enables unsupervised speaker adaptation and prosody modeling without the need for parallel or labeled data. It achieves strong performance with as little as 5 min of adaptation data and retains effectiveness after quantization, making it suitable for edge deployment. In addition to delivering high-quality speech, the model supports both TTS and VC within a unified architecture, enabling broad applicability across real-world scenarios such as virtual assistants and educational platforms. To demonstrate its practicality, VIEACT-TTS-VC has been designed for integration into a web-based application, as illustrated in Appendix E. The results collectively suggest that the framework is not only effective but also scalable for broader usage.

However, several limitations remain. The system still exhibits limited generalization in zero-shot adaptation to unseen speakers. It also lacks support for multilingual synthesis and emotional expressiveness. Furthermore, while quantization reduces model size, additional optimization is needed for ultra-constrained environments.

Future work will focus on enhancing zero-shot adaptability, extending the framework to multilingual and emotion-adaptive voice modeling, and further compressing the model for efficient deployment in low-power settings.

Appendices

A. Model Configuration

Table 8 summarizes the architectural components and parameter distribution of the VIEACT-TTS-VC framework for both TTS and VC modes.

B. Evaluation Techniques for Speech Synthesis Systems

Word Error Rate. Intelligibility measures the degree to which synthesized speech can be accurately transcribed by listeners. A common metric used to evaluate this is the *Word Error Rate* (WER), which reflects the number of insertions, deletions, and substitutions needed to convert the hypothesis (transcribed speech) into the reference (original text). WER is computed as shown in Eq. 5.

$$\text{WER} = \frac{S + D + I}{N}, \tag{5}$$

where,

Table 8. Model configuration of VIEACT-TTS-VC

Layer Name	N-layer TTS	N-layer VC	Number of Parameters
Phoneme-BERT	1	0	12.7M
Pre-encoder	1	2	8.8M
Style encoder	1	1	2.63M
Duration predictor	1	0	1.5M
Flow block	1	1	8.7M
Post-encoder	1	1	14.4M
Total parameters	48.7M	43.3M	48.7M
Model size	540.5MB	456.5MB	–

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- N is the number of words in the reference sentence.

A lower WER value indicates better intelligibility of the generated speech.

Comprehensibility and Naturalness. Comprehensibility refers to how easily a listener can understand the intended meaning of synthesized speech. Unlike intelligibility, it focuses on the cognitive effort required to decode the message. Naturalness evaluates how human-like the synthesized speech sounds, considering prosody, rhythm, and fluency. Both metrics are subjectively evaluated on a five-point Likert scale:

- *Comprehensibility:* 1 (very difficult to understand) to 5 (very easy to understand),
- *Naturalness:* 1 (very unnatural) to 5 (very natural).

The *Mean Opinion Score* (MOS) is used to aggregate listener ratings. First, the average score of each listener is computed as shown in Eq. 6.

$$\text{MOS}_j = \frac{1}{N} \sum_{i=1}^{N} \text{Score}_{ij}. \tag{6}$$

Then, the overall MOS across M listeners is calculated as in Eq. 7.

$$\text{MOS} = \frac{1}{M} \sum_{j=1}^{M} \text{MOS}_j, \tag{7}$$

where

- Score_{ij} is the score given by the j-th listener for the i-th sentence,

- N is the number of test sentences,
- M is the number of listeners,
- MOS_j is the average score for listener j,
- MOS is the overall average score.

Equation 6 and Eq. 7 provide a standardized method to convert subjective assessments into quantitative values.

Mel Cepstral Distortion

Mel Cepstral Distortion. (MCD) is an objective measure that evaluates the spectral distance between synthesized and reference speech using *Mel-Frequency Cepstral Coefficients* (MFCCs). It quantifies how similar the synthesized audio is to the ground truth in the frequency domain. Three main variations are commonly used:

- *MCD (Plain):* Assumes both input sequences have equal length; shorter utterances are zero-padded.
- *MCD-DTW:* Applies *Dynamic Time Warping* (DTW) to align MFCC sequences of varying lengths.
- *MCD-DTW-SL:* An improved version from [26] that adds a penalty factor based on the length mismatch. The distance between two MFCC frames is given in Eq. 8.

$$d(c_t, c'_t) = \sqrt{\sum_{k=1}^{K} (c_{t,k} - c'_{t,k})^2}, \tag{8}$$

where, $c_{t,k}$ and $c'_{t,k}$ are the k-th MFCC coefficients of frame t from the synthesized and reference audio, respectively.
The recursive DTW cost is calculated in Eq. 9.

$$\gamma_{i,j} = d(c_i, c'_j) + \min(\gamma_{i-1,j-1}, \gamma_{i-1,j}, \gamma_{i,j-1}). \tag{9}$$

The final MCD-DTW-SL score is obtained using Eq. 10.

$$\text{MCD-DTW-SL}(C, C') = \left(\frac{\eta}{R}\right) \cdot \gamma_{M,N}, \tag{10}$$

where,

- $\gamma_{M,N}$ is the cumulative DTW alignment cost between M and N frames,
- R is the length of the optimal warping path,
- $\eta = \frac{\max(M,N)}{\min(M,N)}$ is the penalty factor for sequence length difference.

Table 9. Speech datasets used for training

Dataset	Sex	Duration	Sample Rate	Accent
InFore	Female	30 h	22050 Hz	Northern
VLSP 2020	Male/Female	20 h	22050 Hz	Northern/Southern
VinBigData	Male/Female	96 h	16000 Hz	Various

C. Speech Datasets

The datasets utilized in this study include *InFore*, *VLSP 2020*, and *VinBigData*, each contributing to different aspects of the TTS and speech recognition tasks. Table 9 provides a summary of their characteristics.

The InFore dataset is primarily used for the TTS synthesis task, offering 30 h of high-quality recordings from a female speaker with a Northern accent. Meanwhile, the VLSP 2020 and VinBigData datasets - comprising both male and female speakers - are used for the speech recognition and pretraining tasks. These datasets include a variety of accents and speech patterns, thereby improving model robustness. The pretrained VIEACT-TTS-VC model was trained on a merged corpus of VLSP 2020 and VinBigData datasets, totaling approximately 125 h of speech. The training was conducted on an NVIDIA Tesla V100 GPU and required about one week to complete.

For the style adaptation experiments and baseline comparisons, subsets of the InFore dataset were used. These subsets ensure consistency in benchmarking with existing SOTA models. The configuration and hyperparameters used for fine-tuning the VIEACT-TTS-VC model on different durations of data are detailed in Table 10.

Table 10. Fine-tuning configuration of VIEACT-TTS-VC

Subset of InFore	Number of Samples	Fine-tuning Time	Batch Size
5 mins	60	10 mins	5
10 mins	125	15 mins	5
30 mins	325	20 mins	12
45 mins	500	20 mins	12

D. Evaluation Sentences

The following 15 Vietnamese sentences were used for evaluating the intelligibility, naturalness, and comprehensibility of synthesized speech produced by the VIEACT-TTS-VC system. These sentences were selected to cover a range of linguistic structures, vocabulary complexity, and prosodic variation.

1. Lá lành đùm lá rách.
2. Xử lý ngôn ngữ tự nhiên.
3. Khoa học và công nghệ.
4. Công nghệ chuyển đổi văn bản thành giọng nói.
5. Nhánh sông cổ hé lộ cách người Ai Cập xây kim tự tháp.
6. Công cụ chuyển đổi văn bản thành giọng nói là công nghệ có khả năng hiểu văn bản, cũng như ngôn ngữ tự nhiên, dựa trên trí tuệ nhân tạo, từ đó tạo ra âm thanh tổng hợp hoàn chỉnh với nhịp điệu và ngữ điệu phù hợp, giống như giọng nói tự nhiên của con người.
7. Mô hình ngôn ngữ lớn là một loại mô hình ngôn ngữ được đào tạo bằng cách sử dụng các kỹ thuật học sâu trên tập dữ liệu văn bản khổng lồ.
8. Ngôi đền này được xây dựng từ thế kỷ thứ X, tương truyền là nơi thần linh cai quản, bảo vệ cho hòn đảo khỏi những linh hồn quỷ dữ.
9. Việc nhất trí nâng cấp quan hệ hai nước lên Đối tác Chiến lược Toàn diện vì hòa bình, hợp tác và phát triển bền vững được kỳ vọng sẽ mở ra một chương mới trong quan hệ Việt Nam - Hoa Kỳ.
10. Cây cô đơn là một địa danh rất quen thuộc với những người từng đến, ở lại và lỡ phải lòng với vùng đất ngàn hoa Đà Lạt.
11. Khi ấy, tôi còn ở Tây Nguyên làm phiên dịch cho một công ty cà phê của Đức đặt văn phòng ở Ban Mê Thuột.
12. Chúng ta cần phải tập trung vào việc bảo vệ môi trường sống chung của chúng ta để tương lai con cháu được hưởng một hành tinh xanh hơn.
13. Điều quan trọng nhất trong cuộc sống là biết đặt mình vào vị trí của người khác để hiểu và chia sẻ.
14. Sự hiểu biết, tự tin sẽ giúp chúng ta vượt qua mọi khó khăn trong cuộc sống.
15. Einstein còn đề xuất nguyên lý nổi tiếng về sự tương đương giữa khối lượng và năng lượng. Công thức này cho thấy khối lượng và năng lượng có thể chuyển hóa lẫn nhau.

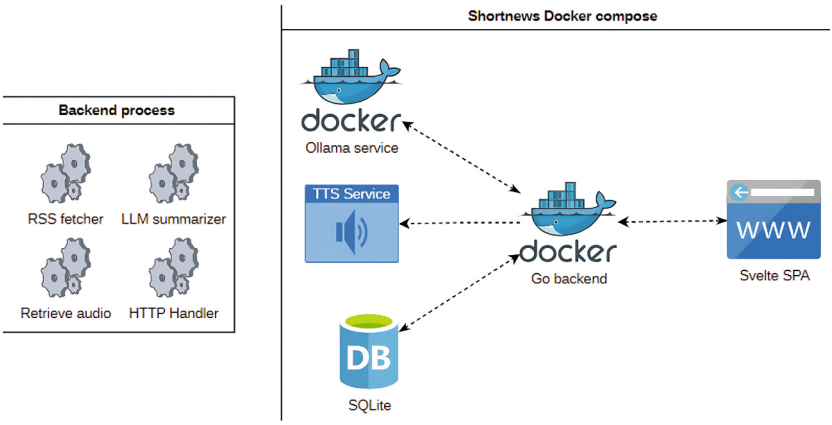

Fig. 11. Architecture of the ShortNews application integrating VIEACT-TTS-VC

E. Web Application

The VIEACT-TTS-VC framework is integrated as a TTS module in a real-world application named *ShortNews*, an RSS-based news reader platform. Users can input RSS feed URLs, and the system automatically crawls, summarizes the content using LLMs, and synthesizes audio via the VIEACT-TTS-VC module. The architecture of this pipeline is illustrated in Fig. 11.

References

1. Allen, J., Hunnicutt, S., Carlson, R., Granstrom, B.: MITalk-79: the 1979 MIT text-to-speech system. J. Acoustical Soc. Am. **65**(S1), S130–S130 (1979)
2. Arik, H.D., Gulsen, N., Hayirli, A., Alatas, M.S.: Efficacy of Megasphaera elsdenii inoculation in subacute ruminal acidosis in cattle. J. Animal Physiol. Animal Nutrition **103**(2), 416–426 (2019)
3. Charpentier,F., Stella, M.: Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In: ICASSP 1986. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 11, pp. 2015–2018 (1986)
4. Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML2017, vol. 70, pp. 933–941. JMLR.org (2017)
5. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using Real NVP. In: International Conference on Learning Representations (2017)
6. Hunt, A.J., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 1, pp. 373–376 (1996)
7. Jia, Y., Zen, H., Shen, J., Zhang, Y., Wu, Y.: PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS. In: Interspeech (2021)
8. Kim, J., Kong, J., Son, J.: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: Meila, M., Zhang, T (eds.) Proceedings of the 38th International Conference on Machine Learning, vol. 139 of Proceedings of Machine Learning Research, pp. 5530–5540. PMLR (Jul 2021)
9. Kong, J., Kim, J., Bae, J.: HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 17022–17033. Curran Associates, Inc. (2020)
10. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: a lite BERT for Self-supervised learning of language representations. In: International Conference on Learning Representations (2020)
11. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
12. Li, Y.A., Han, C., Jiang, X., Mesgarani, N.: Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023)
13. Li, Y.A., Zare, A.A., Mesgarani, N.: StarGANv2-VC: a diverse, unsupervised, non-parallel framework for natural-sounding voice conversion. In: Interspeech (2021)

14. Nguyen, L.T., Pham, T., Nguyen, D.Q.: XPhoneBERT: a pre-trained multilingual model for phoneme representations for text-to-speech. In: Proceedings of the 24th Annual Conference of the International Speech Communication Association (INTERSPEECH) (2023)
15. Nhan, D.T., Tri, N.M., Nam, C.X.: Vietnamese speech synthesis with end-to-end model and text normalization. In: 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), pp. 179–184 (2020)
16. Phung, V.L., Phan, H.K., Dinh, A.T., Nguyen, Q.B.: Data processing for optimizing naturalness of vietnamese text-to-speech system. In: 2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 1–6 (2020)
17. Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., Kudinov, M.: Grad-TTS: a diffusion probabilistic model for text-to-speech. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, vol. 139 of Proceedings of Machine Learning Research, pp. 8599–8608. PMLR (June 2021)
18. 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), pp. 3617–3621 (2019)
19. Ren, Y., Hu, C., Qin, T., Zhao, S., Zhao, Z., Liu, T.-Y.: FastSpeech 2: Fast and High-Quality End-to-End Text-to-Speech (June 2020)
20. Shadle, C.H., Damper, R.I.: Prospects for articulatory synthesis: a position paper. In: 4th ISCA Workshop on Speech Synthesis, August/September 2001 (01/01/02), pp. 121–126 (2002). Address: Pitlochry, Scotland
21. Taylor, P.:Text-to-Speech Synthesis. Cambridge University Press (2009)
22. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech parameter generation algorithms for HMM-based speech synthesis. In: 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), vol. 3, pp. 1315–1318 (2000)
23. van den Oord, A., et al.: WaveNet: a generative model for raw audio. In: 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9), pp. 125 (2016)
24. Weiss, K., Khoshgoftaar, T.M., Wang, D.D.: A survey of transfer learning. J. Big Data **3**(1), 1–40 (2016). https://doi.org/10.1186/s40537-016-0043-6
25. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In: 6th European Conference on Speech Communication and Technology, pp. 2347–2350 (1999)
26. Zhang, G., et al.: Mixed-Phoneme BERT: Improving BERT with Mixed Phoneme and Sup-Phoneme Representations for Text to Speech. In: Interspeech (2022)