TEXT TO SPEECH REPORT

1. Introduction

Text-to-Speech (TTS) technology converts written text into natural-sounding speech. It plays a vital role in accessibility (e.g., screen readers for visually impaired users), entertainment (audiobooks, podcasts), and human-computer interaction (virtual assistants).

For Vietnamese, TTS development is particularly interesting due to linguistic challenges. The language is tonal with six distinct tones (ngang, sắc, hỏi, ngã, nặng, huyền), and words are often compounds with no explicit boundaries. These features make Vietnamese TTS more complex compared to languages like English, but also more valuable for local users.

2. TTS Pipeline

2.1 Input text

- The system begins with raw text provided by the user.
    - Example: “Xin chào, bạn có khỏe không?”

2.2 Text Analysis

- Text Normalization: Expands numbers, abbreviations, and symbols into spoken form.
    - Example: “Dr.” → “Doctor”, “2025” → “hai nghìn không trăm hai mươi lăm”.
- Linguistic Analysis: Breaks down sentences into syntactic and semantic components, helping the system handle punctuation, pauses, and sentence boundaries.

2.3 Phonetic Analysis

- **Grapheme-to-Phoneme (G2P) Conversion**: Maps Vietnamese text into phoneme sequences while encoding tones.
    - Example: “má” (mother, rising tone) vs. “mà” (but, falling tone).

2.4 Prosodic Analysis

- Adds pitch, duration, stress, and intonation to make speech sound natural.
- Example: Rising intonation for questions, longer pauses at commas.

2.5 Acoustic Modeling

- Converts phoneme and prosody information into mel-spectrograms.
- Candidate models:
    - **Tacotron 2**: High-quality, attention-based but slower.
    - **FastSpeech 2**: Non-autoregressive, faster, more stable → **preferred for Vietnamese**.

2.6 Vocoder (Waveform Generation)

- Transforms spectrograms into actual audio waveforms.
- Options: WaveNet, WaveGlow, HiFi-GAN.
- HiFi-GAN is recommended for Vietnamese due to its speed, lightweight design, and high fidelity.

2.7 Speech Output

- The system produces the final speech signal, which can be played back or stored as an audio file.

3. Algorithms & Models

3.1 Text Frontend

- **Normalization & Tokenization**: Rule-based expansion of numbers, abbreviations, segmentation of compound words or Text Normalization Models.
- **G2P Conversion**: Hybrid rule-based + ML approach to map text to phonemes or Vietnamese G2P models.
- **Tone Encoding**: Explicitly represents tones to avoid meaning errors.

3.2 Prosody Modeling

- Uses **FastSpeech 2 predictors** to add pitch, duration, and energy, ensuring natural intonation.

3.3 Acoustic Model

- **FastSpeech 2**: non-autoregressive, efficient, and stable

3.4 Vocoder

- **HiFi-GAN**: Fast, lightweight, and high fidelity

4. Challenges

Developing a Vietnamese TTS system involves several language-specific challenges. The most critical issue is **handling tones**. Vietnamese has six tones, and a single tone change can alter the meaning of a word entirely. This makes tone prediction errors especially harmful, as even minor mistakes can produce confusing or misleading speech. Ensuring tones are represented

accurately during grapheme-to-phoneme conversion and preserved in later modeling stages is therefore essential.

Dialectal variation adds another layer of complexity. Northern, Central, and Southern Vietnamese differ in pronunciation and intonation, and users often expect a system to reflect the variety they are most familiar with. Meeting this expectation requires either separate models for each dialect or a unified multi-speaker system capable of adapting styles while maintaining intelligibility.

Data scarcity is also a significant limitation. Compared to major languages such as English or Chinese, there are far fewer large-scale, high-quality Vietnamese speech datasets. Existing resources like VIVOS or VLSP are helpful but insufficient, making additional data collection and careful cleaning necessary. Techniques such as forced alignment are often required to improve dataset quality and consistency.

Finally, achieving **natural prosody** remains a persistent challenge. Even when tones and phonemes are correct, speech that lacks variation in pitch, rhythm, or energy can sound robotic. Effective prosody modeling is essential to produce speech that not only conveys meaning accurately but also feels expressive and engaging to listeners.

5. Conclusion

This proposal presented a Vietnamese TTS pipeline, covering text normalization, phoneme conversion, prosody modeling, acoustic modeling, and waveform generation. The system relies on a strong text frontend, FastSpeech 2 for efficient spectrogram prediction, and HiFi-GAN for natural audio synthesis.

For Vietnamese, tone accuracy and prosody are the most critical factors. Correct tones ensure intelligibility, while natural prosody makes the voice sound human rather than robotic. By addressing these challenges, the system can deliver clear and natural speech, supporting applications such as accessibility tools, audiobooks, and virtual assistants.