Nonparallel Expressive TTS for Unseen Target Speaker using Style-Controlled Adaptive Layer and Optimized Pitch Embedding

Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh

Budapest University of Technology and Economics

malradhi@tmit.bme.hu

October 27, 2023

SPEECH is not just text that can be heard

SPEECH is not just text that can be heard

- Speech is
  - Style
  - Social
  - Cultural
  - Emotional
  - Memorable
  - Meaningful
  - Dynamic
  - Expressive
  - Prosodic
  - Interactive
  - Linguistic
  - Articulated
  - Adaptive
  - Symbolic
  - Transformative
  - Problem solving

https://www.tmit.bme.hu/node/3418

SPEECH is not just text that can be heard

List of Speech features

Linguistic Parameters

Acoustic Parameters

SPEECH is not just text that can be heard

List of Speech features

Linguistic Parameters

- Phonemes
- Duration
- Pause
- Phrase
- Token
- Intonation
- Lexicon
- Stress

Acoustic Parameters

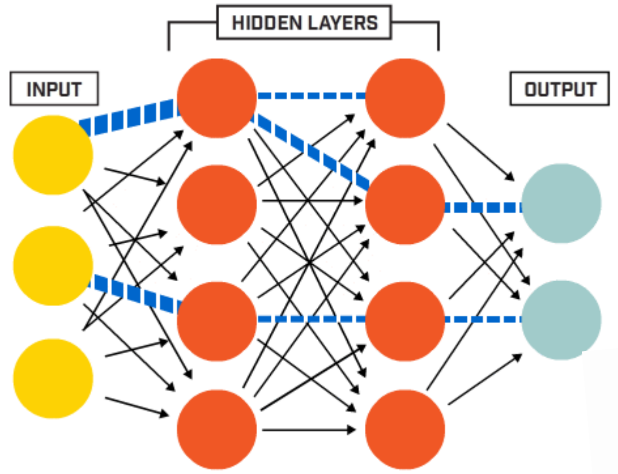
- Fundamental Frequency (F0)
- Spectral Envelope
- Amplitude
- Maximum Voice Frequency (MVF)
- Articulatory Features
- Aperiodicity
- Energy
- Formants

What is “Neural TTS” ?

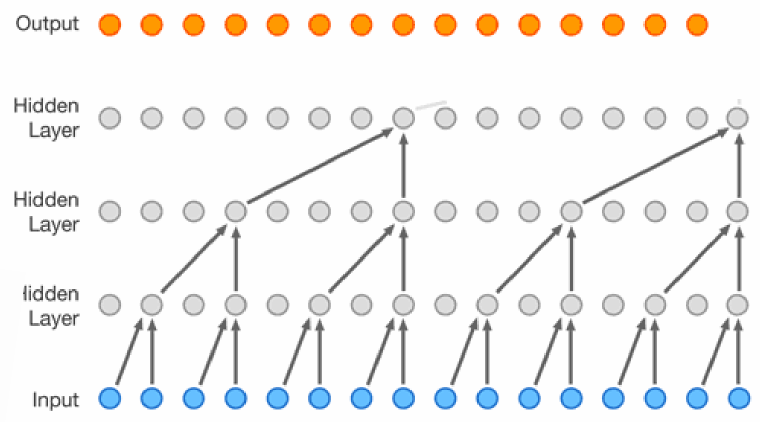
What is "Neural TTS" ?

Neural Text processing

Neural Acoustic model

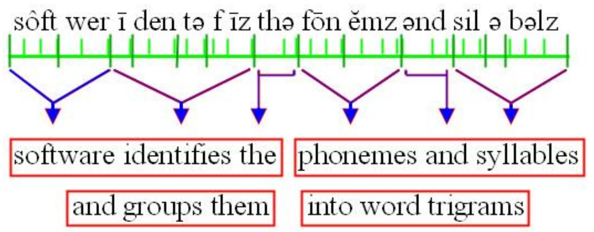


Neural Vocoder

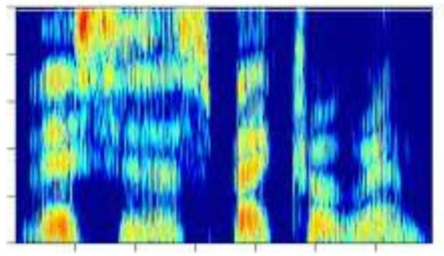


Text

Linguistic features



Acoustic features



Speech

What is “Expressive TTS” ?

What is “Expressive TTS” ?

Expressivity

1. Style
- Reading, news, information providing
- Dialog, Informal & Formal conversation
- Speed

2. Accents
- Native speakers
- Foreign speakers

3. Mood
- Request, Acquisition
- Affirmation, Apology

What is “Expressive TTS” ?

Approaches toward adding expressivity (different emotions and speaking styles) to synthetic speech

- In **linguistics**, expressivity may change the choice of words or syntactic structures.
- In **acoustics**, it impacts various characteristics like energy, pitch, duration, etc.

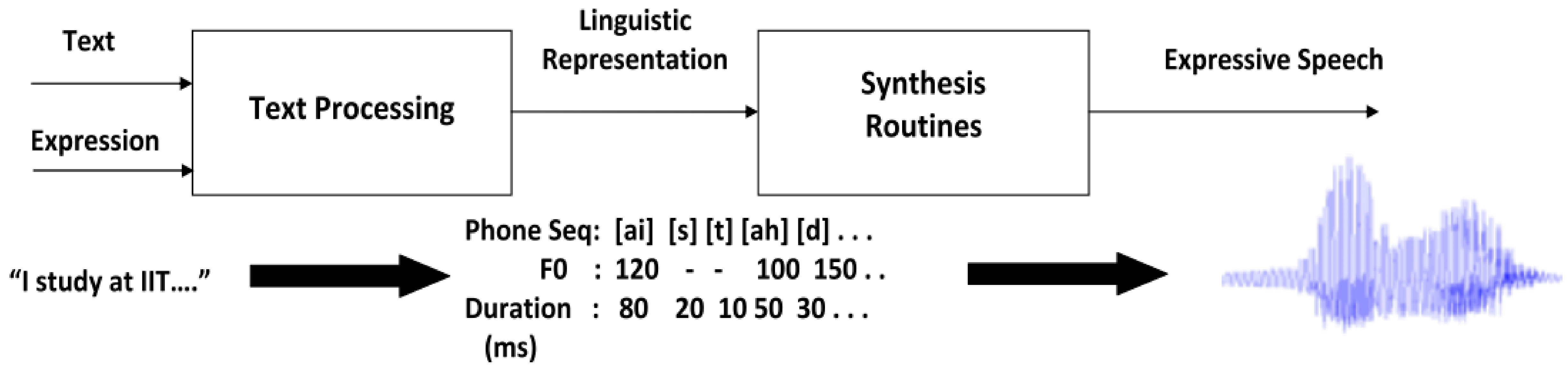
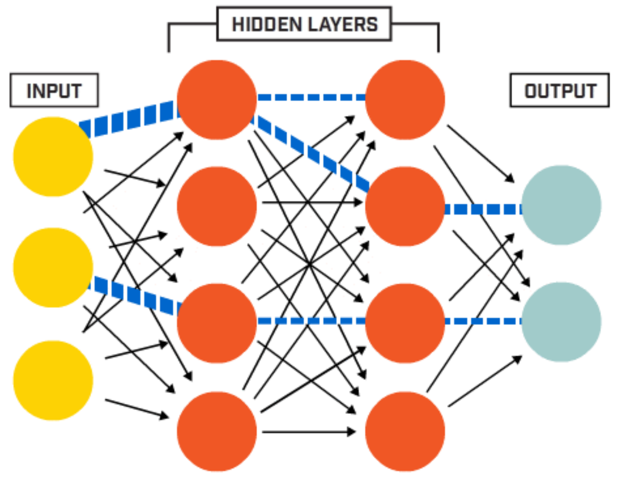


Figure 1: Neural Expressive TTS system **[Govind and Prasanna, 2013]**

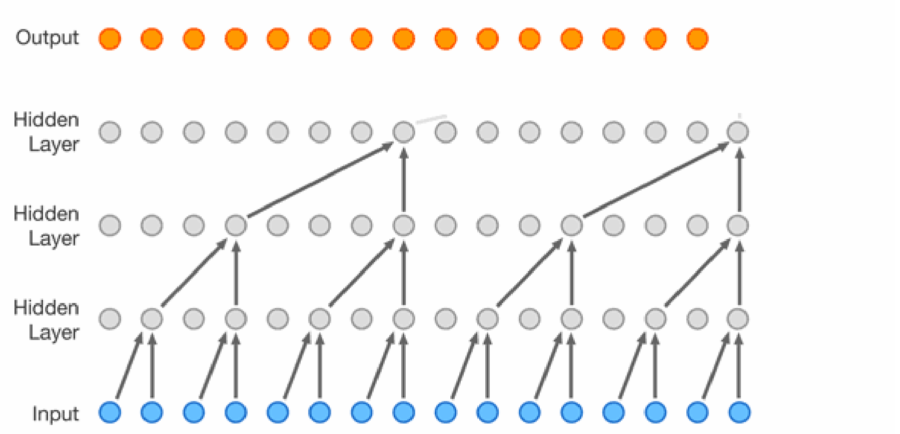
What is “Expressive TTS” ?

Neural Text processing

Neural Acoustic model



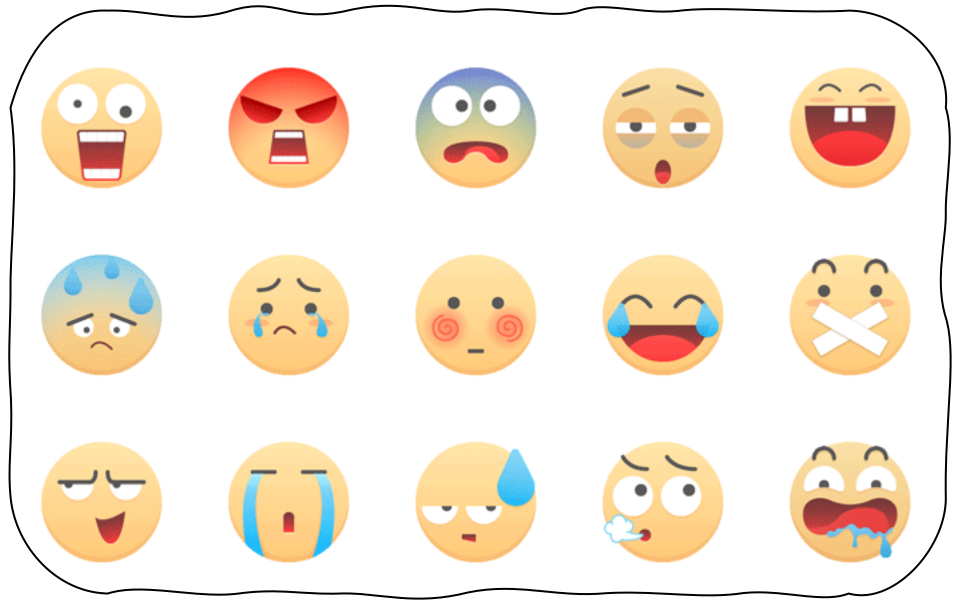
Neural Vocoder



Style

Accent

Speed



Mood

Emotion

Conversation

Text

Speech

Expressive

Limitations of Expressive TTS

1. Weak control over speaking style and difficulty capturing nuanced prosody

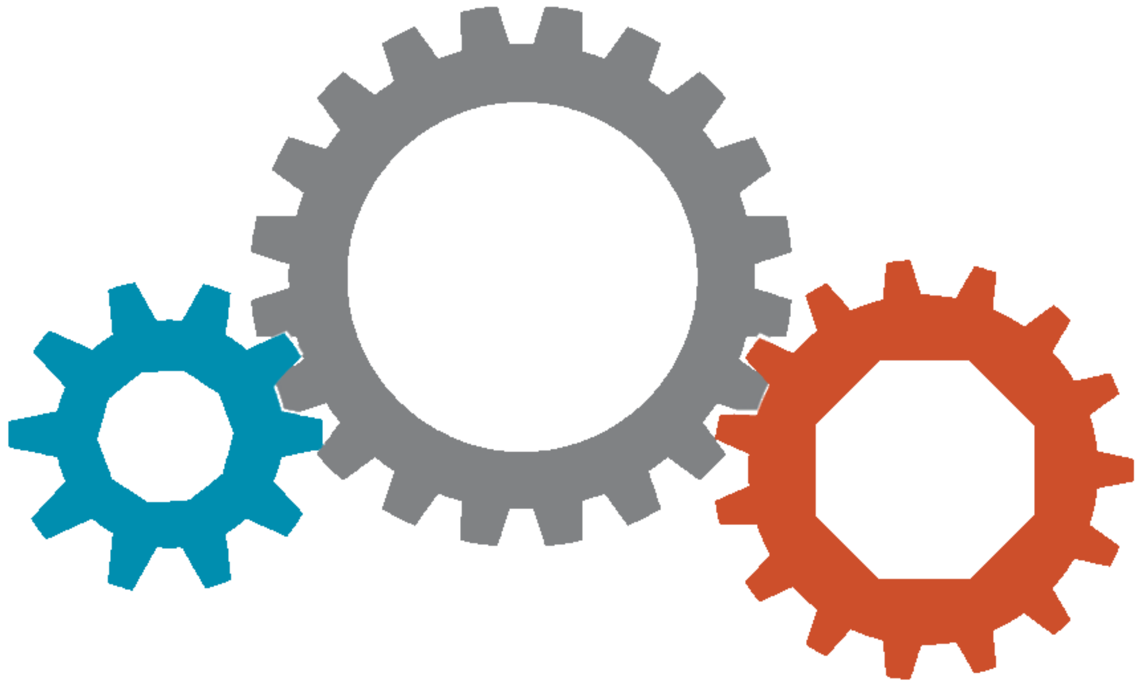
2. Complex training due to the need to disentangle style and content features

- Flexible and appropriate **expressivity in a synthetic voice is still out of reach**:
  - making a voice sound happy, friendly or uncertain is beyond what can be done today.

Research Challenges

1. The challenge is to build a non-parallel, expressive, multi-speaker TTS model
2. Focus on improving training performance while maintaining quality and flexibility

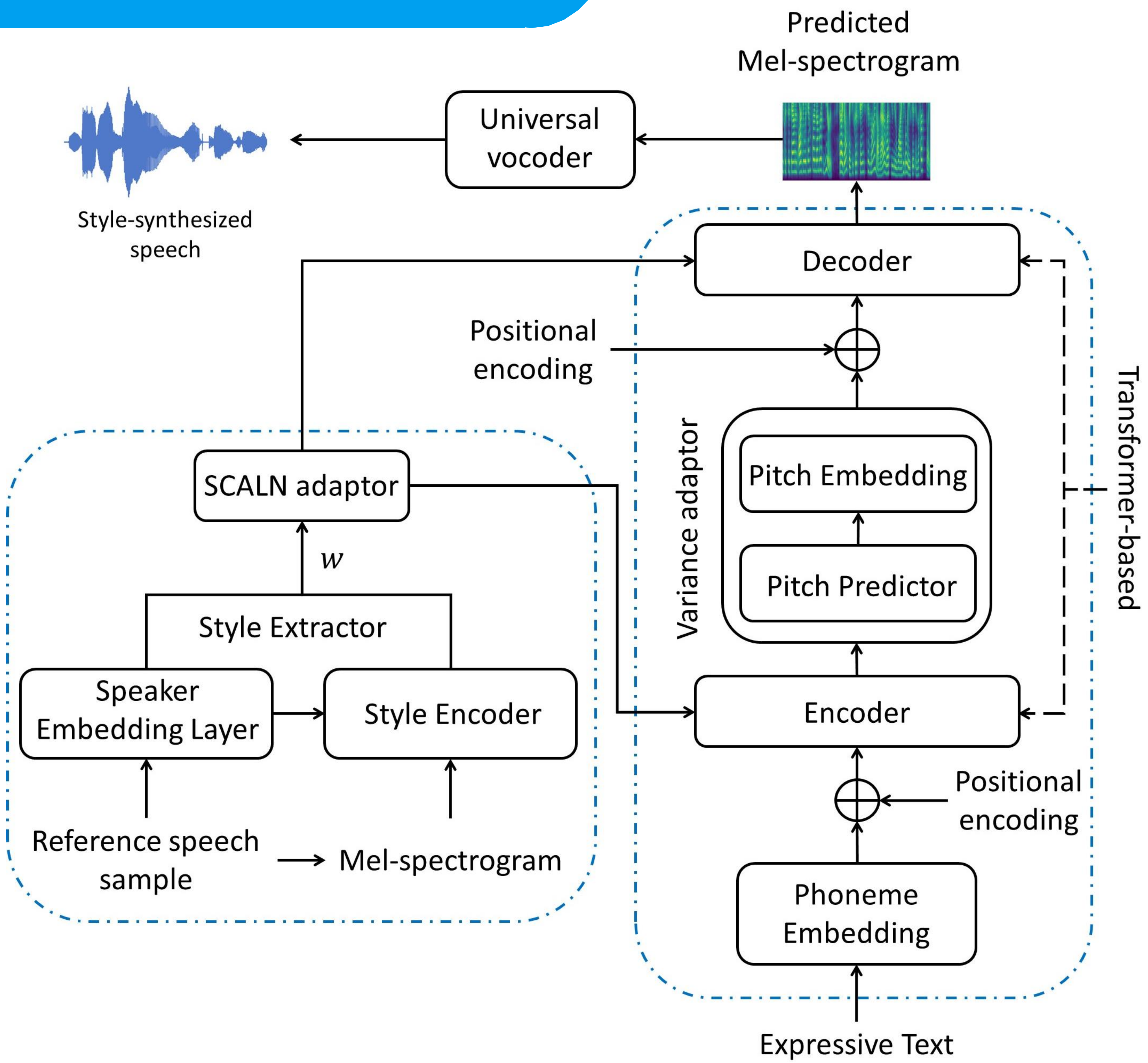
PROPOSED METHODOLOGY



Overall proposed model

The model architecture employs a transformer and includes:

1. phoneme encoder to map phoneme sequences to high-level representations, capturing linguistic nuances.
2. style encoder is used to extract style attributes from reference speech, including speaker identity and prosody
3. compute a style vector, utilizing the Mel-spectrogram and speaker-specific information
4. universal vocoder to synthesize speech from predicted Mel-spectrograms
   - speaker encoding is not required to train a Speaker-Independent (SI) WaveRNN-based neural vocoder



1. Style-Controlled Adaptive Layer Normalization

1. SCALN is introduced to enhance the controllability and stability of training in style TTS systems.
2. SCALN employs an adaptive layer normalization mechanism, which includes normalizing input features and adaptively scaling and shifting them based on style information.
3. The input feature vector $x = (x_1, x_2, \ldots, x_N)$ is normalized $y = (y_1, y_2, \ldots, y_N)$ to have tensor with zero mean and unit variance using mean and standard deviation statistics.

$$y = \frac{x - \mu}{\sigma} \qquad (1)$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2} \qquad (2)$$

1. Style-Controlled Adaptive Layer Normalization

4. Bias and gain parameters for normalization are learned through an affine transformation (sometimes called as linear layer or fully connected layer), using a style embedding as input
5. The bias and gain are then used to scale and shift the normalized features, resulting in the output tensor

$$SCALN(x, w) = g(w).y + b(w) \qquad (3)$$

6. Dropout and gradient clipping are applied to ensure training stability
7. Regularization loss is incorporated to prevent overfitting and promote generalization
8. Weight decay is a common form of regularization

Comparison of normalization techniques

	Layer Normalization (LN) [27]	Style-Adaptive Layer Normalization (SALN) [10]	SCALN (proposed)
Normalizes input features	subtracting the mean and dividing by the standard deviation	extends LN by introducing an affine transformation controlled by an external weight vector	incorporates a learned affine transformation controlled by a style embedding
Stable/Adaptability	provides stable activation distributions during training	limited adaptability to different speech styles	employs dropout and gradient clipping to prevent overfitting, ensuring better control and stability in training

[27] J. Ba, J. Kiros, and G.E. Hinton, "Layer normalization," ArXiv, abs/1607.06450, 2016.
[10] M. Dongchan, B.L. Dong, Y. Eunho, H. Sung, "Meta-StyleSpeech : Multi-Speaker Adaptive Text-to-Speech Generation," in Proc. International Conference on Machine Learning (ICML), pp. 7748-7759, 2021

2. Optimized pitch control embedding

Pitch Embedding: definition

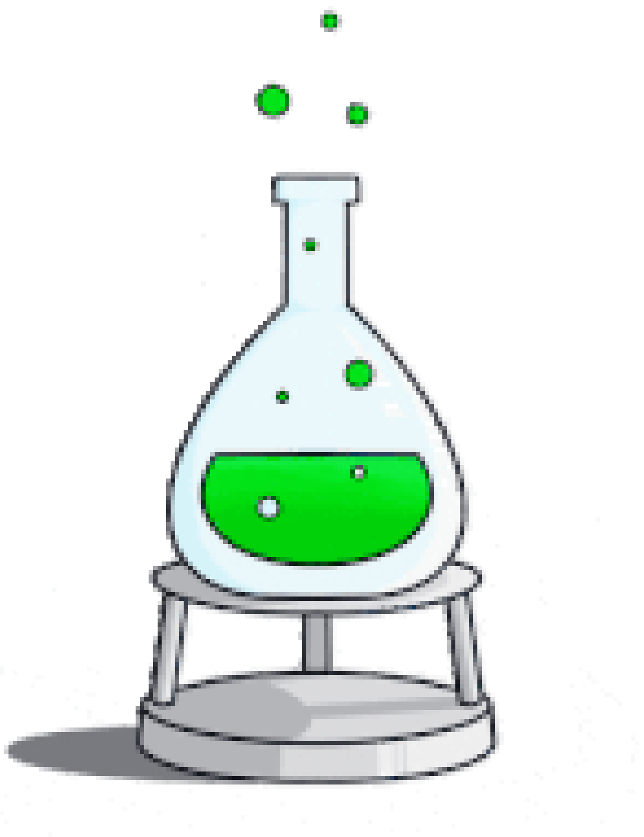
- capture and represent pitch information in a more continuous and nuanced manner
- Instead of discretizing pitch values into buckets, it works with continuous pitch curves.
- predicting pitch values as a function of time, resulting in a curve that can represent variations in pitch more accurately.
- types of embeddings, such as continuous embeddings and one-hot encodings, to provide more flexibility in controlling pitch.

One-Hot Encoding

- removal of the bucketization step (supporting only a fixed range of pitch values and assuming evenly spaced buckets) → improves accuracy and robustnes
- each pitch value at a given time is represented by a vector in which only one element is set to 1, and the rest are set to 0.

$$P(t) = [p_1(t), p_2(t), \ldots, p_N(t)] \tag{6}$$

EXPERIMENTAL SETTING AND EVALUATION



Setup

1. **Data Source**
   - multi-speaker LibriTTS dataset
   - 110 hours of audio from 1151 speakers, along with corresponding text transcripts
   - Data is split into training (80%), validation (10%), and test (10%) sets

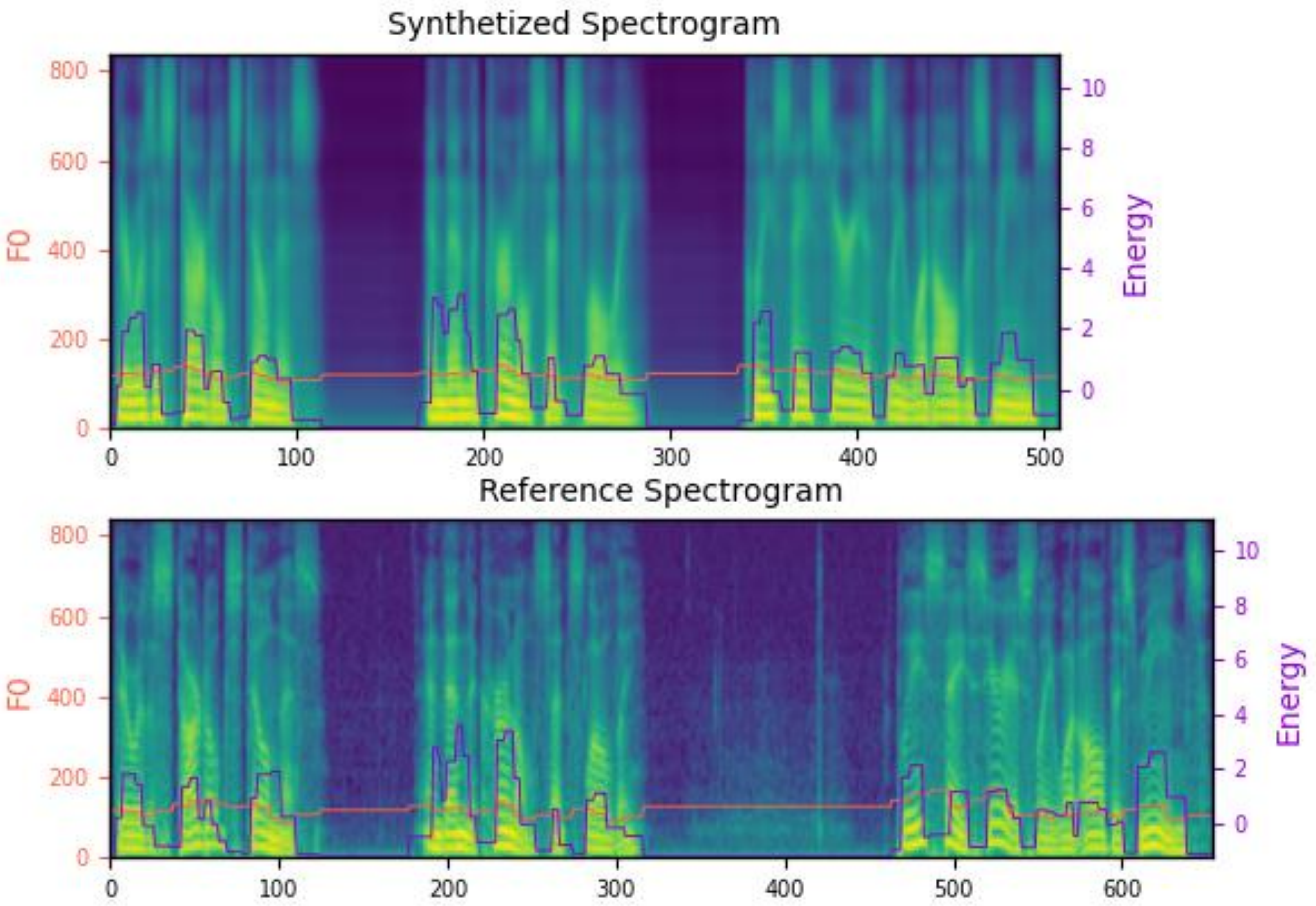
2. **Preprocessing**
   - converting the sampling rate to 16kHz, extracting Mel spectrograms

3. **Training Details**
   1. trained for 200,000 steps with a minibatch size of 48
   2. Adam optimizer
   3. universal vocoder

4. **Comparison**
   - ESPnet2-VITS and CFS2-PWGAN

Evaluation Metric

- Orange line: Pitch contour
- Purple line: Energy
- Sample text: "well, you're not so good looking," spoken by a female speaker
- Model evaluated on a different style of the same speaker.
- Detailed frequency information
- Style closely aligns with reference pitch contours.
- Improves natural sound quality



Evaluation Metric

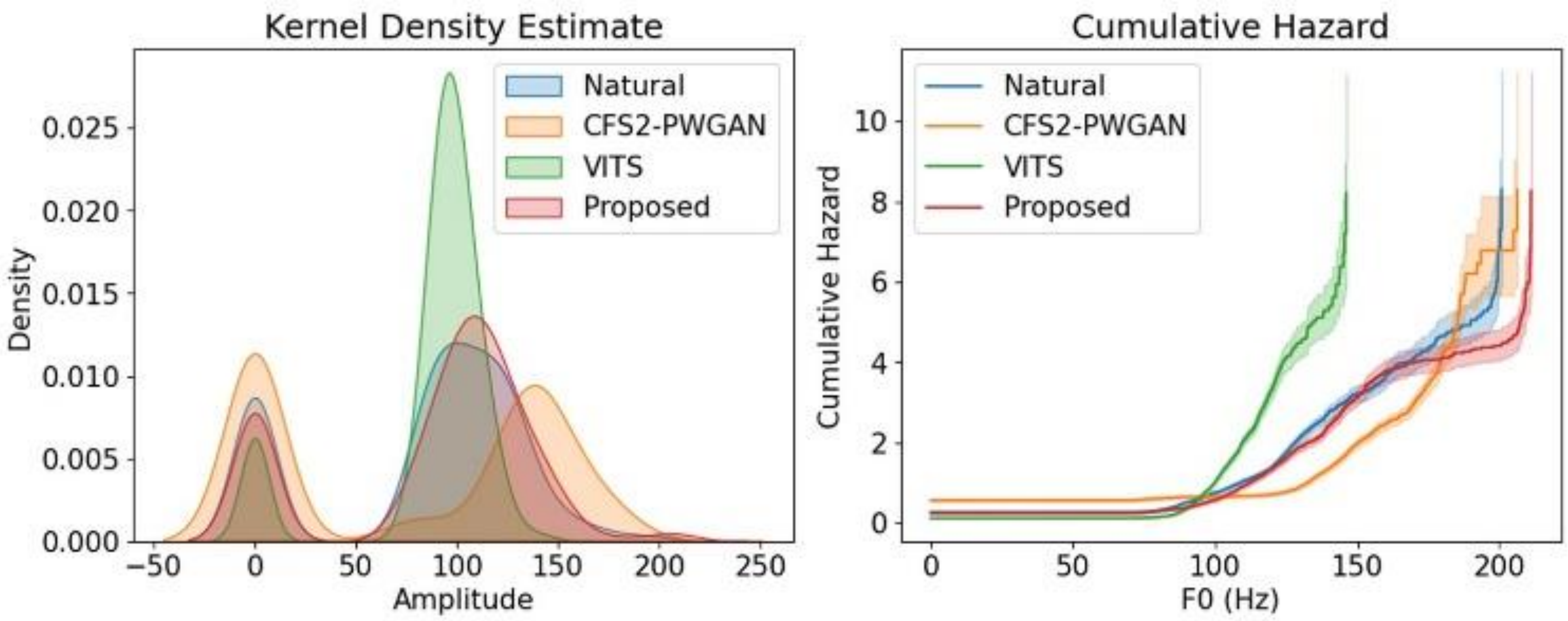
TABLE I. AVERAGE SCORES PERFORMANCE OF SYNTHESIZED SPEECH SIGNALS. THE BOLD FONT SHOWS THE BEST PERFORMANCE.

Model	MCD		F0-RMSE	
	Female	*Male*	*Female*	*Male*
CFS2-PWGAN	5.14	5.00	11.53	10.47
VITS	5.22	4.97	11.44	**10.05**
Proposed	**5.08**	**4.94**	**11.37**	10.26

- VITS better in F0-RMSE for male speakers due to training on a larger dataset with more male voices.

Pitch Control Assessment

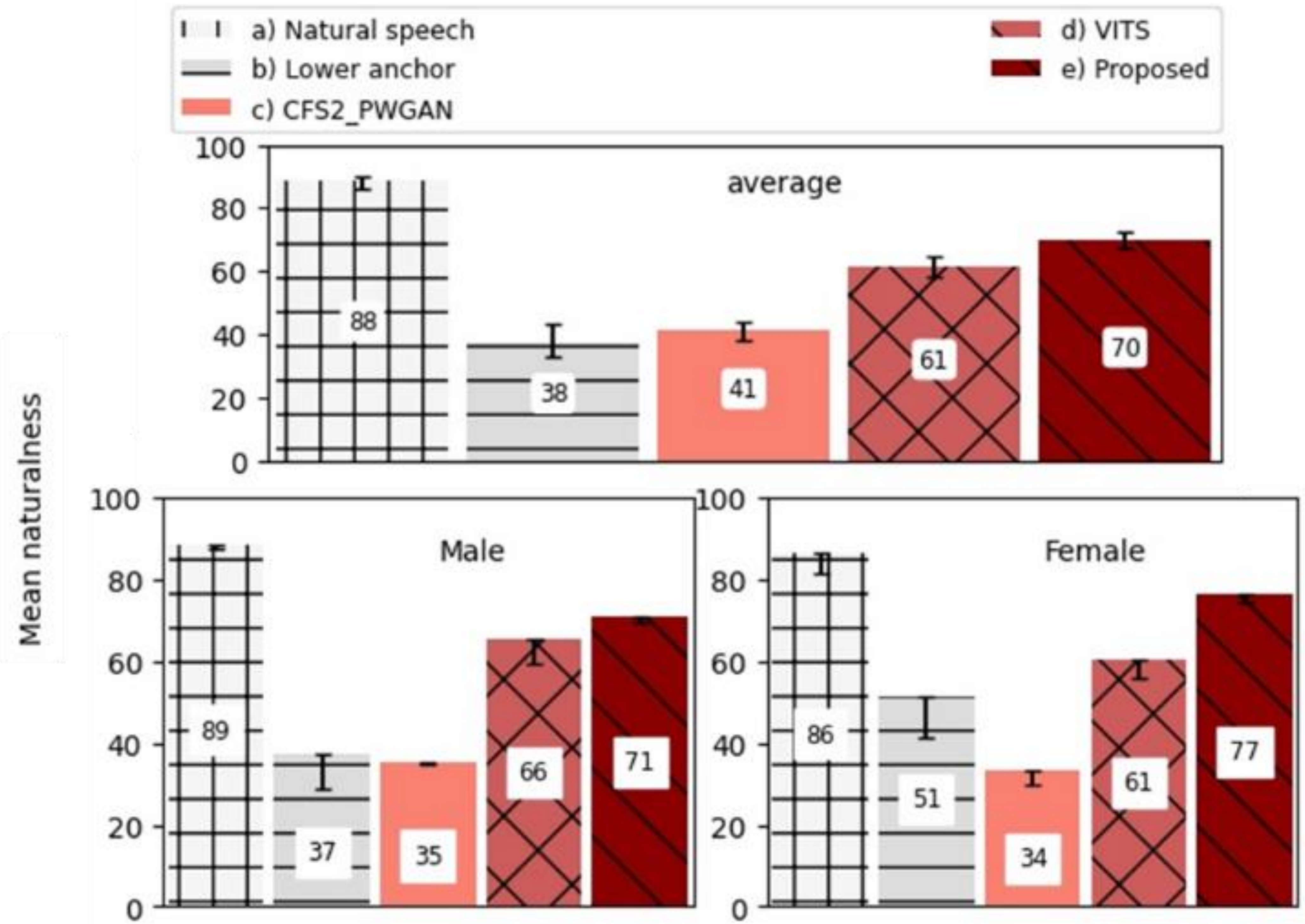
- **Kernel Density Estimate:** Analyzed pitch value distribution for each system.
- **Cumulative Hazard Function:** Evaluated the risk associated with pitch values.



Subjective listening test

- 15 participants (9 males, 6 females)
- 200 utterances included in the test
- Female speakers received higher naturalness scores than male speakers, possibly due to gender-related variations in speaking style

❑ Samples



Demo Samples

Natural | **CFS2_PWGAN** | **VITS_ESPnet2** | **Proposed**

Text: "and the older one answered back, "well, you're not so good looking" (which was also true)"

Conclusion and Future Works

Key Achievements

1. introduced a non-autoregressive non-parallel expressive TTS framework designed for multi-speaker reading-style speech.
2. conducted experiments on an English reading-style corpus, demonstrating superior speech quality and expressiveness compared to baseline models.
3. the model successfully synthesized correct, natural, and expressive speeches based on contextual information from synthetic datasets.

Limitations and Future Work:

1. the need to improve inference speed
2. construct a large-scale dataset with multi-lingual speakers for training
3. extend the applicability of the method to synthesizing spontaneous speech

Thank you for your attention

malradhi@tmit.bme.hu