MŰEGYETEM 1782

# Advancing Limited Data Text-to-Speech Synthesis: Non-Autoregressive Transformer for High-Quality Parallel Synthesis

Mohammed Salah Al-Radhi, Omnia Ibrahim, Ali Raheem Mandeel, Tamás Gábor Csapó, Géza Németh

Budapest University of Technology and Economics, Saarland University

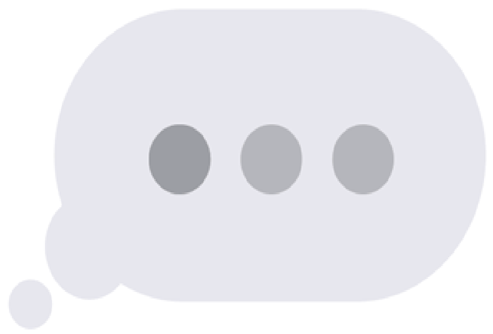
malradhi@tmit.bme.hu

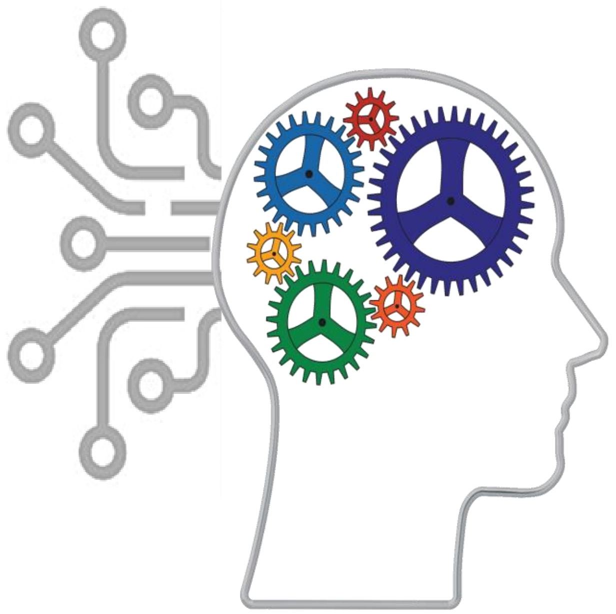
SpeD 2023
Bucharest, Romania
October 25-27, 2023

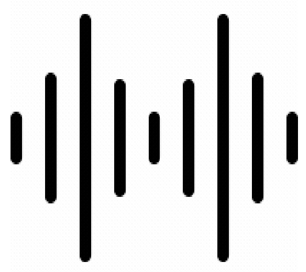
October 27, 2023

# Why Use Text to Speech?

- ❑ producing synthesized speech from text - revolutionary applications
- ❑ help people with visual impairments
- ❑ connect with users at a different platform

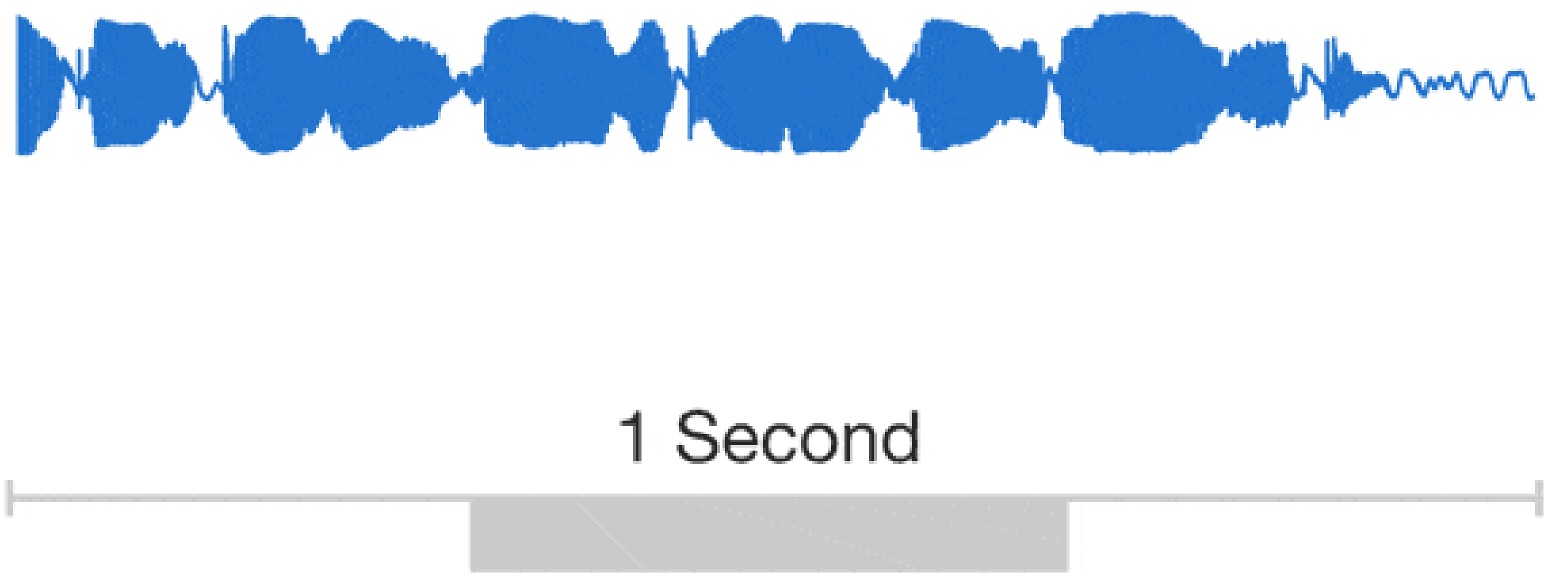






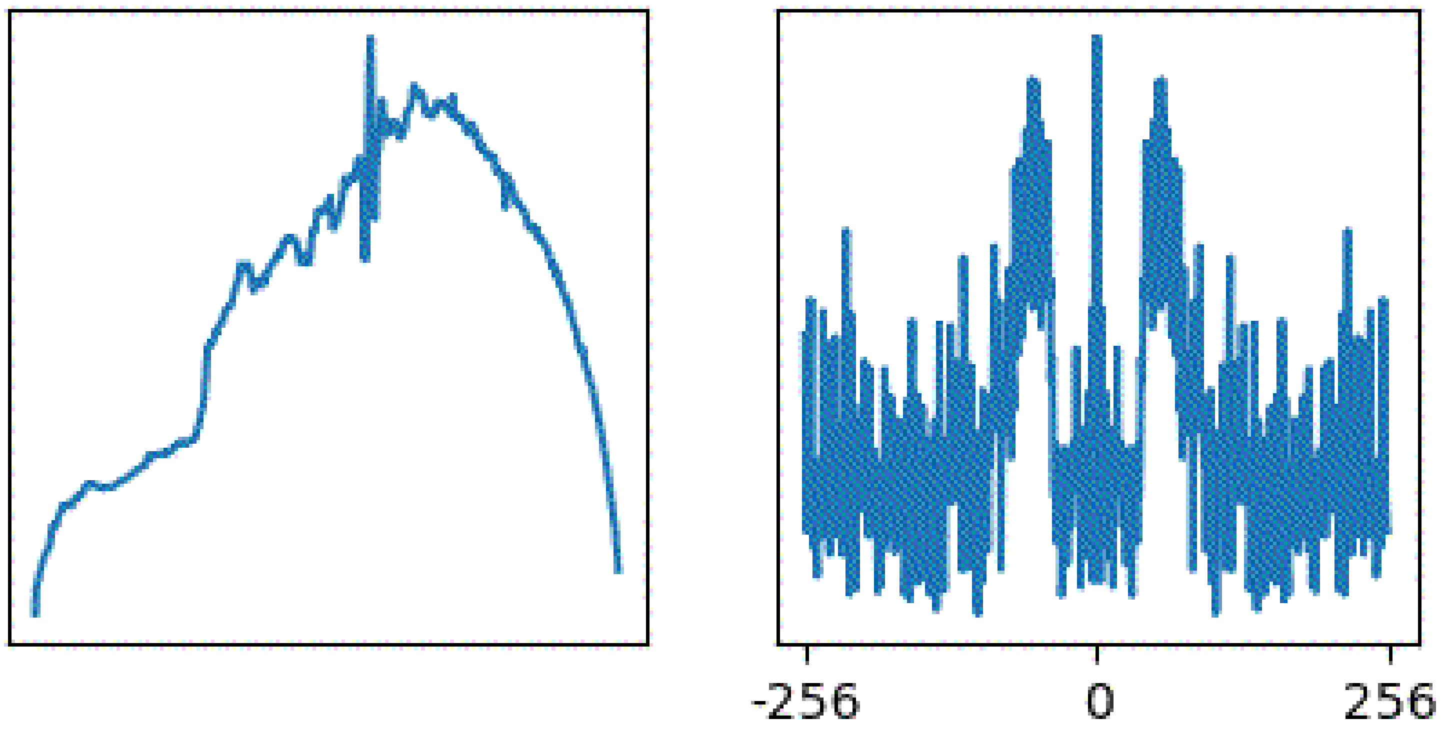
# Successes of Autoregressive Models

- generates speech one element at a time, character by character or phoneme by phoneme
- predicts future values from past values
- can model multiple time scales



# Limitations of Autoregressive Models

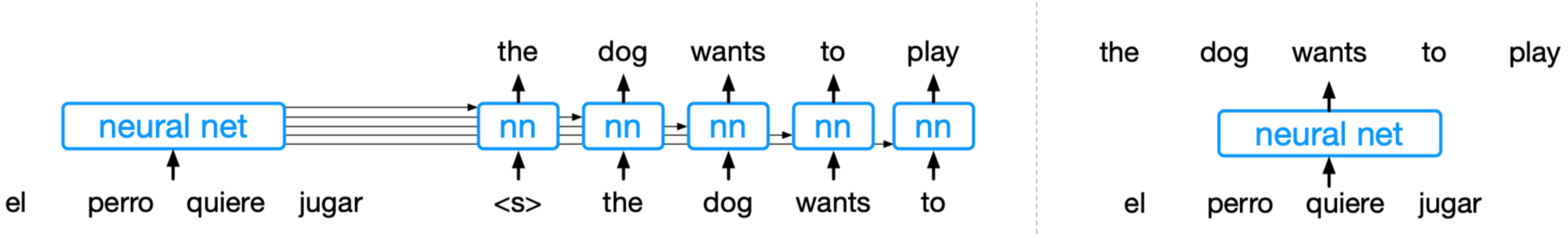
- Slow Inference Speed → time-consuming
- Parallelization Challenges → limiting efficiency
- Error Propagation → affecting overall quality



# Non-Autoregressive Models

- generate all the tokens in parallel by removing the sequential dependencies within the target sequence

- Generation $P(\mathbf{Y}) \quad \mathbf{Y} = y_1, y_2, \cdots, y_T$
- Autoregressive $$P(\mathbf{Y}) = \prod_{i=1}^{T} P(y_i|\mathbf{Y}_{<i})$$
- Non-autoregressive $$P(\mathbf{Y}) = \prod_{i=1}^{T} P(y_i)$$




autoregressive non-autoregressive

# Challenges in Arabic TTS Synthesis

# Challenges in Arabic TTS Synthesis

- **Phonology**
  - different consonants, vowels, and diacritics
- **Morphology**
  - difficult to analyze and generate words (so many morphology)
- **Dialects**
  - Arabic is spoken in many countries with differences in pronunciation, vocabulary, grammar
- **Text Normalization**
  - different writing styles, variations in the representation of characters, or in spelling

# The Need for Faster and Higher-Quality Arabic TTS

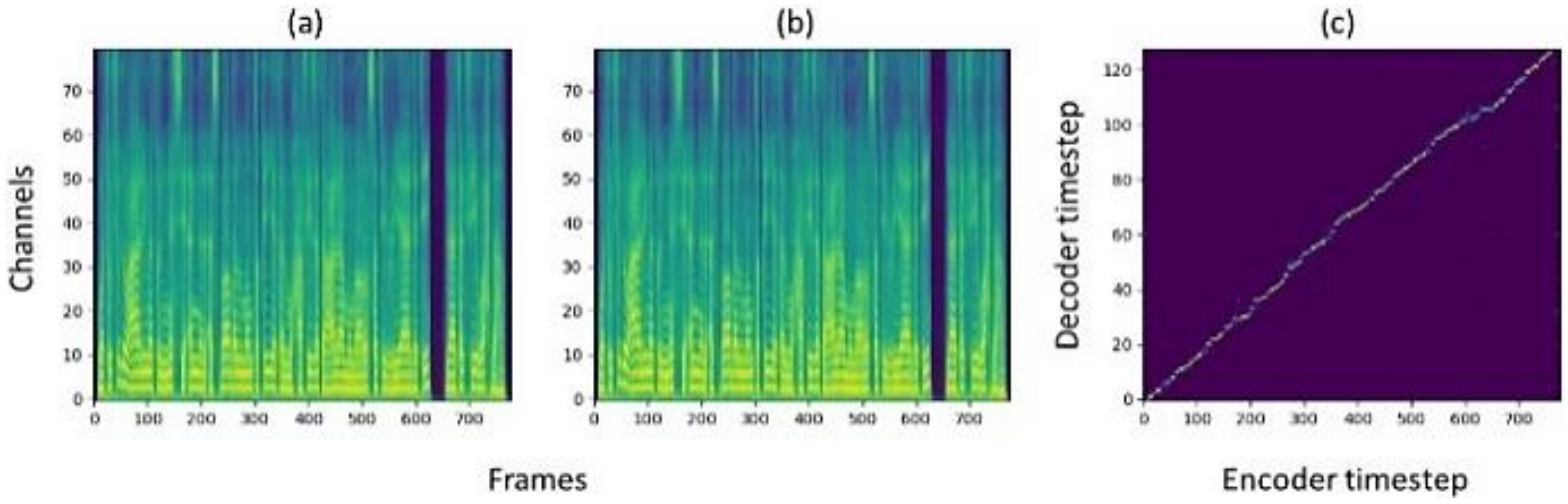
- present a novel non-autoregressive TTS model, that offers faster inference
- enhances alignment and fidelity in synthesized speech
- reducing errors
- improved speech quality

# PROPOSED METHODOLOGY

# 1) Few-Shot Adaptation of Tacotron2

1. **Speaker Adaptation**
   - easily adapted to a speaker with limited training data
   - requiring only 400 sentences (57 minutes) of Arabic speech data
2. **Model Components**
   - encoder that maps diacritic Arabic text to a fixed-dimensional state vector
   - attention-based decoder for predicting an 80-dimensional Mel spectrogram
   - integrate convolutional layers to capture local spectral patterns
   - residual connections within the component to facilitate gradient flow
3. **Waveglow Integration**
   - flow-based implementation using only a single network, trained using only a single cost function
   - enables real-time inference speed, and enhancing the generation of natural-sounding speech

# 1) Few-Shot Adaptation of Tacotron2




1. The spectrogram and alignment results for the Arabic orthographic-transcript: *watu&ak~idu EaAlimapu Aln~afosi >an~a Alo>asobaAba AlomanoTiqiy~apa AlomuHaf~izapa EalaY mumaArasapi Alr~iy~aADapi - laA takofiy waHodahaA*. a) The Mel spectrogram of the ground-truth, b) The Mel-spectrogram of the synthetic speech, c) the attention alignments between the steps of the encoder and decoder.

# 2) Parallel Transformer-based FastSpeech2

## Baseline:

**1. Encoder**

- generates phoneme-level hidden features

**2. Variance Adapter**

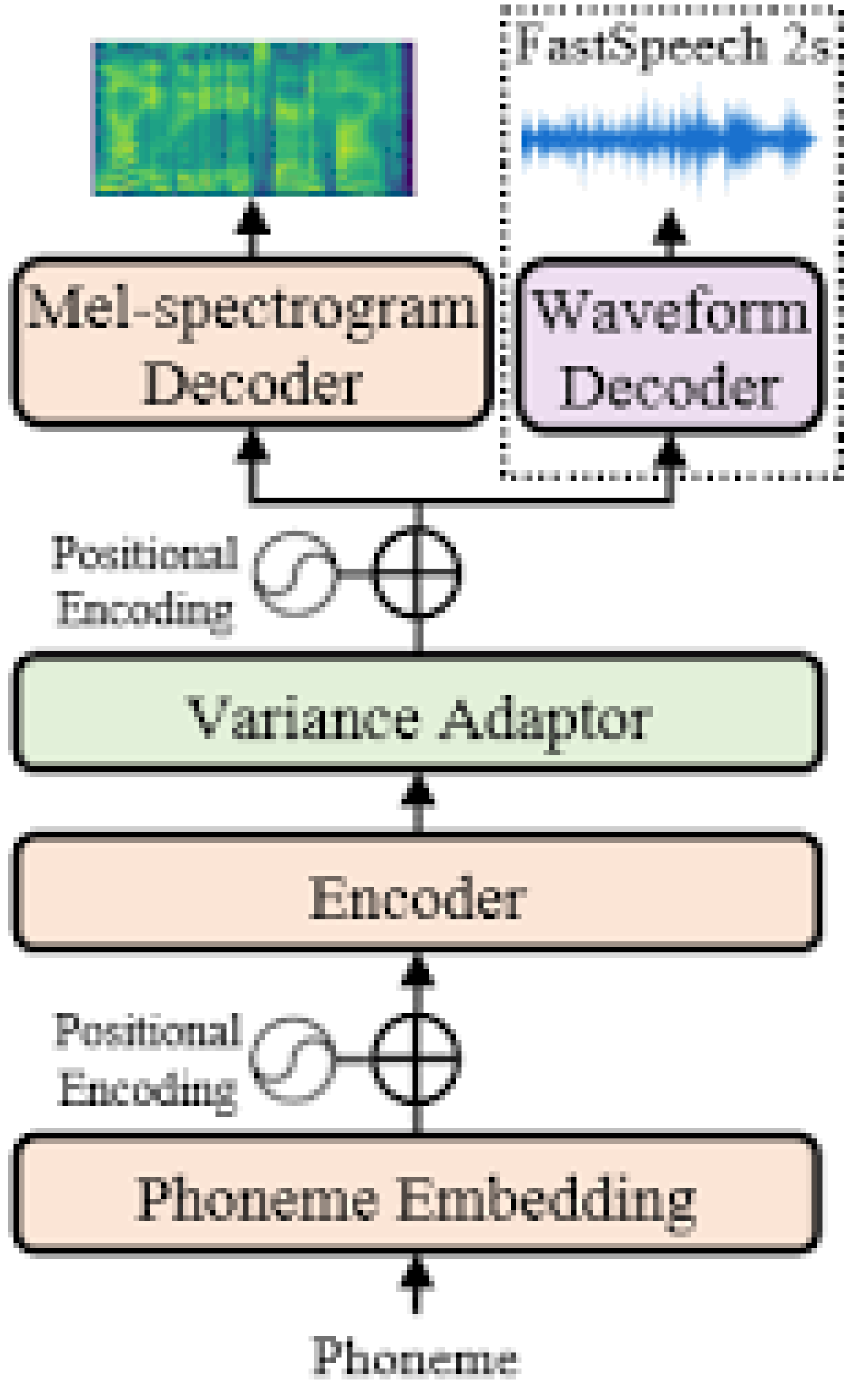
- includes pitch, energy, and duration predictors to control prosody features

**3. Decoding**

- generates Mel-spectrograms

**4. Vocoder**

- HiFi-GAN enhances speech quality, resolution, and fidelity



# 2) Parallel Transformer-based FastSpeech2

## Alternative Architecture:

**1. Positional Encoding**:

- sine/cosine functions is added to the input embeddings and hidden states
- helps the model capture temporal order, capturing prosody, and process speech data effectively

**2. Parallel WaveGAN:**

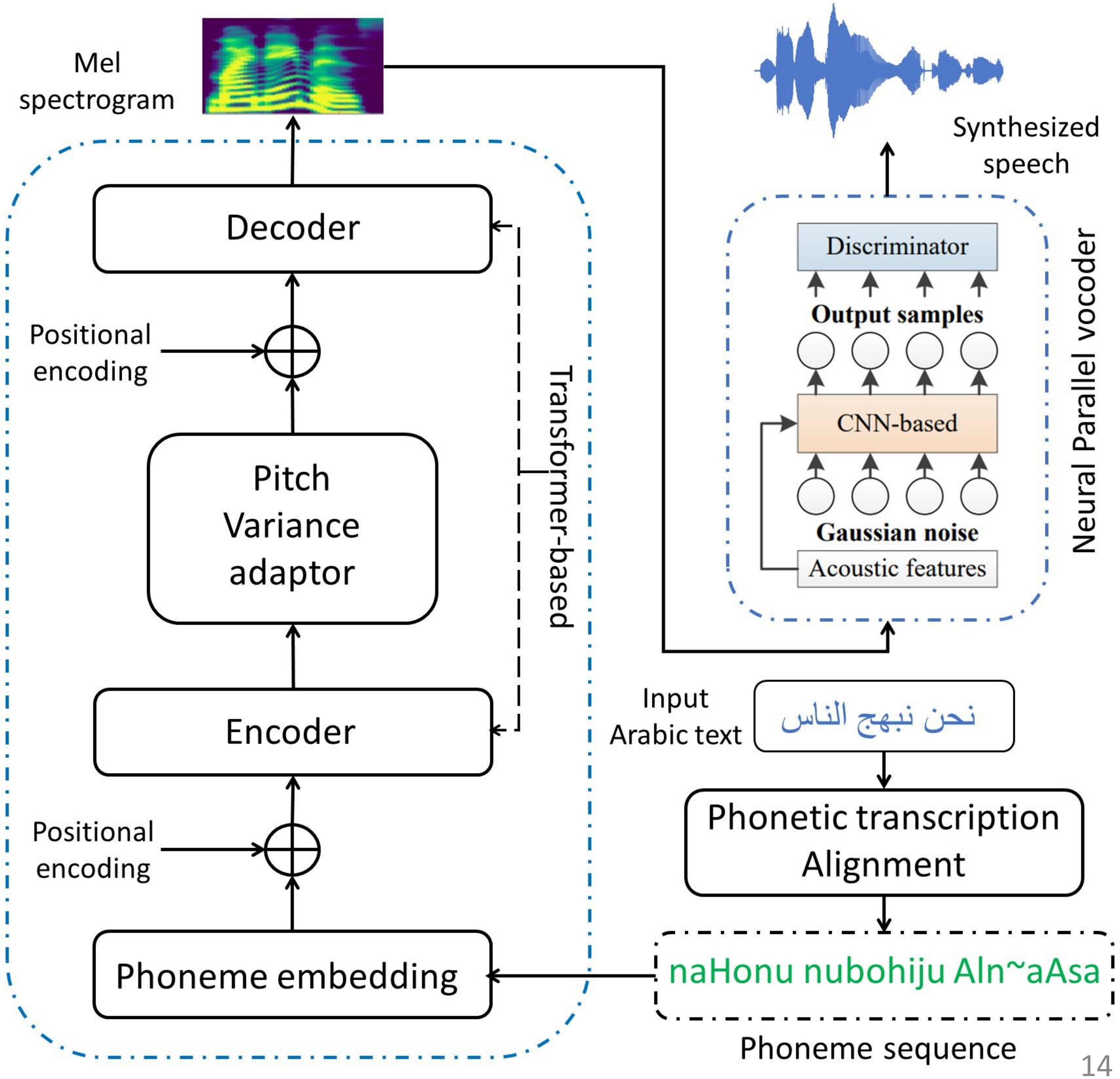
- PWGAN is used for non-autoregressive, non-causal waveform generation
- employs adversarial loss and multiresolution STFT loss to generate realistic waveforms
- avoidance of Artifacts and Word Skipping
- doesn't rely on a teacher-student framework, reducing training and inference time

# 2) Parallel Transformer-based FastSpeech2

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

- where $pos$ represents the position of the frame in the input sequence, $i$ represents the dimension index of the PE vector, $d_{model}$ is the dimensionality of the model.



# EXPERIMENTAL SETUP

# Corpus

1. **Arabic Speech Corpus** [21]
   - single male speaker
   - approximately 2.41 hours
   - 1813 spoken wave files, text utterances, and phoneme labels

2. **Data Preprocessing**:
   - all Arabic characters are replaced with the corresponding Unicode character symbols

3. **Training Data Subset**:
   - 400 sentences, divided into a 90% training set, a 5% test set, and a 5% validation set

TABLE I. ARABIC CHARACTERS REPRESENTATION BY UNICODE CHARACTER SYMBOLS.

| Arabic characters | Unicode character symbols |
|---|---|
| صباح الخير | SabaAHu Aloxayoro |
| نحن نبهج الناس | naHonu nubohiju Aln~aAsa |
| وننقل تراث الاباء والاجداد | Wananoqulu turaA^a Alo\|baA'i waAlo>ajodaAdi |

[21] N. Halabi and M. Wald, "Phonetic inventory for an Arabic speech corpus," in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 734–738, 2016.

# Data Analysis and Acoustic Features

- Acoustic features: F0 contour, F0 mean, F0 standard deviation, F0 coefficient of variability, and F0 slope

- F0 variability, calculated as (F0 standard deviation) * 100 / (F0 mean)

- The F0 distribution of 1813 utterances in the training corpus was visually analyzed, providing insights into pitch characteristics and guiding modeling decisions.

# Data Analysis and Acoustic Features

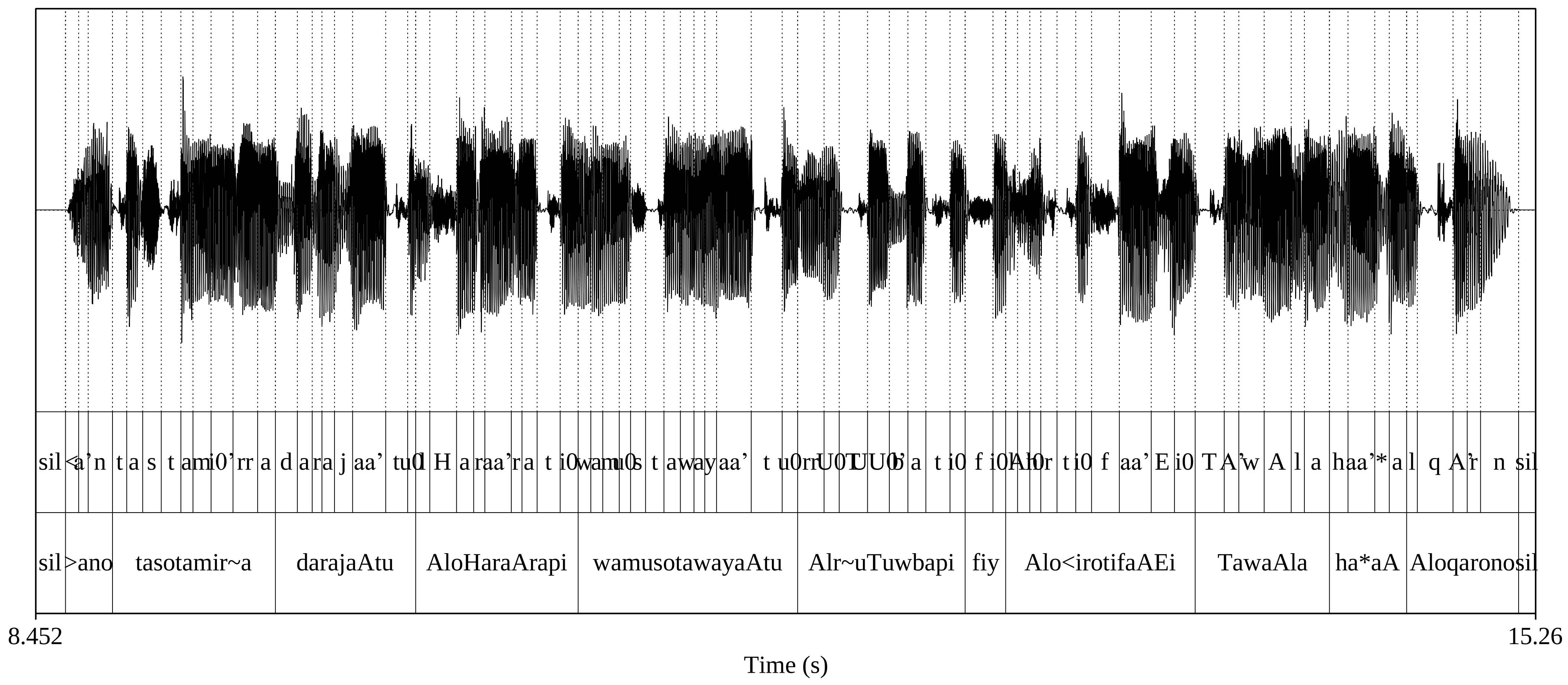



Fig. 3: Example of the recorded utterance with the combined annotation; the first tier is phonemic transcription, and the second tier is word segmentation.

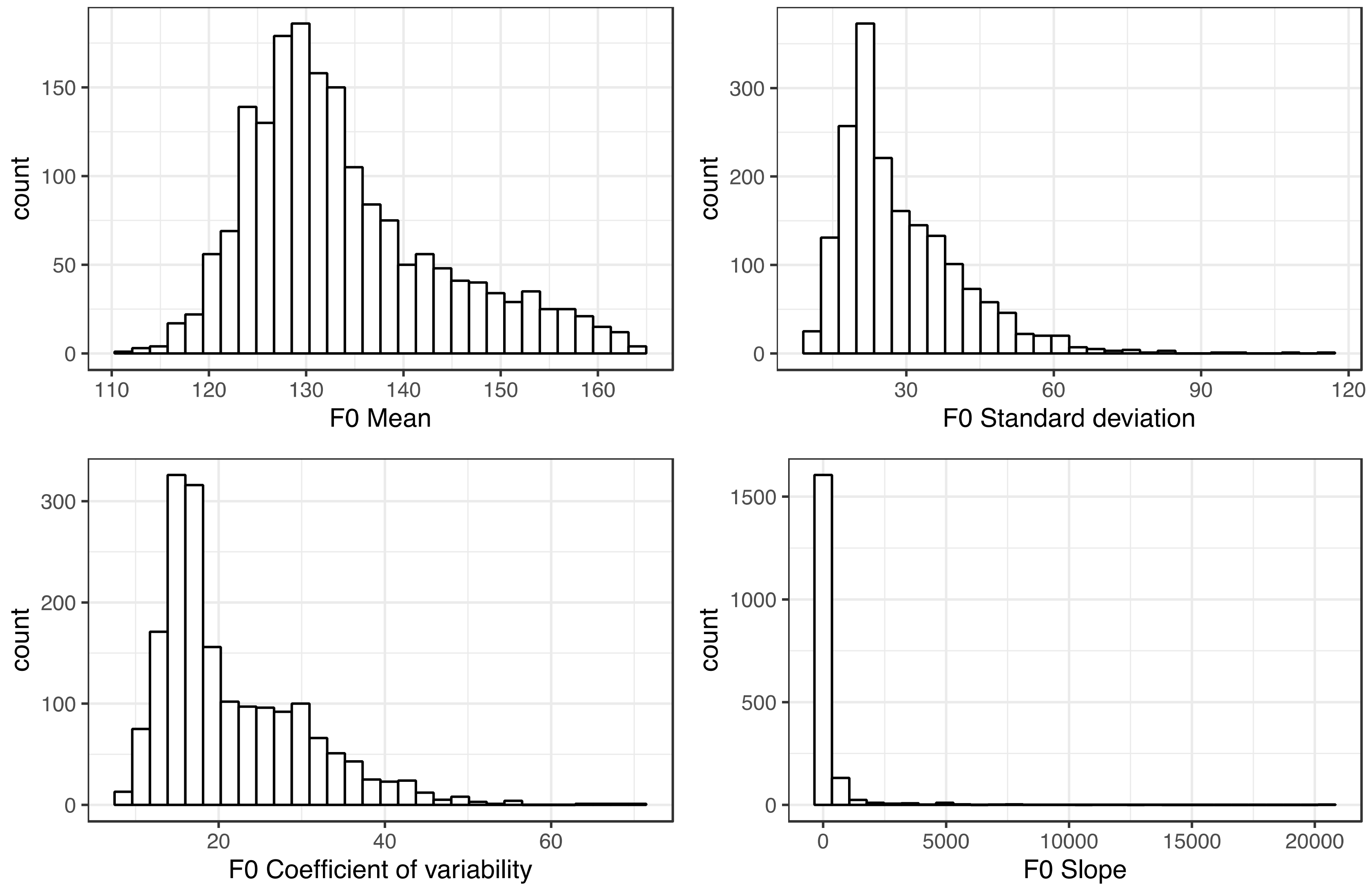
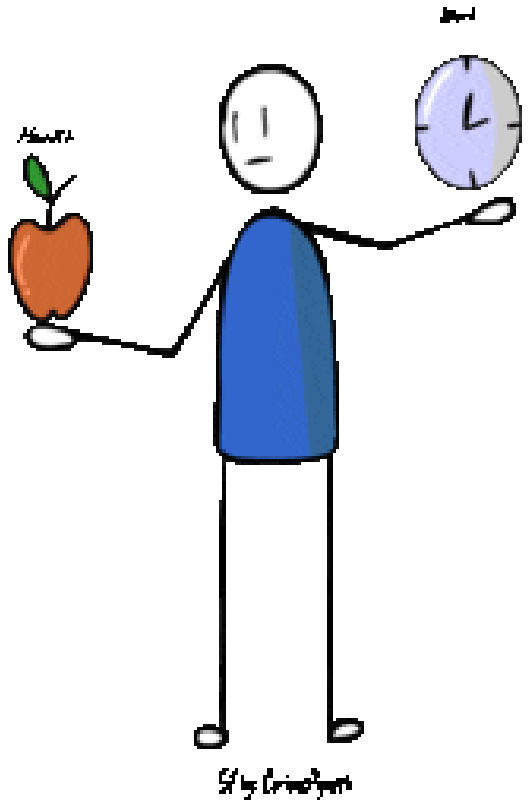



Fig. 4: F0 distributional analysis of the training corpus.

# Models Configuration

1. Limited adaptation data, divided into 90% training data (400 sentences) and 5% validation test data (20 sentences).
2. using the Adam optimizer, 100 iterations per checkpoint, batch size of 4, and checkpoints at 100
3. weight decay of 0.000001, learning rate of 0.001, frame size of 1024, and a hop length of 256
4. Training utilized an NVIDIA Titan X GPU.
5. Arabic character sequences were encoded as 512-dimensional character embeddings for the Tacotron2 encoder.
6. Parallel WaveGAN hyperparameters followed configurations from a reference paper [14].

# RESULTS AND EVALUATION



# Metric evaluation

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

| Model | *MCD* | *NCM* | *SNRseg* | *# Paras.* | *MosNet* |
|---|---|---|---|---|---|
| Taco2-Glow | 5.20 | - | 3.61 | 91 M | 2.63 |
| FastSp2-HiFi | 3.45 | 0.21 | 4.72 | 27 M | 2.91 |
| FastSp2-PWG | **1.55** | **0.93** | **13.89** | **16 M** | **3.48** |

- Proposed system outperformed other methods, exhibiting reduced Mel Cepstral Distortion (MCD), indicating improved speech quality.
- Tacotron2 with WaveGlow (Taco2-Glow) also achieved acceptable results in comparison to non-autoregressive models using the full dataset.
- A computational complexity analysis showed that FastSp2-PWG struck a balance between complexity and speech quality, indicating efficiency in synthesis.

# Metric evaluation

- ❑ proposed system matches the ground-truth pitch, while Tacotron and HiFi have noticeable differences

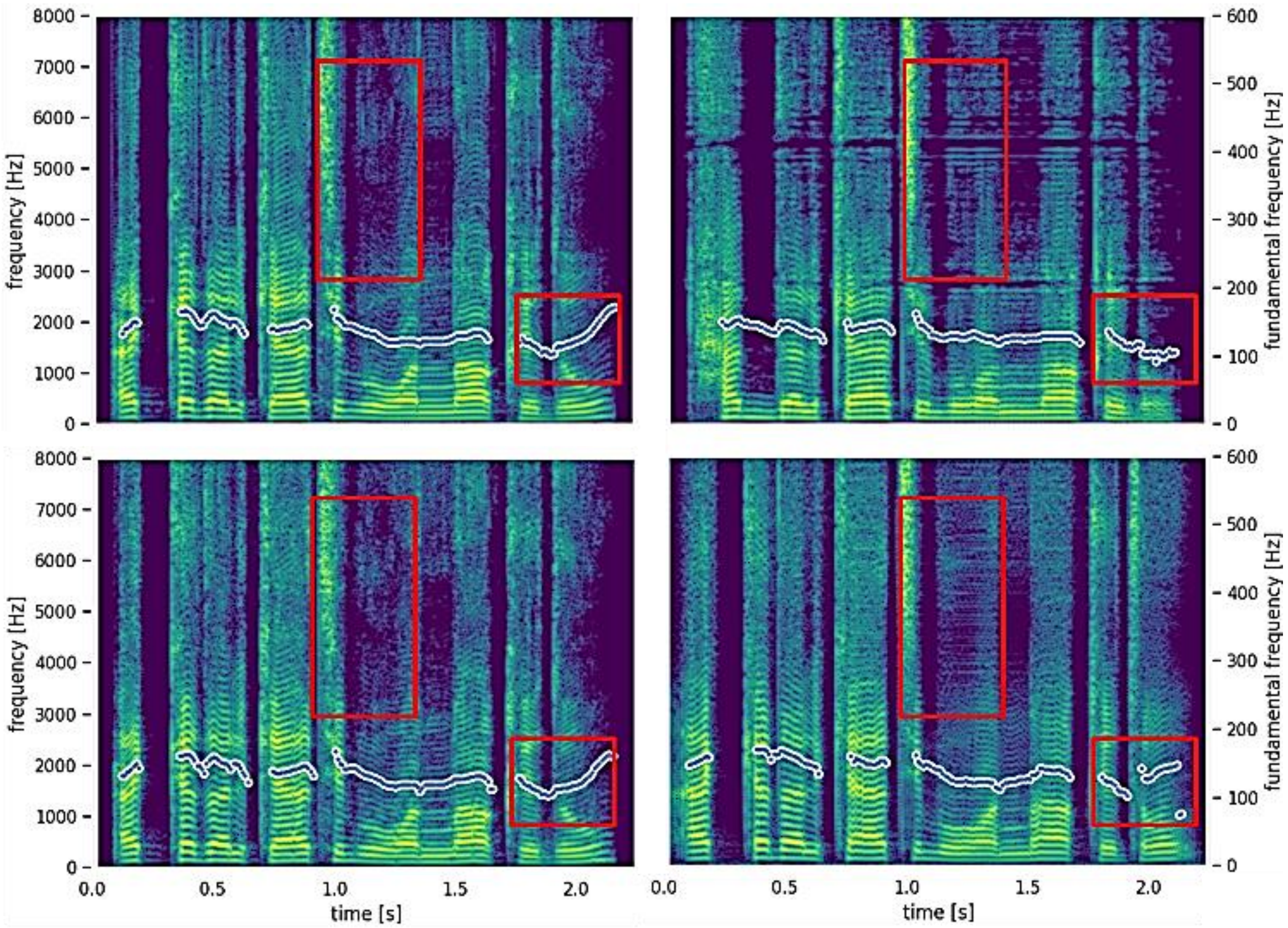
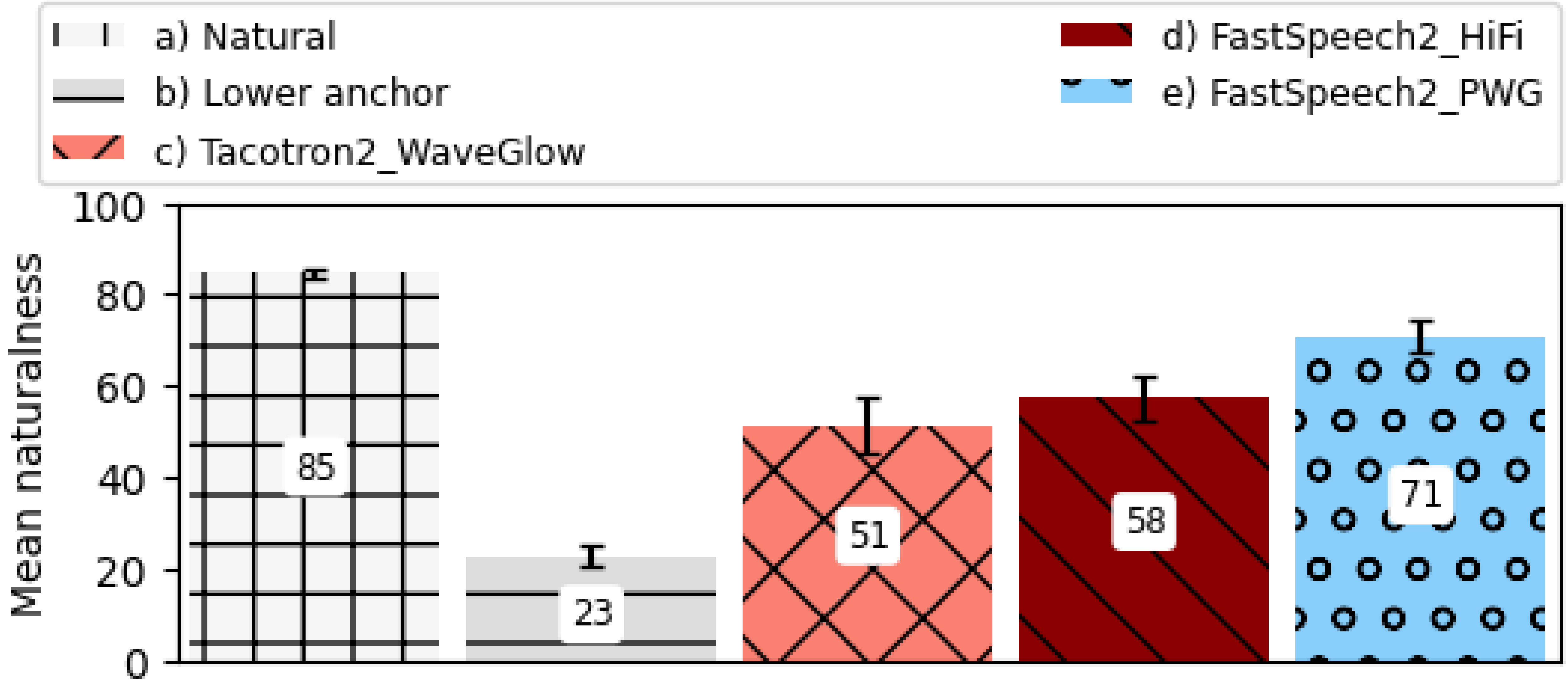



Fig.5: Mel-spectrograms and pitch contours extracted from synthesized speech samples. Top-left: Ground-truth; Top-right: Tacotron2; Bottom-right: FastSp2-HiFi; and Bottom-left: Developed FastSp2-PWG. Only the voiced portion of the pitch contour is plotted (shown as blue curves). The orthographic transcript is: “>ak~ada AlokaAtibu waAln~aAqidu”.

# Subjective listening test

- test took 15 minutes to fill
- 29 participants (11 males, 18 females)

❑ Samples



# Conclusion and Future Works

## Key Achievements

1. Efficient Arabic TTS system with components like speaker adaptation, Parallel waveform, and text-to-mel spectrogram generation
2. Better alignment, audio quality and speed

## Future Directions

1. speaker embeddings to personalize voices and accents
2. AutoVocoder models to enhance speech quality and naturalness
3. Support diverse accents

# Demo Samples

**Natural** | **Tacotron_WaveGlow** | **FastSp2_HiFi** | **FastSp2_PWG**

**Phoneme Sequence:**

"lakin~a diraAsatahumo - >a^abotato >an~a Alomu$okilapa bimiSora - layosato faqaTo fiy kam~iy~api AlT~aEaAmi"

# Thank you for your attention

**12:50 – 13:10**

- 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***

malradhi@tmit.bme.hu