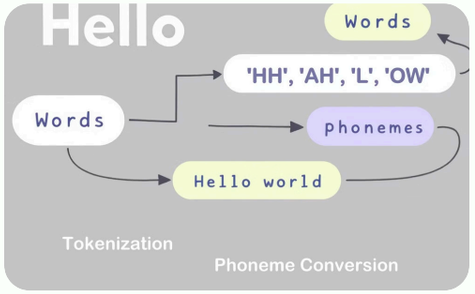
Text-to-Speech
Lab-4

This lab provides examples and exercises to equip you with the knowledge and skills needed to use text-to-speech systems.

Natalia Tomashenko

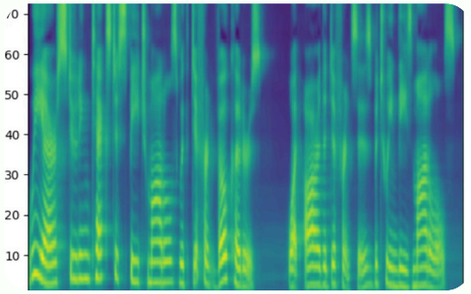
Fundamentals of Text-to-Speech



Text Processing

Tokenization and Phoneme Conversion

First, **tokenization** breaks the text into individual units like words and punctuation. Then, each word is converted into a sequence of **phonemes**, the basic sounds of a language (e.g., "cat" becomes /kæt/). This conversion relies on linguistic knowledge and pronunciation rules, often using dictionaries and phoneme-to-grapheme models.



Acoustic Modeling

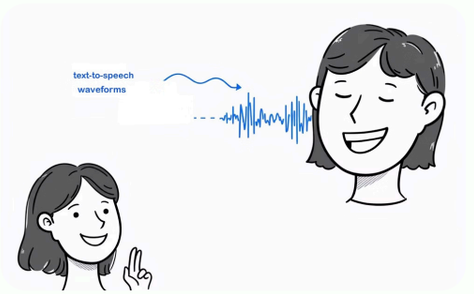
Generating a Spectrogram from Phoneme Sequences

Acoustic modeling takes the phoneme sequence and transforms it into a spectrogram.

Speech Synthesis

Waveform Generation from a Spectrogram

This step involves generating a continuous audio waveform from the spectrogram.

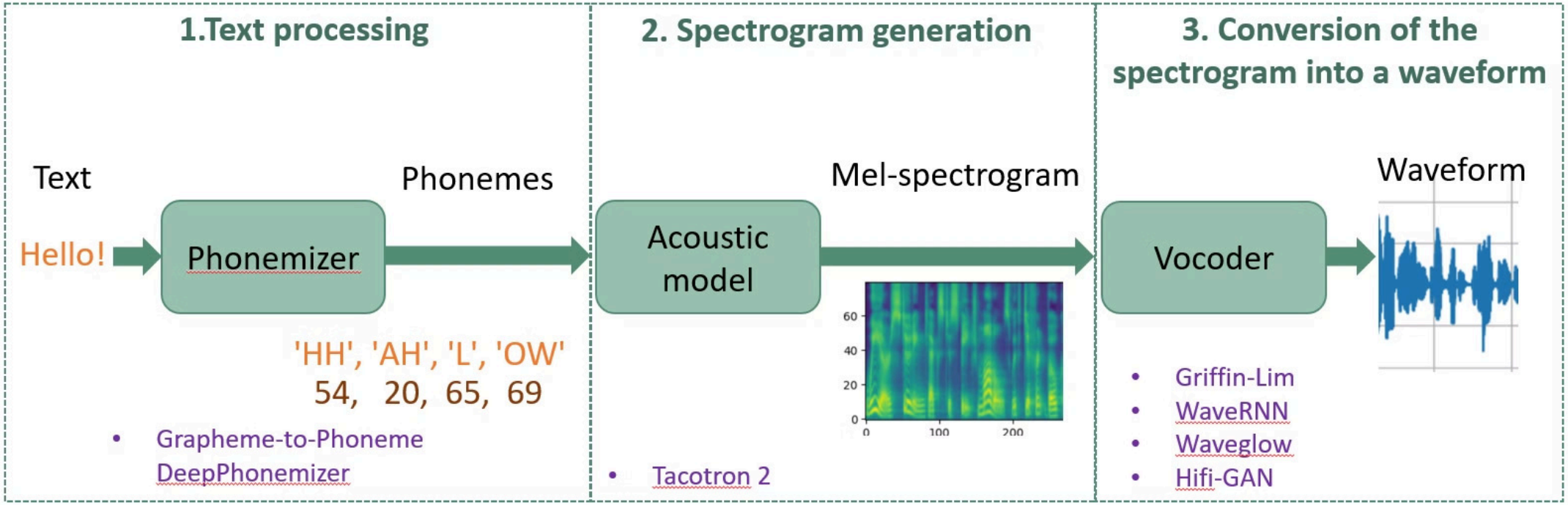


Evaluation

Naturalness and Intelligibility

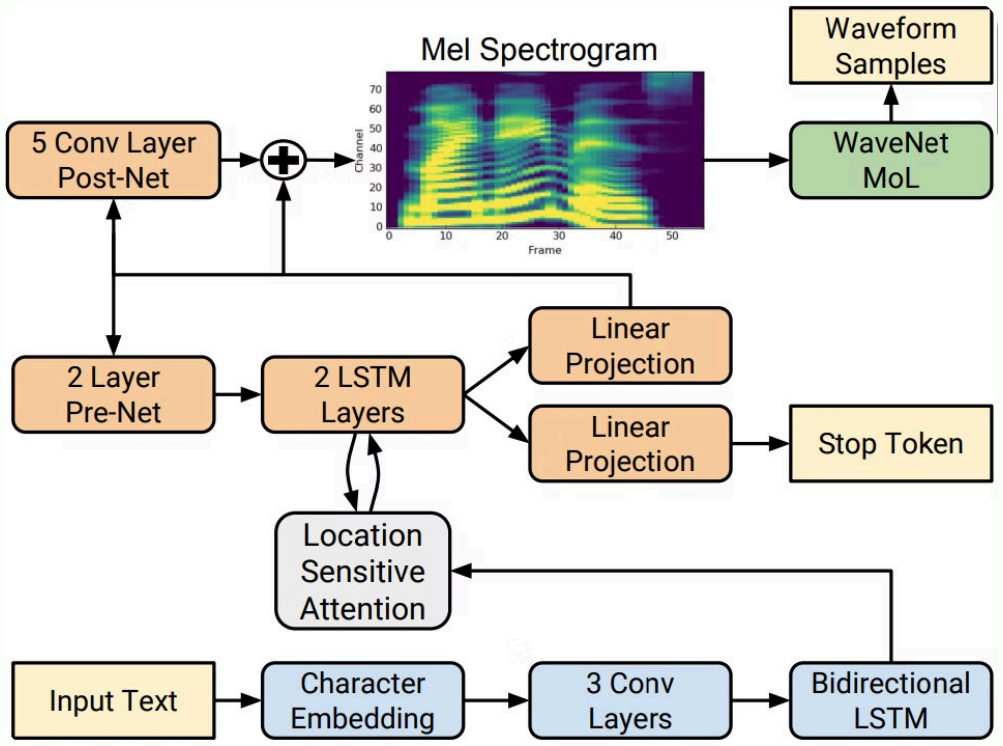
A successful TTS system is both natural-sounding (**naturalness**) and easily understood (**intelligibility**). Achieving this balance requires careful attention to every stage of the process.

Key components of neural TTS



Tacotron 2 model

The **Tacotron 2** model produces mel spectrograms from input text using encoder-decoder architecture. **WaveGlow** is a flow-based model that consumes the mel spectrograms to generate speech.



Subjective Evaluation

The **mean opinion score (MOS)** is a **subjective test** for the evaluation of **speech synthesis** systems where a group of **human listeners** is asked to rate the quality: **speech naturalness** and **intelligibility** of synthesized speech produced by a TTS system on a scale from **1** to **5**. The ratings are then averaged across all listeners to obtain the final MOS.

Subjective Evaluation

Noise	Raters evaluated the level of background noise or artifacts in the audio.	• **None**: No detectable noise or artifacts. • **Low**: Minor noise, such as a faint hiss, but not distracting. • **Medium**: Noticeable noise or artifacts that may be somewhat distracting. • **High**: Significant noise or artifacts that interfere with comprehension.
Context awareness	Raters assessed the TTS system's ability to adjust to context, including tone, emphasis, and punctuation.	• **High**: Excellent adaptation to context, with clear tonal shifts and pauses. • **Medium**: Adequate adaptation but misses some nuances. • **Low**: Little to no adaptation; reads text monotonously without context cues.
Prosody accuracy	Raters evaluated the rhythm, stress, and intonation (prosody) of the generated speech.	• **High**: Natural-sounding rhythm and intonation that closely mimics human speech. • **Medium**: Mostly natural but with occasional awkwardness in stress or rhythm. • **Low**: Monotonous or robotic intonation with unnatural pauses and emphasis.

(source: https://labelbox.com/guides/evaluating-leading-text-to-speech-models/#evaluation-process)

Subjective Evaluation

Word errors	Raters evaluated the presence of word errors in the generated TTS audio by counting the following: • **Word Insertion Errors**: Additional words not present in the original text. • **Word Deletion Errors**: Words from the original text that were omitted. • **Word Substitution Errors**: Incorrect replacements for original words.	• **High:** No word errors or only minor errors (0-1 error). • **Medium:** Some errors (2-3 errors). • **Low:** Significant errors (4 or more errors).
Speech naturalness	Raters assessed how human-like the generated speech sounded.	• **High**: Realistic lighting, textures, and proportions • **Medium**: Somewhat realistic but with slight issues in shadows or textures • **Low**: Cartoonish or artificial appearance
Pronunciation accuracy	Raters evaluated how clearly and correctly words were pronounced in the TTS output.	• **High: A**ll words are pronounced clearly and correctly. • **Medium:** 1-2 words are mispronounced or unclear. • **Low:** 3 or more words are mispronounced or difficult to understand.

(source: https://labelbox.com/guides/evaluating-leading-text-to-speech-models/#evaluation-process)

Objective Evaluation

Objective metrics provide automated, quantitative assessments of speech synthesis quality.

Perceptual Evaluation of Speech Quality (PESQ) measures the perceived quality by comparing the synthesized speech to a reference recording

Short-Time Objective Intelligibility (STOI) assesses how well the synthesized speech can be understood.

Word Error Rate (WER) measures the accuracy of the generated speech by calculating the number of *insertions*, *deletions*, and *substitutions* compared to a reference transcription. The lower the WER, the better the model performed in terms of generating accurate and coherent speech. Automatic speech recognition (ASR) model can be used to obtain transcripts of the synthesised speech and compute WER.

Neural network models trained on subjective scores from human listeners to predict **MOS,** i.e. MOSNet, **wv-MOS**, **Non-Intrusive Speech Quality Assessment (NISQA)**,...

In lab 4, we will use wv-MOS and NISQA.

Exercises

Simple TTS and Subjective Evaluation

Synthesize a simple text message using the proposed TTS library and models.

Objective evaluation

Compare quality of the synthesized speech obtained by different vocoders (using wv-MOS and NISQA)

Multi-speaker TTS

Experiment with different voices.

Supplementary reading and resources

1. **SpeechBrain** toolkit: **https://github.com/speechbrain/speechbrain**
2. **Tacotron 2** model in pytorch:**https://pytorch.org/hub/nvidia_deeplearningexamples_tacotron2/**
3. **Griffin-Lim Vocoder**: D. Griffin and J. Lim. "Signal estimation from modified short-time Fourier transform." *IEEE Transactions on acoustics, speech, and signal processing* 32.2 (1984): pp 236-243.
4. **WaveRNN**: N. Kalchbrenner et al, "Efficient Neural Audio Synthesis" arXiv:1802.08435, 2018
5. **WaveGlow**: R. Prenger et al, "WaveGlow: A Flow-based Generative Network for Speech Synthesis" In Proc. ICASSP (2019), pp 3617 - 3621.
6. **Tacotron**: Y. Wang et al, "Tacotron: Towards end-to-end speech synthesis." *arXiv preprint arXiv:1703.10135* (2017).
7. **Tacotron 2**: J. Shen et al, "Natural TTS synthesis by conditioning Wavenet on Mel spectrogram predictions" In Proc. ICASSP (2018), pp. 4779-4783
8. **HiFi-GAN Vocoder**: J. Kong et al, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis" Advances in neural information processing systems (2020).
9. **NISQA (speech quality evaluation)**: G. Mittag et al, "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets" *arXiv preprint arXiv:2104.09494* (2021).
10. **NISQA for TTS (speech naturalness evaluation)**: G. Mittag et al, "Deep Learning Based Assessment of Synthetic Speech Naturalness. Proc. Interspeech (2020), pp 1748-1752.