Your-TTS: Achieving Zero-Shot Multilingual TTS

Aviad Korall (I.D. 31638511), Eviatar Saadon (I.D. 316321603), Nir Shalem (I.D. 207447657), Tomer Sorany (I.D. 206484859), Or Pinchasov (I.D. 209223726)

{koralla, saadone, nirshale, sorany, orpinc}@post.bgu.ac.il
School of Electrical & Computer Engineering, Ben-Gurion University of the Negev

February 6, 2025

Abstract

Your-TTS is an advanced text-to-speech (TTS) model designed for zero-shot generation of speech for previously unseen speakers. Building upon the VITS architecture, our model incorporates chaining the YourTTS model with transcription of Hebrew text, translating the text from Hebrew to English, and then generating English speech while using the original Hebrew speech as reference audio. This project specifically explores the model's zero-shot capabilities with Hebrew, a language excluded from the original training dataset, demonstrating the model's ability to generalize across linguistic boundaries.

1 Relevant Papers

- J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," arXiv preprint arXiv:2106.06103, 2021.
- E. Casanova et al., "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," arXiv:2112.02418, 2023.
- X. Tan et al., "NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality," arXiv:2205.04421v2, 2022.
- Jadoul, Y., Thompson, B., & de Boer, B. "Introducing Parselmouth: A Python interface to Praat. Journal of Phonetics," 71, 1-15. https://doi.org/10.1016/j.wocn.2018.07.001, 2018.

2 Problem Description

The project addresses the challenge of generating speech that sounds like the original speaker, while translating the speech from one language to another (in our case from Hebrew to English). The end goal here is enabling people to communicate naturally despite language barriers - the translation models together with TTS could eventually be used for real-time translation and generation of translated speech to a language of choice, so that any two people would be able to communicate with each other ”natively”. This problem has many challenges, the main ones are inference time - which has to be extremely short for real-time communication, and the ability to translate from and to as much languages as possible. In this work we attemt to tackle the latter, by demonstrating zero-shot capabilities of the YourTTS model on new languages and therefore support for languages which it was not trained on.

3 Chosen Method for Solving

3.1 Preprocessing

The method involves several preprocessing steps:

- Sentence splitting to improve speech naturalness
- Translation of Hebrew sentences to English
- Transcription of Hebrew recordings into Hebrew text

Our implementation consists of three main components:

1. Speech-to-text conversion using Google Web Speech API
2. Text-to-text translation using googletrans
3. Text & speech-to-speech synthesis using Your-TTS

3.2 Architecture

The Your-TTS architecture comprises several key components:

- **Text Encoder:** Transformer-based with 10 blocks and 196 hidden channels, incorporating 4-dimensional trainable language embeddings
- **Flow-based Decoder:** Utilizes 4 affine coupling layers with WaveNet residual blocks
- **Vocoder:** Modified HiFi-GAN v1 with enhanced discriminator
- **Variational Autoencoder (VAE):** Based on VITS framework
- **Posterior Encoder:** 16 non-causal WaveNet residual blocks
- **Stochastic Duration Predictor:** Enables diverse speech rhythm synthesis

3.3 Loss Functions

The training process utilizes multiple loss components, where the main one was based on the cosine similarity loss:

$$L_{SCL} = \frac{-\alpha}{n} \cdot \sum_{i}^{n} cos\_sim(\phi(g_i), \phi(h_i)), \tag{1}$$

where g and h represent, respectively, the ground truth and the generated speaker audio.
The rest of the loss functions used during training are:

- **Reconstruction Loss:** Mean Squared Error (MSE) between the original and reconstructed mel-spectrograms
- **Adversarial Loss:** Generated by multiple discriminators:
  - Multi-Period Discriminator (MPD)
  - Multi-Scale Discriminator (MSD)
- **KL Divergence Loss:** Applied to the VAE posterior
- **Duration Loss:** For alignment between text and speech
- **Feature Matching Loss:** To stabilize adversarial training

3.4 Inference Process

The inference pipeline consists of the following steps:

1. Text input is encoded through the transformer-based text encoder
2. Reference audio is processed to extract speaker embeddings
3. The flow-based decoder generates mel-spectrograms conditioned on:
   - Encoded text
   - Speaker embeddings
   - Duration predictions
   - Language embeddings
4. The HiFi-GAN vocoder converts mel-spectrograms to waveforms
5. Post-processing is applied to enhance audio quality

4 Novelty of Method

Our approach differs from existing methods in several key aspects:

- Integration of cross-lingual zero-shot capabilities, particularly for languages not present in the training data
- Application of Your-TTS for Hebrew language processing
- Unique combination of speech-to-text, translation, and speech synthesis in a single pipeline

5 Definition of Success and Training Goals

The success of our model was evaluated against the following criteria:

- **Speaker Similarity:** High correlation scores between generated and reference audio using SpeechBrain verification
- **Acoustic Features:** Strong match in pitch and formant characteristics as measured by Parselmouth
- **Cross-lingual Performance:** Successful preservation of speaker identity across language boundaries
- **Natural Prosody:** Appropriate pause placement and rhythm in generated speech

6 Datasets

6.1 Training Datasets

- **English:**
  - VCTK: 44 hours from 109 speakers
  - LibriTTS: Various speakers from audiobooks
- **Portuguese:** Single-speaker dataset (10 hours)
- **French:** M-AILABS (175 hours, multiple speakers)

6.2 Experimental Dataset

For testing, we used Hebrew speech datasets from virtual audiobooks, with each speaker's folder containing 100 audio files ranging from 5 to 30 seconds in length.

7 Evaluation Methods

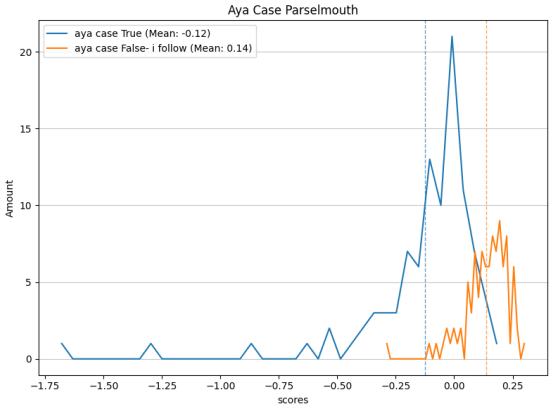
We employed two evaluation approaches:

- **SpeechBrain:** For speaker verification
- **Parselmouth:** For pitch and formant similarity analysis

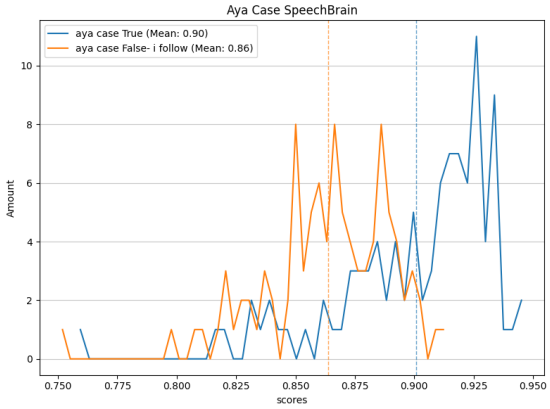
The evaluation process involved:

1. Comparing generated audio with reference recordings
2. Comparing generated audio with unrelated reference recordings
3. Creating distribution histograms for true and false matches

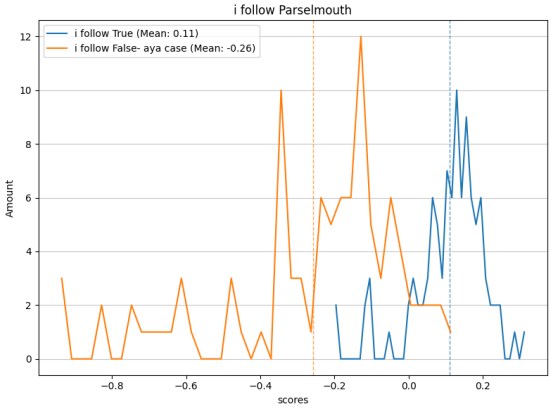
8 Results



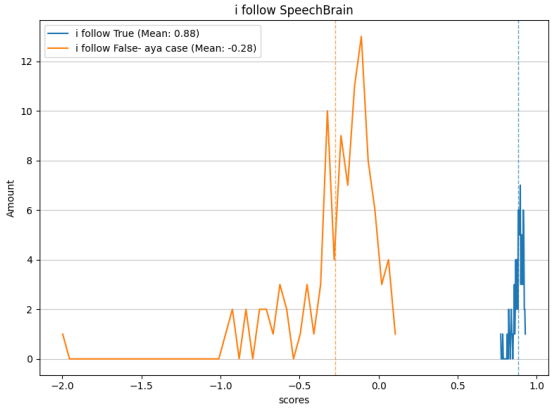
(a) Aya Case Parselmouth Analysis



(b) Aya Case SpeechBrain Analysis



(c) I Follow Parselmouth Analysis



(d) I Follow SpeechBrain Analysis

Figure 1: Evaluation results showing true and false match distributions for both Parselmouth and SpeechBrain analyses across two test cases.

Looking at these histograms, which show evaluation results from both Parselmouth and SpeechBrain for two different cases (”Aya Case” and ”i follow”), several key observations can be made:

1. SpeechBrain Results (Images 2 & 4): - Both cases show high scores for true matches (around 0.88-0.90) - Clear separation between true matches (blue) and false matches (orange) - SpeechBrain appears to be more consistent in differentiating between true and false matches - The distributions have less overlap compared to Parselmouth results

2. Parselmouth Results (Images 1 & 3): - Lower overall scores compared to SpeechBrain (means around -0.12 to 0.14) - More overlap between true and false match distributions - Less distinct separation between true and false matches - Higher variability in scores

3. Comparative Analysis: - SpeechBrain appears to be a more reliable metric for speaker verification, showing clearer distinction between true and false matches - Parselmouth shows more ambiguous results with significant overlap between true and false matches - The results suggest that relying on multiple evaluation methods was a good decision, as they provide different perspectives on the voice cloning quality

4. Performance Implications: - The model appears to maintain speaker identity well (as shown by SpeechBrain scores) - The acoustic features (measured by Parselmouth) show more variation, suggesting that while speaker identity is preserved, some acoustic details might be altered in the process.

Our experiments demonstrated:

- Strong performance in maintaining speaker identity for some cases
- Challenges in proper pause placement due to structural differences between Hebrew and English
- Inconsistent evaluation results across different methods, suggesting the need for multiple evaluation approaches

9 Summary

9.1 Key Achievements

- Demonstrated partially successful zero-shot generalization to unseen languages.
- Developed a system for Hebrew-to-English speech translation with voice preservation.