Quantitative Trade-off Between Inference Latency and Speaker Verification Accuracy in Knowledge-Distilled Multi-Speaker Synthesis

Assignee Research

June 12, 2026

Abstract

Personalizing a speech synthesis system is a highly desired application, where the system can generate speech with the user's voice with rare enrolled recordings. There are two main approaches to build such a system in recent works: speaker adaptation and speaker encoding. On the one hand, speaker adaptation methods fine-tune a trained multi-speaker text-to-speech (TTS) model with few enrolled samples. However, they require at least thousands of fine-tuning steps for high-quality adaptation, making it hard to apply on devices. On the other hand, speaker encoding methods encode enrollment utter

1 Introduction

This paper examines: Meta-TTS: Meta-Learning for Few-Shot Speaker Adaptive Text-to-Speech. Research question: What is the quantitative trade-off between inference latency and speaker verification accuracy when applying knowledge distillation to multi-speaker speech synthesis models?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

4 papers retrieved. 15 claims extracted; 13 independently verified. Quality review score: 8.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Meta-TTS can synthesize high speaker-similarity speech from few enrollment samples.	✓	0.22
Meta-TTS requires fewer adaptation steps than the speaker adaptation baseline.	✓	0.20
Meta-TTS outperforms the speaker encoding baseline under the same training scheme.	✓	0.23
When the baseline speaker encoder is pre-trained with data from 8371 extra speakers, Meta-TTS outperforms the baseline o	✓	0.18
When the baseline speaker encoder is pre-trained with data from 8371 extra speakers, Meta-TTS achieves comparable result	✓	0.16
The code for Meta-TTS is publicly available at https://github.com/SungFeng-Huang/Meta-TTS/.	✓	0.22
Demo audio for Meta-TTS is provided at https://reurl.cc/k7l1nG.	✓	0.16
Text-to-Speech (TTS) systems can synthesize a natural human voice when trained with a large amount of high-quality singl	✓	0.25
Multi-speaker TTS corpora contain a fixed set of speakers.	×	0.12
Speaker encoding methods build a multi-speaker TTS architecture consisting of a speaker encoder and a TTS model.	✓	0.17
Speaker adaptation methods utilize a speaker embedding look-up table instead of a speaker encoder.	✓	0.16
In speaker adaptation methods, the speaker embedding table is jointly trained with the TTS model.	✓	0.22
Previous experiments indicate that the speaker adaptation approach requires thousands of adaptation steps.	✓	0.16
Model-Agnostic Meta-Learning (MAML) consists of two optimization loops: an outer loop to find a meta-initialization and	✓	0.24
The SMOS test rates speaker similarity on a five-point Likert Scale ranging from 1 (Bad) to 5 (Excellent).	×	0.05

References

- http://arxiv.org/abs/1907.08294v1
- http://arxiv.org/abs/2111.04040v3
- http://arxiv.org/abs/2002.12645v2