Edge-Ready Italian TTS via Knowledge Distillation of a Fine-Tuned Transformer Model

1st Alessandro Cagiano
Dept. of Computer Science
University of Milan
Milan, Italy
alesssandro.cagiano@unimi.it

2nd Matteo Zignani
Dept. of Computer Science
University of Milan
Milan, Italy
matteo.zignani@unimi.it

***Abstract*—This ongoing research explores the development of a lightweight Text-to-Speech (TTS) system for the Italian language, based on the fine-tuning of the transformer-based *Sesame CSM-1B* model. The goal is to enable high-quality, low-latency TTS on edge devices for smart city applications, particularly in the domains of digital tourism and accessible cultural services. After language-specific fine-tuning, we will apply knowledge distillation to compress the model for real-time inference. The resulting system is expected to serve a wide range of use cases, from voice assistants in public kiosks to augmented audio guides. The paper outlines the design choices, training pipeline, and deployment strategy, laying the foundation for future experimentation and large-scale evaluation.**

I. Introduction

Smart cities increasingly rely on natural language interfaces to improve citizen engagement, improve accessibility, and enrich cultural experiences. Among these, Text-to-Speech (TTS) technology plays a key role in enabling human-computer interactions, especially in public-facing services. However, state-of-the-art TTS models are often computationally demanding, limiting their usability on edge devices where real-time, offline operation is essential.

This work proposes a TTS development pipeline tailored to the Italian language, designed to meet the constraints of edge deployment. We fine-tune ***Sesame CSM 1-B*** [1] [2], a transformer-based model, on Italian speech data and subsequently compress it through knowledge distillation. This approach aims to preserve voice quality and expressiveness while achieving fast inference on low-resource hardware. The system is intended for smart city scenarios such as multilingual audio guides, inclusive tourism services, and accessible public terminals.

II. Methodology

A. Model Selection and Language Adaptation

Sesame CSM 1-B was selected for its balance between synthesis quality and computational cost. For its adaptation to the Italian language, we constructed a preprocessed dataset combining open-source recordings, audiobooks, and public speech repositories. Special attention is given to phonetic correctness, accented vowels, and handling of numbers and dates—keys for applications in wayfinding in open cultural sites and tourism.

B. Data Collection and Preprocessing

To construct a high-quality Italian dataset suitable for fine-tuning a large-scale TTS model, we combined multiple open-source resources that offer coverage across various speaking styles, domains, and accents. In particular, we utilized audio data from **YODAS** (YouTube-Oriented Dataset for Audio and Speech) corpus [3] and the **Vox Populi** dataset [4].

Unlike YODAS, Vox Populi offers cleaner, domain-specific recordings and is particularly valuable for its structured metadata. We leveraged a fine-tuned Whisper [5] model to automatically transcribe Italian speech segments and extract critical alignment metadata such as **start and end times**, **utterance duration**, and the **number of detected speakers per segment**. This allowed us to filter for high-confidence, single or multi-speaker utterances with suitable length and clarity for speech synthesis.

All collected data underwent a strict cleaning and deduplication phase, followed by automatic phonetic normalization and alignment checks. This dual-source approach ensured a balance between quantity and quality: YODAS contributed large-scale and naturalistic speech, while Vox Populi offered clean, structured and context-rich utterances, leading to a robust and expressive training corpus for the Italian language.

C. Fine-Tuning Pipeline

The fine-tuning process follows the methodology outlined by Knottenbelt [6], whose key steps include:

- *Pre-tokenization*: By utilizing the Mimi audio tokenizer [7], we convert audio waveforms into discrete tokens, facilitating efficient training and better alignment between text and audio representations;
- *Speaker Embedding*: By prepending a speaker ID to the text input, we incorporate speaker identity, allowing the model to learn speaker-specific characteristics and improve multi-speaker handling;
- *Hyperparameter Optimization*: By employing out-of-the-box tools for hyperparameter optimization, such as Optuna, we seek for the best hyperparameter configuration for better model performance.
- *Training Configuration*: By conducting full fine-tuning of the model weights, as opposed to techniques like LoRA (*https://huggingface.co/docs/peft/main/conceptual_guides/lora*),

we might accommodate significant domain shifts inherent in adapting to a new language.

D. Knowledge Distillation for Edge Efficiency

After the fine-tuning pipeline, we apply knowledge distillation [8] to create a smaller, student model optimized for being run on edge devices, to avoid longer latencies typical of cloud services. This process involves training the student model to replicate the outputs of the larger, fine-tuned teacher model, thereby achieving reduced model size and lower latency while maintaining speech quality. This shift toward **on-device TTS inference** aligns with a growing trend in privacy-aware and latency-sensitive applications. By running the model locally, data sovereignty, real-time responsiveness, and offline availability can be guaranteed—key requirements in smart city scenarios with intermittent or constrained connectivity.

III. Expected impact and use cases

The adoption of neural TTS models in real-world smart city infrastructures is often hindered by their demanding computational requirements, especially when targeting real-time, multilingual voice synthesis. Our approach addresses this limitation by producing a compact Italian TTS model that might remain competitive in terms of quality while being executable on low-power devices.

The resulting system is envisioned as a foundational speech technology enabler for next-generation urban services, particularly in contexts where audio interaction plays a critical role in accessibility, personalization, and user engagement. For example, interactive voice-based guides embedded in public kiosks, smart signage, or mobile devices can offer localized and context-aware narration for historical landmarks, exhibitions, and public events. These systems may serve both residents and visitors, promoting cultural heritage, increasing the visitors' engagement, and enhancing inclusivity for visually impaired individuals or people with reading difficulties.

Furthermore, on-device voice synthesis ensures user privacy and robustness in environments where network availability is limited or intermittent. This makes it a suitable choice for public transport information systems, emergency communication, or self-service terminals in civic buildings. The lightweight model architecture also facilitates the integration in resource-constrained embedded systems, enabling developers and municipalities to deploy speech interfaces without relying on cloud infrastructure.

Our work lays the groundwork for a future where natural, expressive Italian speech synthesis can be seamlessly embedded into the urban digital ecosystem, contributing to more intuitive, equitable, and human-centered smart city services.

Current status and future work

As of now, the project has completed the data preprocessing and training pipeline setup. We have collected and normalized a high-quality dataset consisting of Italian speech samples from diverse sources, with attention to coverage of phonetic variation, sentence prosody, and real-world vocabulary. The Sesame CSM-1B model has been successfully configured to support full fine-tuning on this dataset using discrete audio tokenization and multi-speaker conditioning, according to recent advances proposed in Knottenbelt's work.

The next milestone will involve the systematic execution of fine-tuning runs, followed by internal evaluation based on intelligibility and expressiveness metrics. Once the model achieves satisfactory performance, we will begin knowledge distillation to derive a compact version suitable for edge deployment. This process will be iteratively refined to balance performance and latency, guided by profiling on representative edge hardware such as Raspberry Pi or ARM-based mobile chipsets.

Future work includes both technical extensions and deployment-oriented efforts. On the technical side, we plan to: 1) evaluate speech naturalness and prosody quality through Mean Opinion Score (MOS) testing with native speakers; 2) benchmark inference speed and memory usage across various edge configurations; 3) explore speaker adaptation strategies to simulate multilingual or regional variations; and 4) investigate the adoption of quantization and pruning to reduce model footprint.

On the deployment front, we aim to prototype applications that integrate the distilled TTS engine into interactive systems such as mobile apps, wearable devices, or urban IoT platforms. These case studies will help validate usability in real environments and provide feedback for continuous model improvement.

Ultimately, our objective is to contribute an open, efficient, and culturally-aware TTS resource to the Italian NLP and smart city research communities, empowering developers and institutions to offer richer voice-based interactions in public digital services.

References

[1] S. A. Labs, "CSM-1B: Conversational speech model," 2025, https://csm1b.com/.

[2] ——, "CSM - a conversational speech generation model - github repository," 2025, https://github.com/SesameAILabs/csm.

[3] X. Li, S. Takamichi, T. Saeki, W. Chen, S. Shiota, and S. Watanabe, "Yodas: Youtube-oriented dataset for audio and speech," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023*, ser. 2023 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023. IEEE, 2023.

[4] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. ACL, 2021.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[6] W. Knottenbelt, "How to finetune sesame ai's speech model on new languages and voices," 2025, https://blog.speechmatics.com/sesame-finetune/ - Speechmatics Blog.

[7] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," Tech. Rep., 2024. [Online]. Available: https://arxiv.org/abs/2410.00037

[8] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *ArXiv*, vol. abs/1503.02531, 2015.