Immersive Text to Speech

Expressive TTS: Speaking Styles, Techniques & Integration

Sarthak Singh
Ioanna Karagianni

University of Stuttgart
Germany

Motivation

Despite recent advancements in TTS, capturing the full spectrum of human expressive characteristics is challenging. These subtle nuances are difficult to accurately generate using conventional model architectures. Our study focuses on pointing out the latest advancements in the field of **Expressive TTS**:

- **Proposed TTS datasets**
- **Representation and Conditioning of Speaking Styles**
- **Integration of Text Foundation Models**

Datasets & Style Representation

Expressive Dataset

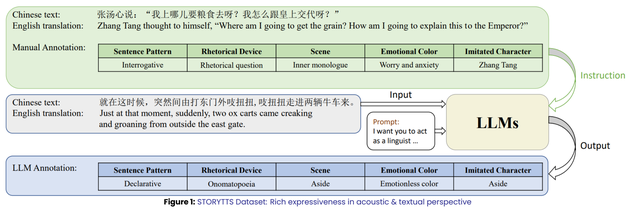

Figure 1: STORYTTS Dataset: Rich expressiveness in acoustic & textual perspective

Timbre Cloning & Style Control

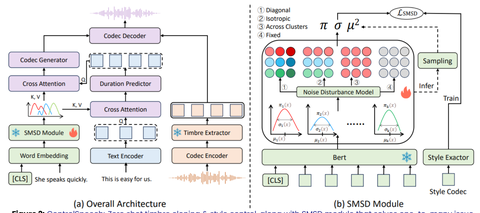

Figure 2: ControlSpeech: Zero shot timbre cloning & style control, along with SMSD module that solves one-to-many issue

Modeling of Voice & Utterance Prosody

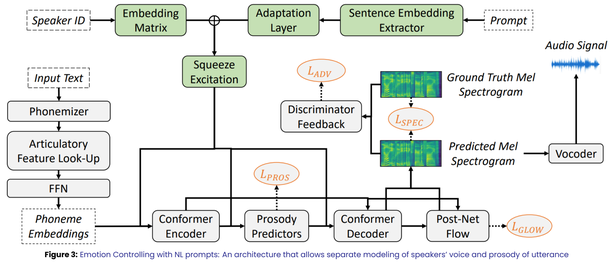

Figure 3: Emotion Controlling with NL prompts: An architecture that allows separate modeling of speakers' voice and prosody of utterance

Extracting Style Representations & Captioning

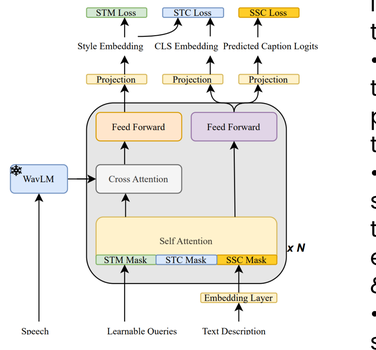

Figure 4: UniStyle: UniConnector is a style extraction and conversion module - Bridge between speech and text modalities

- **STORYTTS**: Rich expressiveness in acoustic & textual perspective.
- **ControlTTS**: Solves the many-to-many issue & proves independent control over content, timbre, style is possible.
- **UniConnector**: Establishes speaking style alignment between speech & text modalities extracting style representations & style captions.
- Combination of emotional speech & text datasets allows precise prosodic & timbral controllability.

Conditioning on Style

Timbre Cross Attention & SALN

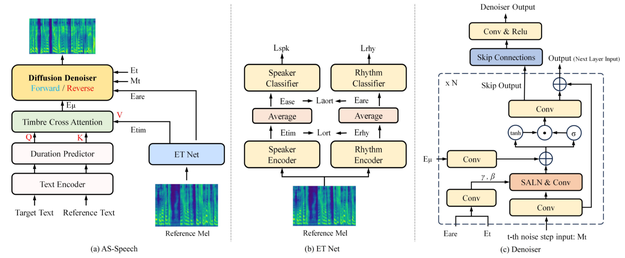

Figure 5: AS Speech Architecture

Conditional Flow Matching

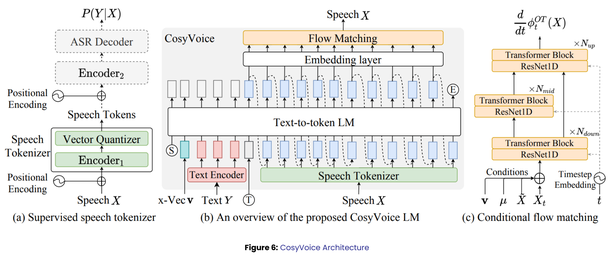

Figure 6: CosyVoice Architecture

Using Text Foundation Models

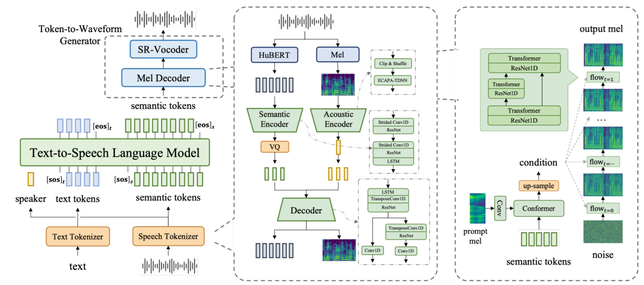

Figure 7: FireRedTTS Architecture

Applications

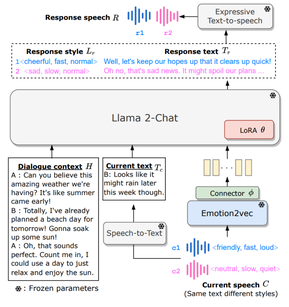

Figure 8: Framework of Spoken LLM

- **AS-Speech**: Integrates timbre and rhythm using SALN for fine-grained style control.
- **CosyVoice**: Employs supervised semantic tokens and conditional flow matching for multilingual zero-shot synthesis.
- **FireRedTTS**: Leverages text foundation models to capture complex semantics and varied speaking styles.
- **Spoken-LLM**: Combines expressive TTS with a Spoken-LLM framework to model speaking styles in conversational applications.

Take Home: Recent advancements highlight the importance of rich datasets and innovative style representation techniques, enabling precise control over prosody, timbre, and expressive alignment between text and speech for more natural and adaptive TTS systems.

References via QR Code

Scan the QR code below to access the main papers and additional resources: