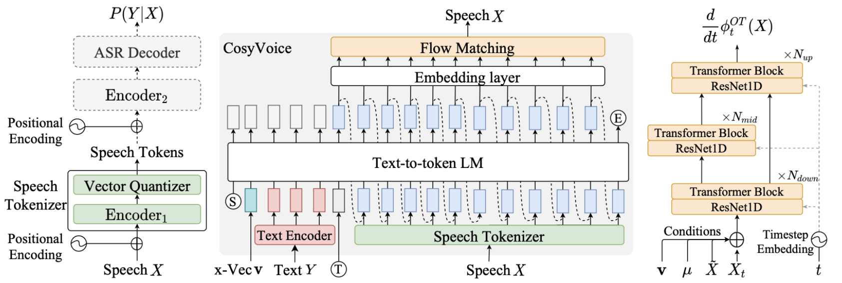
□

home
papers
study
life

About me

CosyVoice



(a) Supervised speech tokenizer (b) An overview of the proposed CosyVoice LM (c) Conditional flow matching

Figure 1: An overview of the proposed CosyVoice model. (a) demonstrates the $\mathcal{S}^3$ tokenizer, where dashed modules are only used at the training stage. (b) is a schematic diagram of CosyVoice, consisting of a text-to-token LLM and a token-to-speech flow matching model. Ⓢ, Ⓔ and Ⓣ denote the "start of sequence", "end of sequence" and "turn of speech" tokens. Dashed lines indicate the autoregressive decoding at the inference stage. (c) provides an enlarged view of our flow matching model conditioning on a speaker embedding $\mathbf{v}$, semantic tokens μ, masked speech features $\tilde{X}$ and intermediate state X_t at timestep t on the probabilistic density path.

- □□□□□CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens
- □□□□□https://arxiv.org/abs/2407.05407
- □□

□□□□□□□□□□□□□□□□□tokenizer□□TTS□□token□□□□□□□□□□□□□□token□□□□□□□□□□□□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□token□□□□□□□□□□□□□□□token□□□token□□□□□□□□□□□□□□□speaker□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□token□□□□□□□□□□□□□□□□□□□□□□token□

Newer Older

2024-07-23
VITS

2024-07-20
OpenVoice

Archive RSS feed GitHub Email QR Code

Made with Montaigne and bigmission