FlEx-TTS: Multilingual Zero-Shot Expressive TTS via Conditional Flow Matching with Hybrid Duration Modeling

Anonymous ACL submission

Abstract

While modern text-to-speech (TTS) systems achieve high intelligibility, they often lack fine-grained prosodic dynamics and temporal coherence essential for expressive, naturalistic speech. Attention-based alignment yields unstable duration modeling, resulting in prosodic inconsistencies and degraded robustness across diverse input lengths. We present FlEx-TTS, a reference-based multilingual, zero-shot TTS system that addresses these challenges using: (1) a hybrid duration predictor combining Monotonic Alignment Search (MAS) with word-level boundary supervision from the Montreal Forced Aligner (MFA), conditioned on global prosodic characteristics including speaking rate, valence, arousal, dominance, normalized voiced phoneme index (NVPI), and pause rate, (2) a trainable temporal prosodic encoder that encodes frame-level prosodic features from a reference utterance to a compact prosody representation for duration conditioning and (3) a conditional flow matching acoustic model built on a Diffusion Transformer (DiT) conditioned on fine-grained prosody. Experiments across multiple languages and speakers demonstrate that FlEx-TTS achieves enhanced naturalness and speaker similarity in zero-shot settings, with substantially improved duration consistency and prosodic expressiveness over strong baselines.

1 Introduction

Recent advances in neural text-to-speech (TTS) synthesis have significantly improved the naturalness and intelligibility of synthesized speech. End-to-end sequence-to-sequence models such as Tacotron (Wang et al., 2017; Shen et al., 2018) demonstrated that direct text-to-melspectrogram modeling with attention-based alignment can generate highly intelligible speech. However, attention-based models suffer from alignment instability, particularly for very long and short utterances, frequently producing repeated or skipped phonemes. Furthermore, they suffer from inconsistent phoneme durations for diverse input text lengths when generating in parallel. To address this, architectures such as FastSpeech (Ren et al., 2019) and Prosody-TTS (Pamisetty et al., 2021) introduced explicit duration modeling for stable parallel generation. Glow-TTS (Kim et al., 2020) introduced Monotonic Alignment Search (MAS) to improve alignment robustness, and employed normalizing flows to enhance the acoustic modeling. VITS (Kim et al., 2021) unified variational inference and adversarial learning in a single end-to-end framework. Diffusion-based models such as Grad-TTS (Popov et al., 2021) further enhanced acoustic fidelity. As a result, modern TTS systems achieve high intelligibility and naturalness under controlled conditions.

With the widespread adoption of conversational agents, audiobooks, and virtual assistants, the focus has shifted from intelligibility to expressiveness. Human-machine interaction increasingly demands speech that conveys emotion, emphasis, rhythm, and speaker intent. However, most systems are optimized primarily for average speech reconstruction and tend to produce prosodically flattened outputs. Early efforts toward expressive TTS introduced style conditioning mechanisms such as Global Style Tokens (GST) (Wang et al., 2018b), which enabled unsupervised modeling of speaking style from reference audio. Hierarchical prosody modeling (Hsu et al., 2018) and variational latent representations (Zhang et al., 2019) further improved style controllability. However, these systems rely on utterance-level global embeddings that capture only coarse style information, fundamentally failing to model fine-grained temporal dynamics such as pitch contour, rhythmic variation, energy envelope, shimmer, jitter, etc. This limitation becomes particularly pronounced in expressive speech, where prosodic dynamics are prominent and temporally structured. Later works explored

reference-based style transfer and speaker adaptation in zero-shot settings (Arik et al., 2018; Jia et al., 2018; Casanova et al., 2022), which achieved promising speaker similarity, but expressive control often remained coarse and globally conditioned. Prosody is inherently time-varying, and global embeddings often smooth out these dynamics.

Parallel to expressiveness, multilingual and code-switched TTS have gained significant attention due to the linguistic diversity of real-world deployment environments, particularly in multilingual societies such as Europe and South Asia. Multilingual TTS models leverage shared phoneme inventories and conditions on the language embeddings to enable cross-lingual parameter sharing (Pamisetty et al., 2023). YourTTS (Casanova et al., 2022) extended VITS for multilingual zero-shot synthesis, conditioning on non-trainable language and speaker embeddings. Despite this progress, code-switched scenarios introduce additional complexity as the languages differ substantially in rhythm, phonotactics, duration distributions, and prosodic realization. Alignment models must handle abrupt phonological transitions, language-dependent duration dynamics, and prosodic patterns that do not transfer across language boundaries. However, attention-based alignment and global style conditioning fail to capture these fine-grained prosodic variations, resulting in perceived unnaturalness at language boundaries.

Zero-shot speaker adaptation introduces a third axis of difficulty. Reference-based style transfer and speaker adaptation in zero-shot settings (Arik et al., 2018; Jia et al., 2018; Casanova et al., 2022) have achieved promising speaker similarity, but expressive control remains coarse and globally conditioned. However, prosody is inherently time-varying as the speakers have moment-by-moment evolution of rate, rhythm, and emphasis that a single embedding vector cannot encode. The latent entanglement of speaker characteristics, linguistic content, and language-specific temporal dynamics further hinders controllability and generalization to unseen speakers. Without an explicit mechanism that relates segmental (phoneme-level) and suprasegmental prosodic factors, zero-shot expressive transfer remains fundamentally limited.

Duration modeling is the important module contributing the TTS naturalness, yet each dominant alignment paradigm carries fundamental limitations. Attention-based alignment (Wang et al., 2017) learns soft attention weights between the linguistic and acoustic characteristics jointly with the acoustic model. They require large amount of data and often suffer from word repetitions in short utterances and word missing in long utterances (Ren et al., 2019; Kim et al., 2020; Pamisetty et al., 2021). These failures are stochastic and difficult to detect at inference time. MAS addresses attention collapse by constraining the alignment to be monotonic and finding the most likely alignment path. However, MAS is self-supervised and derives phoneme durations entirely from the likelihood of the acoustic model. So, early in training, when the acoustic model is poorly initialized, MAS produces noisy alignment estimates that propagate into the duration predictor (Tan et al., 2024; Donahue et al., 2021).

Forced alignment tools such as the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) provide reliable phoneme-level boundaries by conditioning on a fixed, pretrained acoustic model and language-specific pronunciation lexicon. However, MFA cannot adapt to the acoustic characteristics of the target speaker or domain, and its accuracy degrades on expressive, emotional, or heavily accented speech where phoneme realizations deviate substantially from those seen during acoustic model training (Shih et al., 2021). Moreover, MFA is external to the TTS model, and it must learn to reproduce MFA-derived durations with an explicit duration model. So, alignment errors from MFA propagate to the TTS model and accumulates with the duration model error. No existing system has jointly leveraged the complementary strengths of MAS and MFA.

Despite recent advances, there are key challenges: (1) alignment instability in multilingual and code-switched scenarios, (2) fine-grained temporal prosody modeling beyond global style embeddings, which is essential for expressive, naturalistic speech, (3) modeling the conditional distribution of phoneme durations, and (4) robust expressive prosody transfer across languages and to unseen speakers. To address these challenges, we propose **FlEx-TTS**, a Reference-Based Multilingual Zero-Shot Expressive TTS framework. FlEx-TTS introduces a **hybrid duration predictor** that combines MAS with word-level boundary supervision from the MFA. MAS provides fine-grained phoneme-level duration targets derived directly from the statistical modeling, while MFA word boundaries serve as structural anchors that prevent the alignment drift and instability that MAS ex-

hibits. The duration predictor is conditioned not only on speaker characteristics but on a rich set of prosodic characteristics extracted from the reference utterance (speaking rate, phone class, valence, arousal, dominance, normalized voiced phoneme index, pause rate, etc), enabling explicit and interpretable control over temporal prosodic style at inference.

We introduce a **trainable temporal prosodic encoder** that processes frame-level temporal prosodic features from the reference utterance and applies cross-attention with the linguistic representations, producing fine-grained prosodic representations. This representations condition the duration predictor, enabling fine-grained temporal prosody transfer. The acoustic decoder is a **Conditional Flow Matching model** built on a Diffusion Transformer (DiT) backbone, conditioned on speaker-conditioned linguistic representations enriched with fine-grained prosodic information. A shared phoneset with language embeddings enables cross-lingual parameter sharing while preserving language-specific phonetic nuances required for multilingual and code-switched synthesis.

The key contributions of this work are: (i) a hybrid MAS+MFA duration predictor conditioned on global prosodic characteristics and a reference-derived temporal prosody, providing both phoneme-level granularity and word-level structural reliability for stable, expressive duration modeling, (ii) a trainable temporal prosodic encoder that captures fine-grained frame-level prosodic dynamics from a reference utterance for duration conditioning, (iii) a conditional flow matching acoustic model on a DiT backbone with temporal prosodic conditioning, and (iv) a unified multilingual framework with shared phoneset and language conditioning supporting multilingual and code-switched synthesis.

The remainder of the paper is organized as follows. Section 2 describes the proposed architecture, including multilingual encoding, duration modeling, and conditional flow matching. Section 3 presents the experimental setup along with the data curation pipeline used to create expressive speech. Section 4 reports objective and subjective results and compare with the state-of-the art TTS models. Finally, Section 5 concludes the paper and outlines future directions.

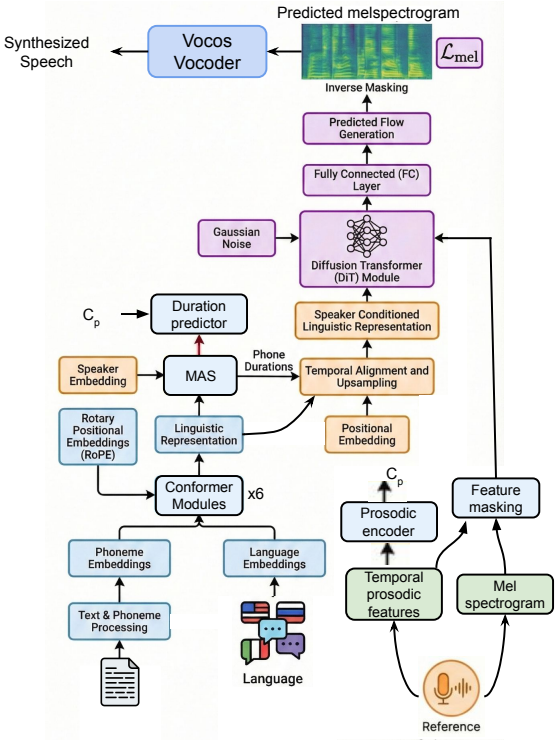

Figure 1: Architecture of the FlEx-TTS system.

2 FlEx-TTS Architecture

The architecture of the proposed reference-based multilingual zero-shot expressive TTS system is illustrated in the figure 1. The proposed FlEx-TTS system addresses key challenges such as alignment instability, limited expressiveness, cross-lingual, and code-switched generalization. Unlike conventional attention-based encoder–decoder systems (Wang et al., 2017), the proposed model eliminates soft alignment learning and instead employs hybrid duration prediction to ensure stable phoneme-to-frame alignment. Furthermore, instead of relying solely on global style embeddings (Wang et al., 2018b), we incorporate frame-level temporal prosodic features directly into acoustic modeling, enabling fine-grained expressive control.

The raw input text is converted to a sequence of language-agnostic international phonetic alphabet (IPA) through a multilingual front-end, and combined with language embeddings to produce a shared multilingual representation that captures both phonetic content and language-specific phonotactic context. These representations are passed through a stack of Conformer blocks that model both local co-articulatory phonetic patterns and global sentence-level prosodic dependencies, pro-

ducing a sequence of linguistic representations that serve as the backbone conditioning signal for all downstream modules.

In parallel, a trainable temporal prosodic encoder extracts frame-level prosodic dynamics that are subsequently combined with the linguistic representations via cross-attention, enabling each phoneme to retrieve the prosodic context most relevant to its temporal position. The linguistic representations, speaker embeddings, and global utterance-level prosodic attributes are passed to a hybrid duration predictor that generates stable phoneme durations supervised jointly by MAS and word-level boundary targets from the MFA. The predicted durations drive a temporal alignment and upsampling module that converts phoneme-rate representations to frame-rate speaker-conditioned linguistic representations. These representations along with the temporal prosodic features condition the Diffusion Transformer acoustic decoder trained with conditional flow matching to generate the melspectrogram. The predicted melspectrogram is finally converted to speech by the Vocos neural vocoder.

The FlEx-TTS model consists of three core components: (1) multilingual text encoder, (2) hybrid duration model conditioned on speaker, language, and prosodic characteristics, and (3) Conditional Flow Matching (CFM)-based acoustic modeling conditioned on frame-level temporal prosody. We describe each component in detail in following subsections.

2.1 Multilingual Text Encoder

Given multilingual or code-switched input text, we first convert the text into a sequence of phonemes $\mathbf{p} = \{p_1, p_2, \ldots, p_N\}$ using epseak (Dunn and Vitolins, 2019). However, phonetic realizations often differ across languages even when the symbolic phoneme is the same. To capture such language-dependent nuances, we condition the phoneme representation on language identity. Each phoneme is mapped to a 128-dimensional fixed (non-trainable) embedding vector. To incorporate language information, a 16-dimensional fixed language embedding $\mathbf{e}^{lang}$ is concatenated with each phoneme embedding:

$$\mathbf{e}_i = [\mathbf{e}_i^{ph}; \mathbf{e}^{lang}]. \tag{1}$$

where e_i denote language-specific embedding of i-th phone and e_i^{ph} denotes i-th phoneme embedding. This conditioning enables the model to learn language-dependent pronunciation, rhythm, and co-articulation patterns while retaining a unified phonetic representation space. The shared universal phoneset enables parameter sharing and avoids redundancy across language-specific phoneme inventories.

Phoneme order is encoded using Rotary Positional Embeddings (RoPE) (Su et al., 2024), which provides relative positional awareness and improves extrapolation to longer sequences compared to absolute positional encoding. The sequence e^{ling} is then processed using stack of Conformer blocks (Gulati et al., 2020):

$$\mathbf{e}^{ling} = [\mathbf{e}_i; \mathbf{e}^{rope}] \tag{2}$$

$$\mathbf{H}^{ling} = \text{ConformerEncoder}(\mathbf{e}^{ling}). \tag{3}$$

Conformer module combines the complementary strengths of convolutional and self-attention layers within a single block. The convolutional submodule captures local phonetic context such as co-articulatory effects and place-of-articulation assimilation across adjacent phonemes, while the multi-head self-attention submodule captures global sentence-level prosodic dependencies. Each Conformer block consists of a feed-forward module, a multi-head self-attention module, a convolution module, and a second feed-forward module, with residual connections and layer normalization applied after each submodule. Unlike purely attention-based encoders, this hierarchical structure emphasizes locality to produce contextualized phoneme-level representations $\mathbf{H}^{ling} \in \mathbb{R}^{T_{ph} \times d}$, where T_{ph} denotes total number of phones in the given utterance and d is the model hidden dimension.

2.2 Hybrid Duration Model

We introduce prosody-conditioned hybrid duration modeling in the FlEx-TTS system. The system departs from both pure attention-based alignment and pure MAS-based training, instead proposing a hybrid framework that combines the acoustic-signal grounding of MAS with the structural reliability of MFA word-level boundary supervision. We also condition the duration predictor on a rich set of prosodic features derived from the reference utterance along with the language and the speaker. Each component of the duration model is described in detail below.

2.2.1 Temporal Prosodic Encoder

. The K-dimensional per-frame temporal prosodic features $P \in \mathbb{R}^{F \times K}$ are extracted from the speech

using OpenSmile tool (Eyben et al., 2010), which comprises of normalized fundamental frequency (F_0), frame energy, spectral flux, shimmer, jitter, etc. They are passed to a trainable temporal prosodic encoder consisting of two dense layers to extract fine-grained frame-level prosodic representations $H^P \in \mathbb{R}^{F \times d_p}$, where F is the number of frames in a given utterance and d_p prosodic encoder hidden dimension. Mean pooling of the prosodic representations will result in the loss of temporal dynamics. So, we apply cross-attention mechanism the prosodic encoder hidden states $\mathbf{H}^p$ and the linguistic representation $\mathbf{H}^{ling}$ from the conformer text encoder. The linguistic representations serve as queries, and the prosodic encoder hidden states serve as both keys and values:

$$\mathbf{C_p} = \sigma\left(\frac{(\mathbf{H}\mathbf{W}_Q)(\mathbf{H}^p\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)(\mathbf{H}^p\mathbf{W}_V) \quad (4)$$

where σ denotes softmax activation, $\mathbf{W}_Q \in \mathbb{R}^{d \times d_k}$, $\mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_p \times d_k}$ are learned projection matrices and d_k is the attention head dimension. This produces a phoneme-aligned prosodic context matrix $\mathbf{C_p} \in \mathbb{R}^{T_{ph} \times d_k}$, where each phoneme representation attends over the full sequence of reference prosodic frames and retrieves the prosodic context most consistent with its phonetic identity and position. The prosodic context $\mathbf{C_p}$ is concatenated with the linguistic representations $\mathbf{H^{ling}}$ to condition the duration predictor on the fine-grained prosodic representations. During inference, the reference utterance and the synthesis text differ in phonetic content, length, and linguistic structure. However, prosodic features are strictly content-agnostic, and cross-attention with the different linguistic representation will not create any adverse effect.

2.2.2 Global Prosodic Attribute Conditioning

In addition to the frame-level prosodic context provided by cross-attention, the duration predictor is conditioned on a set of interpretable utterance-level global prosodic attributes $\mathbf{g} \in \mathbb{R}^G$ extracted from the reference utterance. These attributes capture the overall prosodic character of the reference at a coarser temporal scale and provide explicit, controllable conditioning signals for speaking style transfer. Specifically, $\mathbf{g}$ comprises of speaking rate (syllables per second), phone class identity (vowel, consonant, silence, fricative), NVPI (ratio of voiced phoneme frames to total utterance frames), pause rate (number of silence segments per second), and the three dimensions of the valence-arousal-dominance (VAD) emotional space (Russell, 1980) estimated by a pretrained speech emotion recognition model (Goncalves et al., 2024; Härm and Alumäe, 2024). The global attribute vector $\mathbf{g}$ is also passed as condition to the duration predictor.

2.2.3 Duration Predictor Architecture

The duration predictor takes as input the concatenation of the linguistic representation $\mathbf{H^{ling}}$, the global prosodic attributes $\mathbf{g}$, the speaker embedding $\mathbf{e}_s \in \mathbb{R}^{d_s}$ derived from a pretrained Titanet speaker verification model (Koluguri et al., 2021), fine-grained prosodic context $\mathbf{C_p}$ from the temporal prosodic encoder. These are projected to a common dimension and processed through a stack of Conformer layers that model the sequential dependencies between phoneme durations, capturing sentence-level rhythmic patterns, and produces a sequence of hidden states from which a linear layer predicts the log-domain phoneme duration$\hat{d}_{ph}$ for each phoneme. Operating in log-domain is standard practice in explicit duration prediction (Ren et al., 2020) as it ensures positivity and reduces the dynamic range of the regression target, making training more stable.

2.2.4 Hybrid MAS with MFA Supervision

The duration targets for training the duration predictor are derived from a hybrid alignment procedure that combines MAS (Kim et al., 2020) with word-level boundary supervision from the MFA (McAuliffe et al., 2017). MAS operates by finding the most likely monotonic alignment path $A^* = \arg\max_A \log p(\mathbf{M}|A, \mathbf{H})$ between the phoneme sequence and the mel-spectrogram under the current acoustic model via dynamic programming:

$$A^* = \arg\max_{A \in \mathcal{M}} \sum_{t=1}^{T_{mel}} \log p(\mathbf{m}_t | \mathbf{h}_{A(t)}) \quad (5)$$

where $\mathcal{M}$ is the set of all valid monotonic alignments, $\mathbf{m}_t$ is the mel-spectrogram frame at time t, and $\mathbf{h}_{A(t)}$ is the linguistic representation of the phoneme aligned to frame t. The resulting alignment A^* yields phoneme-level duration targets d_{ph}^{MAS} in frames. However, MAS is self-supervised and therefore susceptible to alignment drift early in training and deviates as the training progresses. To counteract this, we incorporate

a soft word-level boundary prior into the MAS search by adding a Gaussian bias to the log-prior matrix $\log p(\mathbf{m}_t|\mathbf{h}_j)$ before the dynamic programming step. For each word w with MFA-derived frame boundaries $[f^w_{start}, f^w_{end}]$ and phoneme span $[p^w_{start}, p^w_{end}]$, we construct a Gaussian prior centered at the temporal midpoint of the word:

$$\mathbf{B}^w_{j,t} = \lambda_{bias} \cdot \exp\left(-\frac{(t-\mu^w)^2}{2(\sigma^w)^2}\right) \quad (6)$$

where $j \in [p^w_{start}, p^w_{end}]$, $\mu^w = (f^w_{start} + f^w_{end})/2$ is the MFA word centre in frames, $\sigma^w = (f^w_{end} - f^w_{start})/4$ is set to one quarter of the word duration so that the prior covers the frame span of the word at two standard deviations, and λ_{bias} is a scalar weighting hyperparameter. The biased log-prior is then:

$$\log \tilde{p}(\mathbf{m}_t|\mathbf{h}_j) = \log p(\mathbf{m}_t|\mathbf{h}_j) + \sum_w \mathbf{B}^w_{j,t} \quad (7)$$

This soft prior nudges MAS toward the correct word-level temporal regions from the earliest training steps, without hard-constraining the alignment to MFA boundaries, as hard masking causes irrecoverable failures when MFA boundaries are imprecise. The ablation studies on different ways of integrating MFA supervision is detailed in the Appendix. As training progresses and the acoustic model strengthens, λ_{bias} is annealed toward zero, allowing the acoustic likelihood to dominate the alignment decision. The resulting MAS alignment, guided by MFA word priors, yields phoneme-level duration targets d^{MAS}_{ph} that are simultaneously grounded in the acoustic signal and structurally coherent at the word level.

2.2.5 Training Objective

The duration predictor is trained with a two-component hybrid loss. The phoneme-level loss $\mathcal{L}_{phone}$ supervises the per-phoneme log-duration predictions against MAS-derived targets using the Huber loss, which is more robust to outlier durations caused by occasional MAS misalignments than the standard mean squared error (Huber, 1964):

$$\mathcal{L}_{phone} = \frac{1}{T_{ph}} \sum_{j=1}^{T_{ph}} \text{Huber}\left(\hat{d}_j,\ d^{MAS}_j\right) \quad (8)$$

The word-level loss $\mathcal{L}_{word}$ provides additional supervision at the word boundary level using MFA-derived word durations as targets. For each word w spanning phonemes $[p^w_{start}, p^w_{end}]$, the predicted word duration is obtained by aggregating the predicted phoneme log-durations in log space via the log-sum-exp operation:

$$\mathcal{L}_{word} = \frac{1}{N_w} \sum_{w=1}^{N_w} \text{Huber}\Bigg(\log \sum_{j=p^w_{start}}^{p^w_{end}} \exp(\hat{d}_j),\ \log \sum_{j=p^w_{start}}^{p^w_{end}} \exp(d^{MFA}_j)\Bigg) \quad (9)$$

where N_w is the number of words in the utterance and d^{MFA}_j are the MFA-derived phoneme durations aggregated to word level. The combined duration loss is $\mathcal{L}_{dur} = \mathcal{L}_{phone} + \lambda_{word} \cdot \mathcal{L}_{word}$, where λ_{word} is a weighting hyperparameter.

2.3 CFM for Acoustic Modeling

Prosody is inherently time-varying and cannot be fully captured by utterance-level global embeddings (Wang et al., 2018b). To enable fine-grained expressive modeling, we incorporate frame-level temporal prosodic features $\mathbf{P}$. The prosodic features are concatenated with the linguistic representation, $\mathbf{C} = [\mathbf{H}^{ling}; \mathbf{P}]$. This explicit temporal prosody conditioning allows the acoustic model to incorporate localized expressive variations, improving emotional richness and prosody transfer accuracy in zero-shot settings.

Diffusion-based acoustic models (Popov et al., 2021) achieve high quality but require iterative sampling, increasing inference time. We adopt CFM (Lipman et al., 2022), which learns a continuous transformation between a simple base distribution and the target mel distribution through direct vector field regression. In addition to flow matching loss (Lipman et al., 2022), we apply a mel reconstruction loss $\mathcal{L}_{mel} = |\mathbf{y} - \hat{\mathbf{y}}|^2$, where y is the original melspectrogram and $\hat{y}$ is the predicted melspectrogram. The total loss of the model is given by:

$$\mathcal{L} = \mathcal{L}_{mel} + \lambda_{dur}\mathcal{L}_{dur}. \quad (10)$$

where λ_{dur} is the weight parameter deciding the importance of duration loss. During inference, given an unseen speaker reference speech, speaker embedding, global attributes, and temporal prosodic features are extracted. The duration predictor generates phoneme durations, ensuring stable alignment across languages. The CFM acoustic model

then generates expressive mel-spectrograms conditioned on frame-level prosodic features. We have used a pretrained Vocos neural vocoder to convert melspectrograms into speech (Siuzdak, 2023).

3 Experimental Setup

We use two publicly available multilingual datasets released by AI4Bharat: RASA (Varadhan et al., 2024) and IndicVoices-R (Sankar et al., 2024). These corpora provide high-quality studio-recorded speech across multiple Indian languages with transcriptions. The recordings are predominantly read speech collected under controlled acoustic conditions. As the spontaneous expressive speech is limited, we curate a large-scale expressive dataset from unstructured real-world audio sources such as podcasts, interviews, and conversational media. The curation process follows the automated end-to-end pipeline. The pipeline systematically transforms heterogeneous real-world audio into TTS-ready expressive speech by integrating:

- Vocal separation,
- Dereverberation and speech enhancement,
- Voice Activity Detection (VAD),
- Speaker diarization and clustering,
- Overlap detection and SNR-based rejection,
- Loudness normalization and resampling,
- Transcript extraction.

The pipeline processes approximately 13,000 hours of raw unstructured audio collected across 12 Indian languages and English. After source separation, diarization, and quality-based filtering, around 8,100 hours of high-fidelity speech is retained, segmented into approximately 4.5 million utterances with an average duration of 7.2 seconds. The curated dataset contains approximately 2,100 hours from 4,527 speakers (2,412 male and 2,115 female), providing substantial speaker diversity. Unlike acted emotional corpora, this dataset contains naturally occurring spontaneous speech with continuous emotional transitions and realistic prosodic variation, which is critical for training expressive TTS models that generalize to real-world conversational settings. 5% of the data from both sources are kept aside for evaluation.

Word-level ground-truth duration labels are extracted using MFA (McAuliffe et al., 2017). Speaker embeddings are extracted using the Titanet model (Koluguri et al., 2021), which produces discriminative speaker representations trained with additive angular margin loss. Frame-level temporal prosodic features are extracted from reference speech using openSMILE (Eyben et al., 2010). All audio is resampled to 16 kHz and loudness-normalized according to ITU-R BS.1770 standard. Mel-spectrograms are computed using a 1024-point FFT, 256-hop size, and 100 mel filterbanks.

The FlEx-TTS model has around 171 million trainable parameters, trained jointly on studio-recorded corpora (RASA, IndicVoices-R) and the curated expressive dataset. Training is performed on NVIDIA L40S GPUs using mixed-precision training. We use the Adam optimizer with a cosine learning rate scheduler and gradient clipping for stability (Kingma and Ba, 2017). The batch size is dynamically adjusted based on sequence length to maximize GPU utilization while avoiding memory overflow.

4 Results

We compare the proposed approach against several multilingual and zero-shot TTS baselines: YourTTS (Casanova et al., 2022), a multilingual zero-shot multi-speaker TTS system built upon the VITS architecture with speaker embedding conditioning. F5-TTS (Chen et al., 2025), a flow-matching-based TTS model, FastSpeech2 (Ren et al., 2020), and GST-tacotron (Wang et al., 2018a) that uses global style embeddings. All baseline models are trained on the same multilingual corpora and curated expressive dataset for fair comparison. Evaluation is performed under multilingual zero-shot reference-based synthesis settings with unseen speaker references, cross-lingual synthesis, and code-switched input text.

We evaluate objective performance using multiple complementary metrics. Intelligibility is measured using Word Error Rate (WER), computed using a pretrained automatic speech recognition (ASR) model[1]. Speaker similarity is quantified using cosine similarity between synthesized and reference speaker embeddings extracted with Titanet (Koluguri et al., 2021). To assess temporal alignment consistency, we compute the Pearson Correlation Coefficient (PCC) between predicted and reference phoneme durations, reflecting speaking rate preservation. Perceptual speech quality is

[1] https://github.com/AI4Bharat/IndicConformerASR

Table 1: Objective and Subjective evaluation results with 95% confidence interval.

Model	**Objective Metrics**					**Subjective Metrics**	
	WER↓	Spk Sim↑	PCC↑	PESQ↑	STOI↑	MOS (Nat.)↑	MOS (Expr.)↑
YourTTS	12.8	0.81	0.79	3.81	0.89	3.92 ± 0.11	3.35 ± 0.09
GST-Tacotron	13.2	0.76	0.76	3.62	0.88	3.98 ± 0.08	3.63 ± 0.12
FastSpeech2	8.9	0.83	0.81	3.89	0.92	4.01 ± 0.06	3.72 ± 0.05
F5-TTS	8.2	0.81	0.79	3.92	0.91	4.04 ± 0.02	3.82 ± 0.04
FlEx-TTS	**6.8**	**0.92**	**0.89**	**4.11**	**0.95**	$\mathbf{4.18} \pm 0.03$	$\mathbf{4.41} \pm 0.05$

evaluated using Perceptual Evaluation of Speech Quality (PESQ) (Rix et al., 2001), while Short-Time Objective Intelligibility (STOI) is used to measure intelligibility at the signal level (Andersen et al., 2017).

Subjective evaluation is conducted using Mean Opinion Score (MOS) on a 5-point scale to assess perceived naturalness. Additionally, expressiveness similarity MOS is collected, where listeners rate how closely the synthesized prosody matches the reference speech in terms of expressive characteristics. Each model is evaluated on 100 multilingual test sentences across 5 languages, including 50 code-switched utterances. For subjective evaluation, 30 listeners participated to rate 25 randomly chosen utterances.

Table 1 presents the results obtained for subjective and objective tests. The proposed model achieves the highest PCC between predicted and reference speaking rates, validating the effectiveness of explicit speaker and language-conditioned duration modeling. Unlike attention-based baselines, deterministic duration prediction eliminates alignment drift and improves temporal consistency in multilingual and code-switched synthesis. The improvement in cosine speaker similarity further indicates that decoupling duration modeling from acoustic generation reduces entanglement between speaker identity and temporal alignment, allowing the flow-based acoustic model to focus on spectral refinement while preserving speaker characteristics.

Lower WER and improved PESQ/STOI demonstrate that hybrid alignment reduces phoneme skipping and pronunciation inconsistencies. The most significant gains are observed in expressiveness similarity MOS, attributable to frame-level temporal prosodic conditioning rather than global style embeddings. While global style-based models tend to average expressive dynamics across an utterance, explicit frame-level prosodic features enable localized expressive transfer and more natural emotional transitions. In cross-lingual settings, shared phoneset representation combined with language-conditioned duration modeling further enhances stability by explicitly modeling language-dependent temporal characteristics. Overall, these results confirm that adding word boundary supervision to MAS improves alignment robustness, and temporal prosody modeling enhances expressive fidelity. The model has a real-time factor (RTF) of 0.585, and can do 48 parallel generations on L40S with alignment stability across diverse input lengths.

5 Conclusion

We presented FlEx-TTS, a reference-based multilingual zero-shot expressive TTS system that explicitly models temporal prosody to efficiently capture expressive prosodic dynamics. At the core of FlEx-TTS is a hybrid duration modeling framework that combines MAS with word-level boundary supervision from the MFA, conditioned on global prosodic attributes. A trainable temporal prosodic encoder provides fine-grained prosody representations. Frame-level prosodic conditioning integrated within a CFM acoustic model enables fine-grained expressive transfer beyond global style representations, allowing acoustic generation to focus on spectral refinement while temporal dynamics are deterministically modeled. Experiments on studio-recorded and curated spontaneous expressive datasets demonstrate that FlEx-TTS achieves state-of-the-art naturalness and speaker similarity in zero-shot settings, with substantially improved duration consistency and expressiveness over strong baselines. Future work will extend the hybrid alignment framework to leverage multilingual self-supervised speech representations as the alignment backbone, broadening FlEx-TTS coverage to low-resource and truly unseen languages.

Limitations

While FlEx-TTS demonstrates strong performance across multiple languages and zero-shot speakers, there are certain limitations. The hybrid duration predictor relies on MFA-derived word boundary supervision during training, which requires a language-specific acoustic model and pronunciation lexicon. There are pretrained acoustic models available for large number of languages. However, the languages lacking pretrained model requires training an acoustic model. The global prosodic conditioning attributes (valence, arousal, and dominance) are extracted using a pretrained speech emotion recognition model. Errors in these upstream attribute estimates propagate directly into the duration predictor.

Speech disfluencies such as filled pauses (uh, um), false starts, and repetitions remains an open challenge that the current training does not address. Furthermore, FlEx-TTS does not model non-speech vocalizations such as laughter and hesitation sounds, which are pervasive in naturalistic speech and contribute substantially to perceived authenticity and conversational naturalness. The current phoneme-based front-end and duration predictor have no mechanism to generate or transfer such events from the reference utterance. Extending FlEx-TTS to handle non-speech vocalizations and disfluent speech patterns is a critical direction toward conversational TTS systems that can reproduce the full acoustic character of naturalistic human communication.

References

Asger Heidemann Andersen, Jan Mark de Haan, Zheng-Hua Tan, and Jesper Jensen. 2017. A non-intrusive short-time objective intelligibility measure. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5085–5089.

Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural voice cloning with a few samples. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Edresson Casanova, Julian Weber, and et al. 2022. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 2709–2720.

Yushen Chen, Zhikang Niu, Ziyang Ma, and et al. 2025. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *Preprint*, arXiv:2410.06885.

Jeff Donahue, Sander Dieleman, Mikolaj Binkowski, Erich Elsen, and Karen Simonyan. 2021. End-to-end adversarial text-to-speech. In *International Conference on Learning Representations*.

R. H. Dunn and V. Vitolins. 2019. espeak ng speech synthesizer. https://github.com/espeak-ng/espeak-ng. Version 1.50.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA.

Lucas Goncalves, Ali N. Salman, Abinay Reddy Naini, Laureano Moro-Velázquez, Thomas Thebaud, Paola Garcia, Najim Dehak, Berrak Sisman, and Carlos Busso. 2024. Odyssey 2024 - Speech Emotion Recognition Challenge: Dataset, Baseline Framework, and Results. In *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pages 247–254.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020*, pages 5036–5040.

Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, and 1 others. 2018. Hierarchical generative modeling for controllable speech synthesis. *arXiv preprint arXiv:1810.07217*.

Peter J. Huber. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101.

Henry Härm and Tanel Alumäe. 2024. TalTech Systems for the Odyssey 2024 Emotion Recognition Challenge. In *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pages 255–259.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. 2021. Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context. *Preprint*, arXiv:2110.04410.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017*, pages 498–502.

Giridhar Pamisetty, K. Sri, and Rama Murty. 2021. Prosody-tts: An end-to-end speech synthesis system with prosody control. *Circuits, Systems, and Signal Processing*, 42:361–384.

Giridhar Pamisetty, S Chaitanya Varun, and K Sri Rama Murty. 2023. Lightweight prosody-tts for multilingual multi-speaker scenario. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2.

Valentin Popov, Ivan Vovk, and et al. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. *arXiv preprint arXiv:2105.06337*.

Yi Ren, Yangjun Ruan, Xu Tan, and et al. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yi Ren and 1 others. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *ICLR*.

A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752 vol.2.

James A Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Ashwin Sankar, Srija Anand, and et al. 2024. Indicvoices-r: Unlocking a massive multilingual multi-speaker speech corpus for scaling indian TTS. In *NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *Preprint*, arXiv:1712.05884.

Kevin J. Shih, Rafael Valle, Rohan Badlani, Adrian Lancucki, Wei Ping, and Bryan Catanzaro. 2021. RAD-TTS: Parallel flow-based TTS with robust alignment learning and diverse synthesis. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*.

Hubert Siuzdak. 2023. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. *arXiv preprint arXiv:2306.00814*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Sheng Zhao, Tao Qin, Frank Soong, and Tie-Yan Liu. 2024. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4234–4245.

Praveen Srinivasa Varadhan, Ashwin Sankar, and et al. 2024. Rasa: Building Expressive Speech Synthesis Systems for Indian Languages in Low-resource Settings. In *Proc. INTERSPEECH 2024*.

Yuxuan Wang, R.J. Skerry-Ryan, and et al. 2017. Tacotron: Towards end-to-end speech synthesis. In *Interspeech 2017*, pages 4006–4010.

Yuxuan Wang, Daisy Stanton, and et al. 2018a. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International conference on machine learning*, pages 5180–5189. PMLR.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. 2018b. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5180–5189. PMLR.

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949.