Marco-Voice Technical Report

Fengping Tian, Chenyang Lyu*, Xuanfan Ni, Haoqin Sun, Qingjuan Li, Zhiqiang Qian, Haijun Li, Longyue Wang, Zhao Xu, Weihua Luo, Kaifu Zhang
Alibaba International Digital Commerce
* Project Lead and Corresponding Author: lyuchenyang.lcy@alibaba-inc.com

This paper presents a multifunctional speech synthesis system that integrates voice cloning and emotion control speech synthesis within a unified framework. The goal of this work is to address longstanding challenges in achieving highly expressive, controllable, and natural speech generation that faithfully preserves speaker identity across diverse linguistic and emotional contexts. Our approach introduces an effective speaker-emotion disentanglement mechanism with in-batch contrastive learning, enabling independent manipulation of speaker identity and emotional style, as well as rotational emotional embedding integration method for smooth emotion control. To support comprehensive training and evaluation, we construct CSEMOTIONS, a high-quality emotional speech dataset containing 10 hours of Mandarin speech from six professional speakers across seven emotional categories. Extensive experiments demonstrate that our system, Marco-Voice, achieves substantial improvements in both objective and subjective metrics. Comprehensive evaluations and analysis were conducted, results show that Marco-Voice delivers competitive performance in terms of speech clarity and emotional richness, representing a substantial advance in the field of expressive neural speech synthesis. Our code and dataset are publicly available at https://github.com/AIDC-AI/Marco-Voice and https://huggingface.co/datasets/AIDC-AI/CSEMOTIONS respectively.

1. Introduction

The field of text-to-speech (TTS) synthesis has witnessed remarkable progress in recent years, driven by advances in deep learning and the availability of large-scale speech datasets [Zeng et al., 2020, Kim et al., 2021, Shen et al., 2023]. Modern TTS systems now approach or even surpass human-level performance in terms of intelligibility and naturalness, making them indispensable in a wide range of applications, including virtual assistants, audiobook narration, accessibility tools, and entertainment [Li et al., 2024b, Du et al., 2024a,b].

Despite these achievements, truly human-like speech synthesis remains an open challenge [Li et al., 2024a]. In natural communication, human speech is characterized by a rich interplay of speaker identity, prosodic style (intonation, rhythm, emphasis), and nuanced emotional expression [Tan et al., 2021, Zhang et al., 2023, Barakat et al., 2024]. Replicating this diversity and flexibility in synthetic speech requires effective modeling and disentanglement of these factors [Wang et al., 2024, Meng et al., 2025]. The motivation for this work stems from three persistent challenges in the field: 1) Entanglement of Emotion and Speaking Style: Many TTS models intertwine speaker-specific emotion with prosodic style [Chen et al., 2024c], making it difficult to independently control voice identity and manner of speaking. This limitation restricts the personalization and expressiveness of synthesized voices, particularly in applications that require voice cloning or style transfer. 2) Balancing Prosody and Emotion Consistency: Achieving both natural prosody and consistent, expressive emotional content is difficult [Wu et al., 2019, Li et al., 2022]. Systems often excel at one aspect at the expense of the other, resulting in speech that sounds either monotonic or emotionally incongruent [Li et al., 2024c]. 3) Limitations of Conventional Emotion Modeling: Most existing TTS systems represent emotions using discrete categories (e.g., happy, sad, angry), which fails to capture the continuous

and multidimensional nature of real-world emotional expression. Moreover, these methods often struggle to maintain high speaker similarity when synthesizing emotional speech, especially in voice cloning scenarios [Li et al., 2021, Kansizoglou et al., 2022]. 4) Limited Availability of High-Quality Emotional Speech Data: Existing emotional speech datasets often suffer from limited speaker diversity, inconsistent recording conditions, or insufficient emotional coverage, particularly for non-English languages [Tits et al., 2020, Zhou et al., 2022, Ma et al., 2024], which constrains the development and evaluation of emotional TTS systems.

Existing solutions typically address these challenges by deploying separate modules for each function, such as distinct encoders for speaker and emotion or post-hoc prosody adjustment [Park et al., 2023, Jiang et al., 2024]. While modularization simplifies implementation, it often leads to weak interactions among features and degrades the overall synthesis quality [Diatlova and Shutov, 2023, Zhu et al., 2023, Choi et al., 2024]. For instance, the separation of speaker and emotion modules may result in unnatural blending or loss of speaker identity during expressive speech synthesis. Furthermore, the discrete treatment of emotions hinders the generation of subtle or mixed affective states, which are common in natural conversations.

To address these limitations, we built Marco-Voice, a TTS system unified emotional speech generation and voice cloning; and a emotional speech dataset named CSEMOTIONS. Our contributions in this paper is of two main parts:

1. Marco-Voice Model:

- We develop a speaker-emotion disentanglement mechanism that separates speaker identity from emotional expression, enabling independent control over voice cloning and emotional style. We also proposed to employ in-batch contrastive learning to further disentangle speaker identity with emotional style feature.
- We implement a rotational emotion embedding integration method to obtain emotional embeddings based on rotational distance from neutral embeddings. Finally, we introduce a cross-attention mechanism that better integrates emotional information with linguistic content throughout the generation process.

2. CSEMOTIONS Dataset:

- We construct CSEMOTIONS, a high-quality emotional speech dataset containing approximately 10 hours of Mandarin speech from ten professional native speakers (five male, five female), all with extensive voice acting experience. The dataset covers seven distinct emotional categories. All recordings were made in professional studios to ensure high audio quality and consistent emotional expression.
- We also develop 100 evaluation prompts for each emotion class across both existing datasets and CSEMOTIONS in English and Chinese, enabling thorough and standard assessment of emotional synthesis performance across all supported emotion categories.

Marco-Voice combines these innovations to deliver expressive, natural, and highly controllable speech synthesis. By integrating speaker identity, emotional style, and linguistic content within a single framework, our system achieves superior speech quality and emotional richness while expanding the potential applications of TTS technology in multilingual and interactive environments.

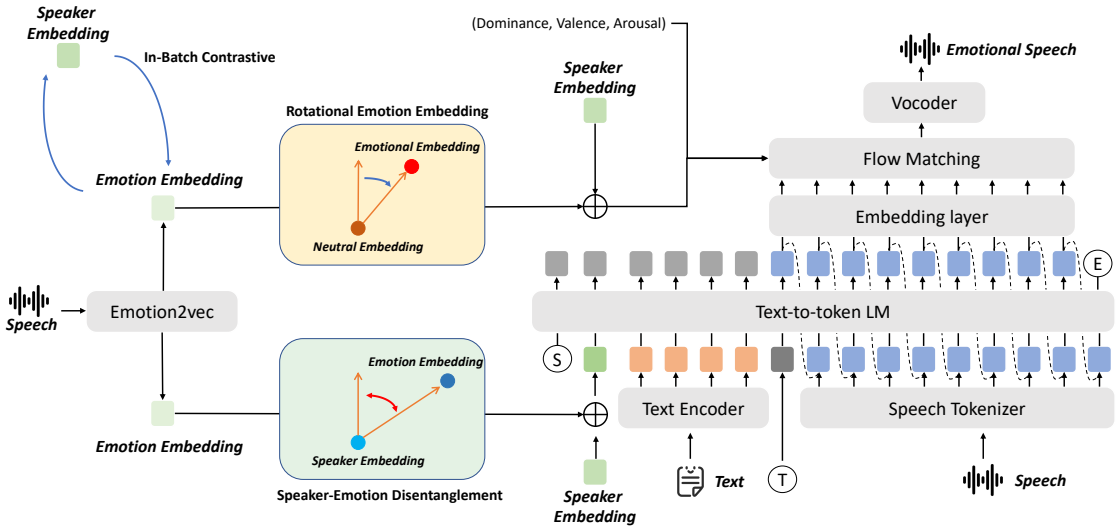

Figure 1 | The overall architecture of our Marco-Voice system, incorporating speaker-emotion disentanglement, in-batch contrastive learning.

2. Methodology

Our Marco-Voice system follows a clear pipeline: input text and reference speech are processed through separate encoders, while speaker and emotion information are embedded as conditioning signals. These features are fed into a language model that generates token representations. The emotion embeddings then interact with the LM outputs through cross-attention before being passed to a flow matching module, which generates high-quality expressive speech. The system incorporates several key innovations including speaker-emotion disentanglement, contrastive emotion learning, and adaptive cross-attention mechanisms.

2.1. System Architecture

The overall architecture of Marco-Voice consists of four main components: (1) input encoders that process text and speech separately, (2) embedding modules that encode speaker identity and emotional style, (3) a text-to-token language model that integrates linguistic and conditioning information, and (4) a flow matching module that generates acoustic parameters for final speech synthesis. The system is designed to handle multiple types of conditioning information while maintaining independent control over different aspects of speech generation.

2.2. Rotational Emotion Embedding Integration

We use an emotion feature extractor module to extract emotion embeddings from speech. We disentangle speaker-specific qualities and speaker-independent emotion representations by using paired samples of emotional speech x_i^e and neutral speech x_i^n from the same speaker. These are encoded using a pre-trained emotional encoder E_e to obtain representations $u_i^e = E_e(x_i^e)$ and $u_i^n = E_e(x_i^n)$.

We adopted the method intorduced by [Chen et al., 2024a] that hypothesizes that the difference between these encodings captures a direction vector in the speaker embedding space corresponding to the emotional content, while removing speaker identity:

$$v_i^e = \frac{u_i^e - u_i^n}{\|u_i^e - u_i^n\|} \tag{1}$$

We then aggregate over N such pairs to obtain a robust emotion embedding:

$$\mathbf{e} = \frac{1}{N}\sum_{i=1}^{N} v_i^e \tag{2}$$

For most cases, we find that single-shot ($N = 10$) suffices to produce high-quality emotion control. The resulting emotion embedding $\mathbf{e}$ serves as a conditioning signal at multiple stages, allowing the system to maintain emotional consistency from text processing through final speech generation.

2.3. Speaker-Emotion Disentanglement

We introduce a cross-orthogonal constraint to separate speaker identity from emotional expression. Given input features, we obtained speaker embeddings $\mathbf{s}$ using encoders E_s and E_e:

$$\mathbf{s} = E_s(\mathbf{x}) \tag{3}$$

where x is the speech and the emotion embedding $\mathbf{e}$ is obtained in Equation. 2. Our implementation computes the cross-orthogonality loss as follows. Given batch-wise speaker and emotion embeddings, we calculate the dot-product matrix, normalize by their vector norms, and compute the squared Frobenius norm. In addition, we calculate the average cosine similarity across all pairs in the batch, also using the squared Frobenius norm. The total orthogonality loss is a weighted sum of both terms:

$$\begin{aligned}
&\text{Let} \quad S \in \mathbb{R}^{B\times D}, \quad E \in \mathbb{R}^{B\times D} \\
&\text{Dot-Product Matrix:} \quad D = ES^T \\
&\text{Norms:} \quad n_E = \|E\|, \quad n_S = \|S\| \\
&\text{Normalized Matrix:} \quad \hat{D} = D/(n_E n_S^T) \\
&\mathcal{L}_{ort} = \|\hat{D}\|_F^2 + \|\text{mean}(\cos_\text{sim}(E,S))\|_F^2
\end{aligned} \tag{4}$$

where B is batch size, S and E consists of a batch of $\mathbf{s}$ and $\mathbf{e}$, $\|\cdot\|_F$ denotes the Frobenius norm, and $\cos\_\text{sim}(E,S)$ computes the cosine similarity between each emotion embedding and each speaker embedding. During training, if batchwise pairwise computation is enabled, we average the absolute dot product between each embedding and all opposing emotion embeddings in the batch, except self-pairs; otherwise, we use the orthogonality loss as above. This constraint forces speaker and emotion embeddings to be perpendicular in the feature space, enabling independent control over voice identity and emotional expression.

2.4. In-Batch Contrastive Learning

To improve the quality of emotion representations, we employ in-batch contrastive learning [Gao et al., 2021]. For each emotion embedding in a training batch, we encourage it to be dissimilar from other emotion embeddings that represent different emotional states.

Concretely, during training, for each pair in the minibatch, the speaker and emotion embeddings are projected and added, then, for all pairs (i, j) in the batch $(i \neq j)$, we accumulate the absolute dot products:

$$\mathcal{L}_{contrast} = \frac{1}{N(N-1)/2}\sum_{i<j} |\langle \mathbf{h}_i, \mathbf{e}_j \rangle| \tag{5}$$

where $\mathbf{h}_i$ is the sum of projected speaker and emotion embeddings for sample i, and $\mathbf{e}_j$ is the corresponding projected emotion embedding in the batch.

This batchwise contrastive learning encourages distinctiveness among emotion embeddings within the batch and enhances the separation of emotion representations.

2.5. Conditional Flow Matching Module

The conditional flow matching module [Lipman et al., 2023, Tong et al., 2024, Du et al., 2024a] processes transforms noise into speech parameters through a continuous flow, conditioned on all the input features. Specifically, we used an additional cross-attention mechanism such that the emotion embedding serves as the query (Q), and the acoustic token outputs from the language model serve as the keys (K) and values (V):

$$\begin{aligned} Q &= W_q(\mathbf{e}) \\ K &= W_k(\mathbf{h}_{LM}) \\ V &= W_v(\mathbf{h}_{LM}) \end{aligned} \tag{6}$$

where $\mathbf{h}_{LM}$ is the linguistic/acoustic token sequence generated by LLM, and W_q, W_k, W_v are learned projections. Then we compute:

$$h_{attn} = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

with optional masking for padding positions. The output dimensions and residual connections align with the input token sequence, allowing the emotion query to dynamically modulate the linguistic representations, thus enabling emotionally coherent speech synthesis.

The flow matching module takes the attended linguistic features $\mathbf{h}_{attn}$, along with speaker and emotion conditions, to generate acoustic parameters. The module uses a combination of Transformer and ResNet1D blocks to handle both sequential dependencies and local acoustic refinements:

$$\mathbf{h}_t = \text{FlowMatch}(\mathbf{h}_{attn}, \mathbf{h}_{LM}, \mathbf{e}) \tag{8}$$

This approach provides stable training while allowing flexible control over the generated speech characteristics.

2.6. Training Objective

The overall training objective combines the main TTS loss with regularization terms for disentanglement and contrastive learning:

$$\mathcal{L} = \mathcal{L}_{TTS} + \lambda_{orth}\mathcal{L}_{orth} + \lambda_{contrast}\mathcal{L}_{contrast} \tag{9}$$

where $\mathcal{L}_{TTS}$ is the main speech synthesis loss, and the λ terms control the relative importance of each constraint. The main TTS loss includes reconstruction, spectral, adversarial, and duration components to ensure high-quality speech generation.

3. Experimental Setup

3.1. Datasets

We combine established public corpora with our newly developed proprietary dataset to ensure diversity and robustness in our modeling.

Training Datasets

- **ESD (Emotional Speech Dataset):** ESD [Zhou et al., 2022] is a large-scale, high-quality resource specifically designed for emotional voice conversion and synthesis. It comprises approximately 29 hours of audio spanning five emotion categories (neutral, happy, angry, sad, and surprise) and features 20 professional speakers (10 native English and 10 native Chinese). Each speaker recorded 350 parallel utterances.
- **CSEMOTIONS (Chinese Speech Emotions):** To supplement existing public resources, we constructed **CSEMOTIONS**, a proprietary emotional speech dataset. This corpus includes about 10 hours of Mandarin speech from ten professional native speakers (five male, five female), all with extensive voice acting experience. CSEMOTIONS covers seven distinct emotional categories, with each speaker reading a curated set of 100 Chinese and English prompts. All recordings were made in professional studios to ensure high fidelity and expressive consistency.

Evaluation Datasets

- **LibriTTS:** For evaluation on English TTS and conversion tasks, we utilize LibriTTS [Zen et al., 2019], a large-scale, multi-speaker speech corpus derived from public domain audiobooks. We sample **400 prompts** from LibriTTS.
- **AISHELL-3:** To assess Mandarin performance, we employ AISHELL-3 [Shi et al., 2021], a multi-speaker Mandarin TTS dataset with approximately 85 hours of neutral speech from 218 native speakers. We select **400 prompts** from AISHELL-3 for evaluation.
- **CSEMOTIONS:** To comprehensively evaluate emotional expressiveness, we construct **100 prompts for each emotional class** (across both ESD and CSEMOTIONS) in both English and Chinese as dedicated evaluation data. This allows for targeted assessment of emotional synthesis and conversion across all supported emotion categories.

ESD and **CSEMOTIONS** serve as our primary training resources, providing rich emotional diversity. **LibriTTS** and **AISHELL-3** are employed for evaluation, with 400 prompts each used to benchmark model performance in English and Mandarin, respectively. Additionally, emotion-specific evaluation is conducted using the eval set of **CSEMOTIONS** - 100 prompts per emotional class, ensuring robust and fine-grained assessment of emotional speech capabilities. All audio was preprocessed to a consistent format (24/48kHz sampling rate, 16-bit depth) and normalized to control for volume variations.

3.2. Implementation Details

The model was implemented based on CosyVoice1 [Du et al., 2024a] and trained on 8 NVIDIA A100 GPUs for approximately couple hours. We used the Adam optimizer with a learning rate of 1×10^{-5} for llm part and 1×10^{-4} for flow matching part and a cosine decay schedule. The batch size was set to 32 per GPU. For the weighting factors in the loss function, we used $\lambda_{orth} = 0.1$ and $\lambda_{rot} = 0.5$. These values were determined through a hyperparameter search on a validation set.

3.3. Evaluation Metrics

We evaluated our system mainly based on human evaluation with additional automatic metrics for analysis to address the challenges for evaluating emotional speech generation with voice cloning:

- Speaker similarity was measured using a pre-trained speaker model [Ravanelli et al., 2024, Chen et al., 2024b] that computes cosine similarity between speaker embeddings.
- Emotional expressiveness was evaluated through human ratings on a 5-point Likert scale.
- Overall speech quality was assessed using mean opinion scores (MOS) from human listeners, as well as objective metrics including Whisper-WER [Radford et al., 2022] and DNS-MOS.

4. Results and Analysis

To comprehensively assess the effectiveness of the proposed Marco-Voice system, which integrates both voice cloning and emotional speech generation, we conduct evaluations targeting these two core capabilities. Given the subjective nature of speech quality, speaker similarity, and emotional expressiveness, we primarily rely on human evaluation, supplemented by automatic metrics where appropriate. This approach ensures a robust and representative assessment of system performance in both naturalness and controllability.

Human evaluations were conducted using a panel of native speakers who rated different systems across several dimensions on a five-point Likert scale (higher is better). For speaker similarity, listeners compared generated speech to reference samples from target speakers. For emotional expressiveness, raters assessed the naturalness, clarity, and emotional content of synthesized utterances. In addition, direct A/B preference tests were performed, where raters listened to paired samples and indicated their preference for each pair. Each system was evaluated using the same set of prompts to ensure fairness and comparability.

4.1. Voice Cloning Evaluation

Table 1 summarizes the results for voice cloning capabilities, including speech clarity, rhythm and speaking speed, naturalness, overall satisfaction, and speaker similarity. Only systems supporting voice cloning are included.

	Speech Clarity	Rhythm & Speed	Naturalness	Overall Satisfaction	Speaker Similarity
CosyVoice1	3.000	3.175	3.225	2.825	0.700
CosyVoice2	3.770	4.090	3.150	3.330	0.605
Marco-Voice	**4.545**	**4.290**	**4.205**	**4.430**	**0.8275**

Table 1 | Human evaluation results for voice cloning systems. Higher scores indicate better performance.

As shown in Table 1, Marco-Voice consistently outperforms existing voice cloning systems across all evaluated dimensions. Notably, our system achieves the highest speaker similarity score (0.8275), demonstrating its effectiveness in preserving speaker identity. Improvements in speech clarity, rhythm, and overall satisfaction further highlight the advantages of our speaker-style disentanglement approach.

4.2. Emotional Speech Generation Evaluation

Table 2 presents the evaluation results for systems supporting emotional speech generation, including speech clarity, emotional expression, rhythm and speaking speed, naturalness, and overall satisfaction.

	Speech Clarity	Emotional Expression	Rhythm & Speed	Naturalness	Overall Satisfaction
CosyVoice2	3.770	3.240	4.090	3.150	3.330
Marco-Voice	**4.545**	**4.225**	**4.290**	**4.205**	**4.430**

Table 2 | Human evaluation results for emotional speech generation. Higher scores indicate better performance.

According to Table 2, Marco-Voice achieves the best performance in all evaluated aspects, especially in emotional expression (4.225) and overall satisfaction (4.430). These results validate the effectiveness of our emotion modeling strategy, enabling more natural and expressive emotional speech synthesis compared to CosyVoice2 [Du et al., 2024b].

4.3. Direct Comparison (A/B) Tests

In addition to rating-based evaluations, we conducted A/B preference tests in which listeners compared pairs of samples from Marco-Voice and competing systems using the same prompts. The results, presented in Table 3, show the percentage of times Marco-Voice was preferred.

Compared Model	Marco-Voice Win Rate
CosyVoice1	60% (12/20)
CosyVoice2	65% (13/20)

Table 3 | A/B preference test results: percentage of times Marco-Voice was preferred in blind listening tests.

Marco-Voice is consistently preferred over all baseline systems in direct listening comparisons, indicating that listeners value the emotional expressiveness and speaker similarity of our system in direct comparisons.

4.4. Analysis Studies

To further investigate the performance of the Marco-Voice system, we conducted detailed analysis studies using objective metrics on both English (LibriTTS) and Mandarin (AISHELL) datasets. We compared multiple versions of Marco-Voice including: **v1** incorporates rotational emotion embeddings as conditioning signals in both the LLM and flow matching module. **v2** adds the cross-orthogonal constraint to enforce speaker-emotion disentanglement, enabling independent control over voice identity and emotional expression. **v3** employs in-batch contrastive learning between emotion and speaker embeddings. **v4** uses a cross-attention mechanism between emotion embeddings and language model tokens to ensure coherent emotion-text integration. The evaluation metrics are Word Error Rate (WER), Speaker Similarity (SS, using both SpeechBrain and ERes2Net), deletion and insertion errors (Del & Ins), substitution errors (Sub), and DNS-MOS scores for perceptual quality.

LibriTTS Results: Table 4 presents the results for the LibriTTS dataset. Across all Marco-Voice versions, the WER remains low and comparable to the best-performing baseline (CosyVoice1), with Marco-Voice-v4 achieving the lowest WER (11.4). Speaker similarity scores (both SS-speech brain and SS-ERes2Net) for Marco-Voice variants are consistently higher than CosyVoice2 and on par with

or slightly exceeding CosyVoice1. DNS-MOS scores for Marco-Voice models are also competitive, indicating strong perceptual quality.

System	CosyVoice1	CosyVoice1*	Marco-Voice-v1	Marco-Voice-v2	Marco-Voice-v3	Marco-Voice-v4
WER ↓	12.1	58.4	12.4	12.5	12.0	**11.4**
SS (SpeechBrain) ↑	64.1	61.3	64.2	**64.7**	64.5	63.2
SS (ERes2Net) ↑	80.1	64.2	**80.3**	79.5	80.1	74.3
Del & Ins ↓	**413**	2437	387.0	398.0	415.0	395.0
Sub ↓	251	2040.0	251.0	286.0	251.0	**242.0**
DNS-MOS ↑	3.899	3.879	**3.926**	3.900	3.923	3.860

Table 4 | Objective evaluation of speech recognition and synthesis quality on the LibriTTS dataset. Metrics include word error rate (WER), speaker similarity (SS) using SpeechBrain and ERes2Net, error counts (Del & Ins, Sub), and DNS-MOS for perceptual quality. Lower WER and error counts, and higher SS and DNS-MOS indicate better performance.

System	CosyVoice1	CosyVoice1*	Marco-Voice-v1	Marco-Voice-v2	Marco-Voice-v3	Marco-Voice-v4
WER ↓	**3.0**	23.3	17.6	15.9	18.2	17.6
SS (SpeechBrain) ↑	10.7	10.6	**11.0**	10.9	10.5	10.4
SS (ERes2Net) ↑	73.5	54.5	**73.8**	73.2	73.7	67.6
Del & Ins ↓	**11.0**	170.0	212.0	211.0	212.0	218.0
Sub ↓	**97.0**	674.0	485.0	408.0	496.0	471.0
DNS-MOS ↑	3.673	**3.761**	3.687	3.701	3.689	3.656

Table 5 | Objective evaluation of speech recognition and synthesis quality on the AISHELL dataset. Metrics are as in Table 4. Results demonstrate the effectiveness of Marco-Voice models for Mandarin emotional TTS, particularly in speaker similarity and perceptual quality. CosyVoice1* indicates that we continue training the base model on the same dataset, which typically leads to degraded WER performance and explains the higher WER observed in the Marco-Voice models.

AISHELL Results: Table 5 shows the results on AISHELL. Here, Marco-Voice variants generally outperform CosyVoice2 in WER, though CosyVoice1 achieves the lowest WER (3.0). Speaker similarity (SS) and DNS-MOS values for Marco-Voice remain strong, with SS-ERes2Net showing clear superiority over CosyVoice2. Notably, deletion and insertion errors are higher for Marco-Voice models, which can be attributed to challenges in emotional prompt synthesis and the presence of vocalized pauses (e.g., “ah,” “um”) that are often included in expressive and emotional speech but are not always reflected in text transcripts.

The observed WER values, while generally low, may appear elevated in some cases due to the inclusion of vocalized fillers and interjections inherent in natural, emotional speech. These elements are frequent in emotional prompts and can increase WER even when the generated speech is perceptually natural and expressive. Additionally, the complexity and variability of emotional text prompts pose extra challenges for TTS systems, potentially leading to more substitution, deletion, and insertion errors. Despite these difficulties, Marco-Voice demonstrates strong speaker similarity and perceptual quality across both English and Mandarin, validating the robustness and generalization of our approach in multilingual and emotionally-rich scenarios. Overall, these objective analysis studies complement our human evaluation findings, further confirming the effectiveness of Marco-Voice in both standard and emotionally challenging TTS scenarios.

Model Performance Progression Figure 2 shows that Marco-Voice-v4 achieves the best performance with 0.78 accuracy on Chinese and 0.77 on English datasets. CosyVoice1 provides a strong baseline (0.72 Chinese, 0.67 English), while CosyVoice2 shows performance degradation. The Marco-Voice

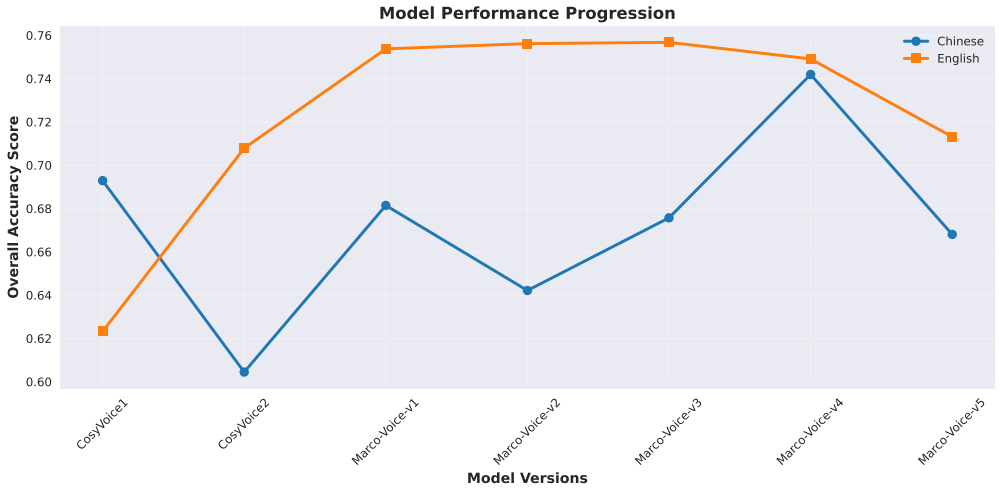

Figure 2 | Overall performance progression across model versions on Chinese and English datasets. The graph shows average accuracy scores across all emotions (excluding Playfulness) for emotion recognition tasks.

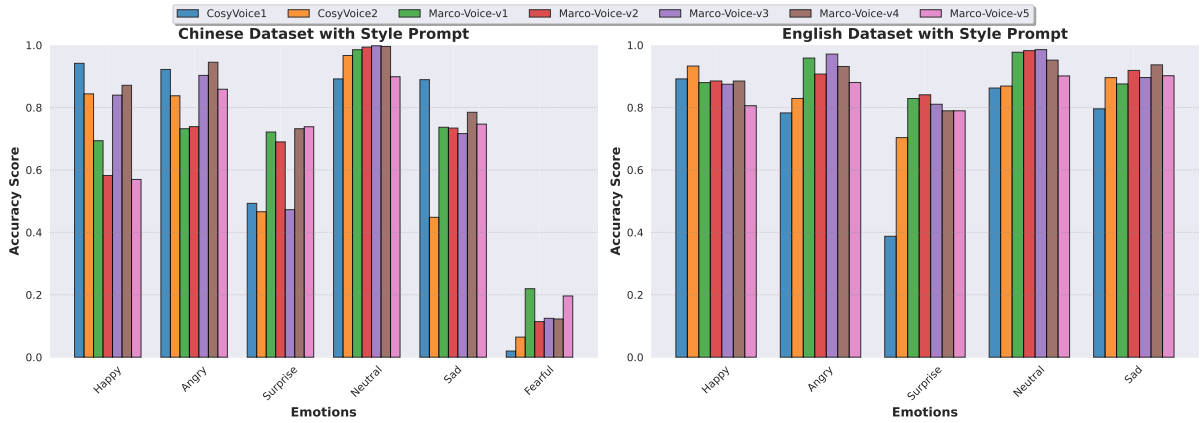

Figure 3 | Emotion recognition performance comparison between Chinese and English datasets with style prompts. Results show accuracy scores for six emotion categories across seven model variants using the emotion2vec_base_finetuned classifier.

series demonstrates clear progression from v1 to v4, with v5 showing slight decline, indicating that v4 represents the optimal balance of architectural improvements.

Crosslingual Emotion Recognition Figure 3 reveals that neutral and angry emotions achieve consistently high performance (>0.85) across both languages, while surprise and sad emotions remain challenging. Marco-Voice-v4 and v5 show superior performance for complex emotions, with accuracy scores above 0.73 for surprise recognition. The relatively balanced performance between Chinese and English suggests effective crosslingual generalization.

Language-Specific Patterns Figure 4 shows that Chinese datasets favor happy and angry emotion recognition, while English datasets perform better for neutral and sad emotions. The convergence of performance in advanced model versions (Marco-Voice-v4 and v5) suggests that architectural improvements can reduce language-specific biases and support more universal emotion recognition

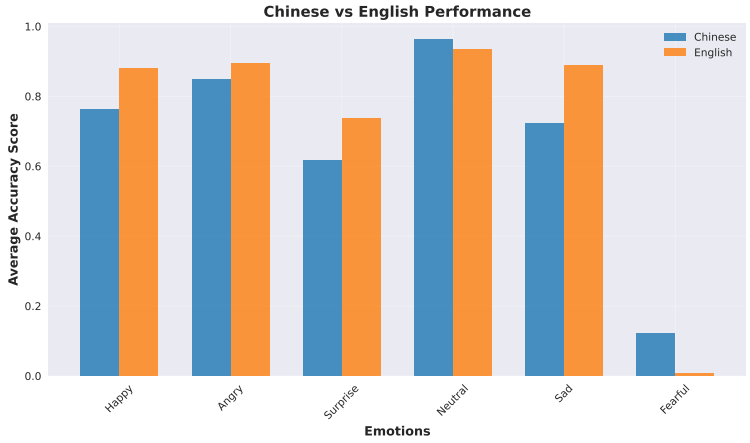

Figure 4 | Cross-language performance comparison showing average emotion recognition accuracy between Chinese and English datasets across all model versions.

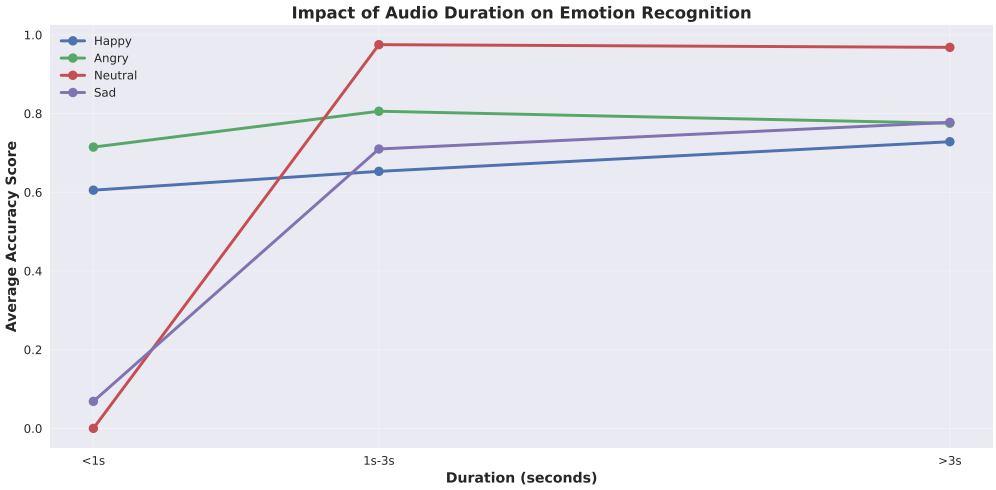

Figure 5 | Effect of audio duration on emotion recognition accuracy. Performance is evaluated across three duration categories: short (<1s), medium (1s-3s), and long (>3s) audio segments for four primary emotions.

systems.

Duration Impact on Recognition Figure 5 demonstrates that recognition accuracy increases substantially with audio duration. Short segments (<1s) show poor performance (<0.6), while medium duration (1s-3s) provides optimal efficiency with 0.6-0.8 accuracy. Long segments (>3s) achieve the highest performance but with diminishing returns, indicating 1s-3s as the practical sweet spot for real-time applications.

Gender Performance Disparity Figure 6 reveals significant gender bias, with male speakers showing substantially lower recognition accuracy across all emotions. Female speakers achieve 0.4+ accuracy for most emotions, while male speakers often fall below 0.2, particularly for surprise and sad emotions. This systematic bias indicates training data imbalances and highlights the need for gender-aware model development.

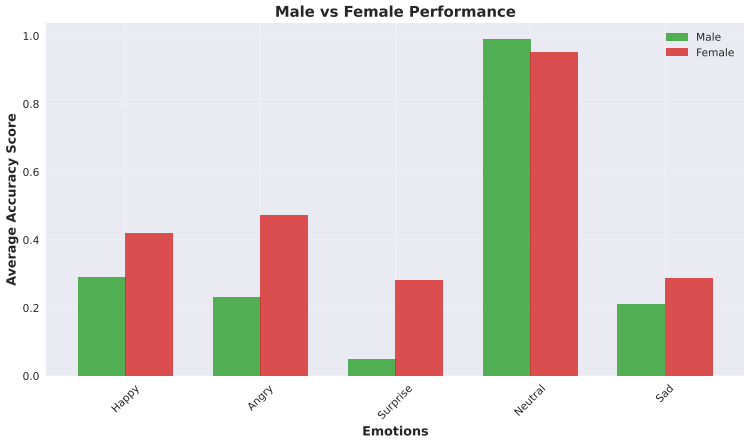

Figure 6 | Gender-based performance analysis showing emotion recognition accuracy differences between male and female speakers on the Chinese dataset.

5. Discussion

5.1. Benefits of Unified Modeling

Our results demonstrate the advantages of addressing voice cloning and emotional expression within a unified model rather than as separate components. The integrated approach allows the model to learn the subtle interactions between speaker characteristics and emotional expressions, resulting in more natural and consistent speech synthesis.

5.2. Limitations and Future Work

Despite the promising results, several limitations remain. First, the current model requires paired emotional speech data, which is scarce for many languages and domains. Future work could explore semi-supervised or self-supervised approaches to reduce this dependency. Second, computational efficiency remains a challenge, particularly for real-time applications. Exploring model compression techniques and optimized inference strategies would make the system more practical for deployment on resource-constrained devices.

6. Conclusion

In this paper, we presented Marco-Voice, a multifunctional speech synthesis system that achieves strong performance in voice cloning and emotion controllable speech generation. Through techniques including Rotational Emotion Embedding Integration and Speaker-Emotion Disentanglement as well as other training strategies, our system demonstrates substantial improvements over existing approaches with particular strengths in speaker similarity and emotional expressiveness. The system's unified approach to modeling various speech factors enables more natural and controllable speech synthesis than previous methods that treat these factors in isolation. This work represents an important step toward more expressive and personalized speech synthesis, with potential applications in virtual assistants, accessibility technologies, content creation, and human-computer interaction. Future research directions include expanding language support, reducing data requirements, and optimizing for real-time applications.

References

H. Barakat, O. Turk, and C. Demiroglu. Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):11, 2024.

H. Chen, R. Chen, and J. Hirschberg. EmoKnob: Enhance voice cloning with fine-grained emotion control. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8170–8180, Miami, Florida, USA, Nov. 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.466. URL https://aclanthology.org/2024.emnlp-main.466/.

Y. Chen, S. Zheng, H. Wang, L. Cheng, Q. Chen, S. Zhang, and J. Li. Eres2netv2: Boosting short-duration speaker verification performance with computational efficiency, 2024b. URL https://arxiv.org/abs/2406.02167.

Z. Chen, X. Li, Z. Ai, and S. Xu. Stylefusion tts: Multimodal style-control and enhanced feature fusion for zero-shot text-to-speech synthesis. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 263–277. Springer, 2024c.

H. Choi, J.-S. Bae, J. Y. Lee, S. Mun, J. Lee, H.-Y. Cho, and C. Kim. Mels-tts: Multi-emotion multi-lingual multi-speaker text-to-speech system via disentangled style tokens. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12682–12686. IEEE, 2024.

D. Diatlova and V. Shutov. Emospeech: Guiding fastspeech2 towards emotional text to speech. *arXiv preprint arXiv:2307.00024*, 2023.

Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma, Z. Gao, and Z. Yan. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens, 2024a.

Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang, F. Yu, H. Liu, Z. Sheng, Y. Gu, C. Deng, W. Wang, S. Zhang, Z. Yan, and J. Zhou. Cosyvoice 2: Scalable streaming speech synthesis with large language models, 2024b. URL https://arxiv.org/abs/2412.10117.

T. Gao, X. Yao, and D. Chen. SimCse: Simple Contrastive Learning of Sentence Embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

X. Jiang, X. Peng, Y. Zhang, and Y. Lu. Universal speech token learning via low-bitrate neural codec and pretrained representations. *IEEE Journal of Selected Topics in Signal Processing*, 2024.

I. Kansizoglou, E. Misirlis, K. Tsintotas, and A. Gasteratos. Continuous emotion recognition for long-term behavior modeling through recurrent neural networks. *Technologies*, 10(3):59, 2022.

J. Kim, J. Kong, and J. Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, 2021. URL https://arxiv.org/abs/2106.06103.

T. Li, S. Yang, L. Xue, and L. Xie. Controllable emotion transfer for end-to-end speech synthesis. In *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE, 2021.

T. Li, X. Wang, Q. Xie, Z. Wang, M. Jiang, and L. Xie. Cross-speaker emotion transfer based on prosody compensation for end-to-end speech synthesis. *arXiv preprint arXiv:2207.01198*, 2022.

W. Li, P. Yang, Y. Zhong, Y. Zhou, Z. Wang, Z. Wu, X. Wu, and H. Meng. Spontaneous style text-to-speech synthesis with controllable spontaneous behaviors based on language models. *arXiv preprint arXiv:2407.13509*, 2024a.

X. Li, F. Bu, A. Mehrish, Y. Li, J. Han, B. Cheng, and S. Poria. Cm-tts: Enhancing real time text-to-speech synthesis efficiency through weighted samplers and consistency models. *arXiv preprint arXiv:2404.00569*, 2024b.

X. Li, Z.-Q. Cheng, J.-Y. He, X. Peng, and A. G. Hauptmann. Mm-tts: A unified framework for multimodal, prompt-induced emotional text-to-speech synthesis. *arXiv preprint arXiv:2404.18398*, 2, 2024c.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Z. Ma, M. Chen, H. Zhang, Z. Zheng, W. Chen, X. Li, J. Ye, X. Chen, and T. Hain. Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. *arXiv preprint arXiv:2406.07162*, 2024.

M. Meng, Z. Yang, J. Yang, Z. Su, Y. Zhu, and Z. Fan. Ds-tts: Zero-shot speaker style adaptation from voice clips via dynamic dual-style feature modulation. *arXiv preprint arXiv:2506.01020*, 2025.

S. Park, M. Mark, B. Park, and H. Hong. Using speaker-specific emotion representations in wav2vec 2.0-based modules for speech emotion recognition. *Computers, Materials and Continua*, 77(1): 1009–1030, 2023.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision, 2022.

M. Ravanelli, T. Parcollet, A. Moumen, S. de Langen, C. Subakan, P. Plantinga, Y. Wang, P. Mousavi, L. D. Libera, A. Ploujnikov, F. Paissan, D. Borra, S. Zaiem, Z. Zhao, S. Zhang, G. Karakasidis, S.-L. Yeh, P. Champion, A. Rouhe, R. Braun, F. Mai, J. Zuluaga-Gomez, S. M. Mousavi, A. Nautsch, H. Nguyen, X. Liu, S. Sagar, J. Duret, S. Mdhaffar, G. Laperrière, M. Rouvier, R. D. Mori, and Y. Estève. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333), 2024. URL http://jmlr.org/papers/v25/24-0991.html.

K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*, 2023.

Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li. AIshEll-3: A multi-speaker Mandarin Tts Corpus and the Baselines, 2021.

X. Tan, T. Qin, F. Soong, and T.-Y. Liu. A survey on neural speech synthesis, 2021. URL https://arxiv.org/abs/2106.15561.

N. Tits, K. El Haddad, and T. Dutoit. Emotional speech datasets for english speech synthesis purpose: A review. In *Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 1*, pages 61–66. Springer, 2020.

A. Tong, K. Fatras, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport, 2024. URL https://arxiv.org/abs/2302.00482.

Z. Wang, L. Ma, Y. Feng, X. Pan, Y. Jin, and K. Zhang. Samoye: Zero-shot singing voice conversion model based on feature disentanglement and enhancement. *arXiv preprint arXiv:2407.07728*, 2024.

P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai. End-to-end emotional speech synthesis using style tokens and semi-supervised training. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 623–627. IEEE, 2019.

H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu. LibriTts: A corpus Derived from LibriSpeech for Text-to-Speech, 2019.

Z. Zeng, J. Wang, N. Cheng, T. Xia, and J. Xiao. Aligntts: Efficient feed-forward text-to-speech system without explicit alignment. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6714–6718. IEEE, 2020.

C. Zhang, C. Zhang, S. Zheng, M. Zhang, M. Qamar, S.-H. Bae, and I. S. Kweon. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai, 2023. URL https://arxiv.org/abs/2303.13336.

K. Zhou, B. Sisman, R. Liu, and H. Li. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18, 2022.

X. Zhu, Y. Lei, K. Song, Y. Zhang, T. Li, and L. Xie. Multi-speaker expressive speech synthesis via multiple factors decoupling. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.