Intelligent Interpretation of Traditional Folk Tales through Vocalization: a Generative Model Based on Emotion-Driven and Lip-Syncing

Sha Li

College of Arts and Sports, Henan Open University, Zhengzhou 450000, Henan, China

Corresponding author: Sha Li (e-mail: 18633448883@163.com)

ABSTRACT This study presents an integrated multi-modal signal generation framework for intelligent vocalization of traditional folk tales, ensuring emotion-driven speech synthesis with precise lip synchronization. The proposed system models continuous narrative emotion trajectories and embeds them into phoneme-level acoustic generation, using differentiable temporal alignment to synchronize audio and visual outputs. Multi-channel signals—including acoustic features and lip motion vectors—are processed through recursive smoothing and alignment matrices to maintain temporal consistency across modalities. Experimental results demonstrate median phoneme-duration deviations of 19.4–22.8 ms and lip-sync errors within 0.031–0.034, indicating stable audio-visual synchronization under complex narrative rhythms. By interpreting the system as a multi-channel signal processing and closed-loop control framework, the method mirrors principles of precision-engineered communication and control systems, providing insights into temporal alignment, cross-modal consistency, and real-time feedback optimization. This approach offers an engineering-oriented perspective for the development of advanced audio-visual generation, multi-sensor signal integration, and temporally coherent generative models.

INDEX TERMS emotion-driven generation, audio-visual signal processing, phoneme-level modeling, lip synchronization

I. INTRODUCTION

ALTHOUGH most generations of folk tales are coherent in terms of narration, current generation systems cannot accurately model the relationship between voice-over speech and lip movements, making the traditional folk tale less realistic and practical in virtual performances [1,2]. Research on multimodal narrative generation based on traditional folk tales will facilitate the development of richer multimodal storytelling systems [3,4].

Narrative emotion evolves over time and may extend across multiple sentences; recorded vocal expression causes fluctuations in pitch and rhythmic structure; and lip synchronization depends on accurate phoneme-level alignment. These three factors therefore interact and may conflict in the temporal domain [5,6]. Research has identified the major technical roadblocks in narrative generation as limited resolution of emotion annotations, separation of audio and video modeling, and the need for post-alignment processing [7,8]. Research has also explored other approaches to overcome these challenges, including emotional spoken-word synthesis [9], vocalization synthesis [10] and speech-driven lip animation [11]. Models creating emotional spoken language [12,13] allow for a greater variety of emotions generated in a single model than existing speech emotion models. Sing-song synthesis techniques [14,15] address the creation of melodies and rhythms while synthesizing lip movement corresponds to phonemes by synthesizing mouth shapes [16,17]. Although each of these areas has achieved significant progress, they have largely been developed separately and have not established a unified framework linking narrative emotion with acoustic and visual sequences across modalities, resulting in inconsistencies among the components [18,19].

This paper presents a sound and image generation approach that synchronizes emotion-driven vocalization and narrative-based lip-syncing. By treating audiovisual coordination as a multi-layered weaving process, the model ensures that digital storytelling and fabric-inspired textures converge into a seamless, co-modeled output. The method is grounded in the traditional structure of folktale texts, in which each narrative unit has a time-dependent representation of its emotional context. The time-dependent representation of emotion is used as a constraint signal during the generation of phoneme-level acoustic predictions to simultaneously generate pitch contours, phoneme duration and energy distribution relative to the temporal

structure of vocalization. This process generates a consistent acoustic intermediate representation and phoneme boundary information for visual lip-sync generation. The acoustic intermediate representation also provides the basis for synchronizing visual articulation units with the acoustic timeline. The differentiable temporal alignment constraints define the synchronization relationship between acoustic changes and lip movements during training. Unlike most models that treat vocal emotion prediction and lip-sync generation as separate tasks, this model provides an integrated framework for generating intelligent vocalizations from traditional folktales.

II. EMOTION-DRIVEN VOCAL-VISUAL JOINT GENERATIVE MODEL

The overall structure of the emotion-driven vocalization-visual joint generation model is shown in Figure 1, which demonstrates the collaborative relationship between narrative emotion modeling, vocalization-visual generation, and lip movement generation under a unified time constraint.

In Figure 1, the model takes traditional folk tale texts as input, constructs an emotional trajectory that evolves continuously as the narrative progresses through narrative semantic parsing, and uses it as a conditional signal throughout the generation process; emotions constrain the generation of vocalizations at the phoneme level, explicit duration modeling unifies phoneme and frame-level temporal structures, acoustic representations and phoneme boundaries directly drive lip-sync generation, and a differentiable temporal alignment mechanism is used to impose joint constraints on acoustic changes and lip movements, thereby forming an integrated generation framework for sound-visual collaboration.

A. NARRATIVE SEMANTIC ANALYSIS AND CONTINUOUS EMOTIONAL TRAJECTORY MODELING

Traditional folk tale texts are represented as character sequences $X = \{x_1, \ldots, x_N\}$, and narrative unit sequences $S = \{s_1, \ldots, s_M\}$ are obtained through a narrative unit segmentation model based on boundary prediction. The segmentation model extracts position-related semantic representations using a bidirectional context encoder and constrains the global consistency of unit boundaries with a conditional random field to obtain the context embedding representation h_m of each narrative unit. On this basis, emotional states are modeled as low-dimensional continuous vectors e_m, which are generated by a mapping function from semantic embedding to the emotional space. This mapping directly learns the deterministic correspondence between semantics and emotions through regression [20]:

$$e_m = W_e h_m + b_e \tag{1}$$

In equation (1), W_e represents the emotion projection matrix; b_e represents the emotion bias vector; and e_m represents the original emotion representation corresponding to the m-th narrative unit.

To characterize the continuous evolution of narrative emotion in the time dimension, a hidden state sequence z_m is introduced to dynamically model the emotion. The emotion state transition adopts a linear state space form, and the time smoothing constraint is achieved by minimizing the high-frequency fluctuations of emotion changes in adjacent units [21]:

$$z_m = A z_{m-1} + B e_m \tag{2}$$

In equation (2), A is the autoregressive emotion matrix; B is the semantic driving matrix; and z_m represents the continuous emotional latent state after time constraint. This modeling method maintains the consistency of emotional trends while suppressing the discrete jumps caused by narrative unit segmentation. To align the narrative-level emotional trajectory with the subsequent phoneme-level generation, linear interpolation [22] is used to map the latent state sequence z_m to the time-continuous emotional trajectory $E(t)$, and its first derivative is constrained by the emotional smoothing regularization term:

$$L_{\text{emo}} = \sum_t \left\| E(t) - \hat{E}(t) \right\|_2^2 + \lambda \left\| \partial_t E(t) \right\|_2^2 \tag{3}$$

In equation (3), $\hat{E}(t)$ represents the target emotion trajectory obtained by interpolation of the latent state of the narrative unit, and λ represents the temporal smoothing coefficient. This loss shares the time axis with downstream acoustics and lip-sync generation during the training phase, ensuring that the emotion trajectory maintains consistency and differentiability between the narrative scale and the phoneme scale.

B. PHONEME-LEVEL EXPRESSIVE SPEECH SYNTHESIS UNDER EMOTIONAL CONSTRAINTS

The expressive speech synthesis process uses a phoneme sequence $P = \{p_1, \ldots, p_K\}$ and its corresponding emotion trajectory $E(t)$ as joint input. The phoneme sequence is mapped to a discrete phoneme representation u_k through an embedding matrix, and conditionally fused with the emotion vector E_k, which is resampled along the phoneme time axis, on the feature dimension to form an emotion-modulated phoneme representation.

$$\tilde{u}_k = u_k + G_e E_k \tag{4}$$

In equation (4), G_e represents the emotion modulation matrix, and E_k represents the emotion vector aligned with the center of the k-th phoneme. This method introduces continuous emotion constraints at the phoneme level, avoiding temporal drift of the emotion signal in frame-level modeling [23,24].

The emotion-modulated phoneme representation sequence is fed into a temporal acoustic student generation network. Explicit duration modeling is used within the network to constrain the song-like rhythmic structure. Phoneme duration prediction is given by conditional regression:

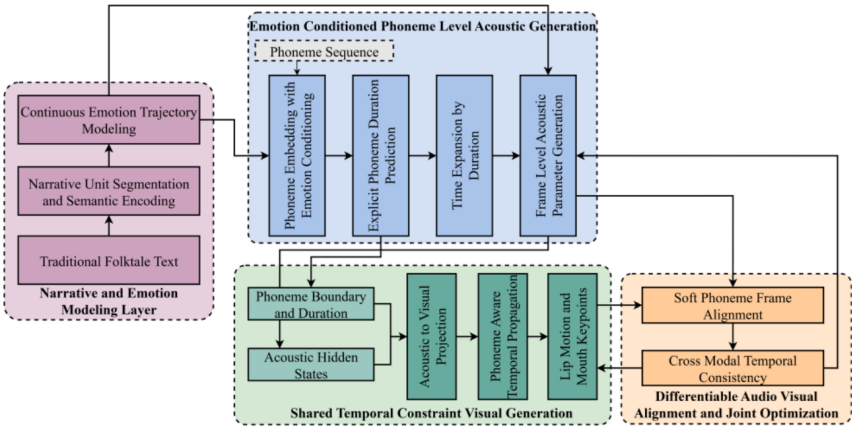

FIGURE 1. Overall Architecture of the Emotion-driven Vocal-visual Joint Generative Model

$$d_k = \phi_d \tilde{u}_k \tag{5}$$

In equation (5), d_k represents the predicted duration of the k-th phoneme, and ϕ_d represents the duration prediction mapping. This prediction result is used in the time-unfolding stage to map the phoneme-level implicit representation to the frame-level time axis, thus providing a unified time reference for the joint generation of pitch and energy [25,26].

On the frame-unfolded implicit representation sequence, acoustic parameters are predicted using a joint regression approach:

$$[a_t, f_t, e_t] = \phi_a h_t \tag{6}$$

In equation (6), h_t represents the acoustic latent state of frame t obtained from phoneme expansion; a_t represents the acoustic spectral feature; f_t represents the fundamental frequency; e_t represents the energy value; and ϕ_a represents the acoustic parameter mapping function. Emotional information is kept consistent throughout the temporal expansion process through modulation using Equation (4), ensuring that the pitch trend and energy distribution change in tandem with the emotional trajectory on the narrative scale [27,28]. Acoustic generation training employs a multi-objective joint loss, uniformly constraining the spectral features, fundamental frequency, and energy:

$$L_{\text{ac}} = \sum_t \|a_t - \hat{a}_t\|_1 + \alpha \left\| f_t - \hat{f}_t \right\|_2^2 + \beta \|e_t - \hat{e}_t\|_2^2 \tag{7}$$

In equation (7), $\hat{a}_t$, $\hat{f}_t$, and $\hat{e}_t$ represent the corresponding real acoustic parameters, and α and β represent weighting coefficients. This loss shares the time unfolding result with the emotion trajectory modeling, which makes the generation constraints of the singing speech consistent in terms of phoneme duration, prosodic structure and emotion modulation [29,30].

C. GENERATION OF LIP SHAPE MOTION WITH SHARED TEMPORAL CONSTRAINTS

The lip-sync generation takes the frame-level acoustic latent state sequence h_t and phoneme boundary information d_k obtained during the acoustic generation stage as input. The phoneme boundary information d_k represents the predicted duration of the k-th phoneme, and its cumulative form implicitly determines the boundary position of the phoneme on the frame-level time axis. To avoid temporal drift in the acoustic-to-visual mapping, the phoneme duration prediction results are directly used to construct the temporal mask for lip-sync generation, ensuring that the acoustic latent state maintains consistent temporal alignment weights within the phoneme duration interval [31,32]. Based on this constraint, the acoustic latent state is linearly mapped to obtain an intermediate representation of the lip-sync generation:

$$g_t = W_v h_t + b_v \tag{8}$$

The matrix W_v is used to associate audio and video with the bias term represented by b_v. The acoustic time axis of the lip-sync latent representation is shown as g_t. W_v and b_v make the two modalities of audio and video have the same time index (at the frame level) when trained together, making them remain synchronous to each other when trained together on [33,34].

The lip-sync keypoint sequence is represented as v_t, which is defined as a continuous vector sequence. The prediction of v_t is based on the assumption that the lip-sync latent representation changes smoothly over time. Therefore, to reduce jitter caused by random fluctuations at phoneme transitional points, a recursive structure is introduced based on phoneme masking.

$$v_t = \phi_v g_t + C v_{t-1} \tag{9}$$

In equation (9), ϕ_v is the mapping for lip-sync regression; C is the temporal smoothing matrix; and v_t is the vector of lip-sync keypoints at time t. The recursive expression governs how strongly the state will reset at the phoneme's boundary, depending on d_k. This ensures that lip-sync

transitions are synchronized with phoneme switches, and that no large discontinuities are created during this process [35,36].

The lip-sync generation training employs a temporal consistency-based regression loss, jointly constraining keypoint positions and motion smoothness:

$$L_{\text{lip}} = \sum_t \|v_t - \hat{v}_t\|_2^2 + \gamma \|v_t - v_{t-1}\|_2^2 \tag{10}$$

In equation (10), $\hat{v}_t$ represents the real lip-sync keypoints, and γ represents the motion smoothing weight. This loss shares the time index and phoneme boundary information with the acoustic generation stage, so that the lip-sync motion is always constrained by the acoustic temporal structure during the generation process, providing a consistent temporal basis for subsequent audio-visual joint optimization.

D. SOUND-VISUAL JOINT OPTIMIZATION MECHANISM BASED ON DIFFERENTIABLE ALIGNMENT

The acoustic parameter sequences a_t, f_t, e_t and the lip-sync keypoint sequence v_t have been assigned a unified temporal expansion index in the preceding stages, but the implicit propagation of the phoneme duration d_k in the acoustic and visual branches still exhibits cumulative bias. This cumulative bias originates from the nonlinear transformation of discrete phoneme boundaries and continuous acoustic features in the cross-modal temporal mapping process. Traditional attention mechanisms may attenuate temporal information through multilayer nonlinear transformations. The differentiable alignment mechanism proposed in this article constructs a soft alignment matrix from phoneme index to frame index as an intermediate variable, and its continuous processing keeps the gradient propagation smooth in the time dimension. This matrix directly constrains the correspondence between the acoustic change rate and the lip movement change rate on the phoneme scale, avoiding the accumulation of errors caused by multi-layer transformations. Unlike standard transformer attention, this method explicitly models the dynamic mapping between phoneme duration and lip movement, ensuring temporal consistency between acoustic features and visual performance in long sequences and effectively suppressing cross-modal drift. To directly constrain the consistency of the two modalities' responses to phoneme temporal sequences during generation, a joint optimization mechanism based on differentiable alignment is introduced. This mechanism uses the soft alignment matrix $M_{k,t}$ from the phoneme index to the frame index as an intermediary variable. This matrix is obtained by converting the temporal expansion results within the acoustic generation network into a continuous representation and satisfies the normalization constraint on the temporal dimension.

$$\sum_t M_{k,t} = d_k \tag{11}$$

In equation (11), $M_{k,t}$ represents the alignment weight of the k-th phoneme to the t-th frame, and d_k represents the predicted duration of the corresponding phoneme. This constraint ensures the interpretability and differentiability of the alignment relationship on the time scale.

Based on this, a consistency constraint is defined between the acoustic rate of change and the lip movement rate of change on the phoneme scale. The acoustic rate of change is represented by the first-order time difference of the spectral features, and the lip movement rate of change is represented by the time difference of the keypoint vectors. Both are mapped to a unified phoneme domain through an alignment matrix:

$$\Delta a_k = \sum_t M_{k,t}(a_t - a_{t-1}), \quad \Delta v_k = \sum_t M_{k,t}(v_t - v_{t-1}) \tag{12}$$

In equation (12), Δa_k represents the acoustic change vector corresponding to the k-th phoneme, and Δv_k represents the corresponding lip-sync change vector. This mapping enables direct comparability of cross-modal dynamic information at the phoneme scale. The transformation matrix $H \in \mathbb{R}^{D_c \times D_v}$ is a learnable parameter initialized with Xavier uniform distribution where D_c and D_v denote the dimensionalities of acoustic change vectors and lip-sync change vectors respectively ensuring dimensional compatibility between spectral features and geometric representations during cross-modal alignment.

Based on the above representation, an acoustic-visual temporal consistency loss is constructed to constrain the dynamic synchronization of acoustics and lip-sync at the phoneme boundary:

$$L_{\text{av}} = \sum_k \|\Delta a_k - H\Delta v_k\|_2^2 \tag{13}$$

Equation (13) indicates that the acoustic variation space is transformed by H, which maps lip-shape differences into the same modal space. This loss influences both the acoustic generation network and the lip-shape generating network during backpropagation, allowing phoneme duration, audio rhythm changes, and lip movements to align and converge within the same alignment structure.

The total loss consists of the acoustic loss (L_{ac}), lip-shape loss (L_{lip}) and audio-visual alignment loss (L_{av}):

$$L = L_{\text{ac}} + L_{\text{lip}} + \eta L_{\text{av}} \tag{14}$$

In equation (14), η represents the acoustic-visual alignment weight coefficient. This joint optimization maintains the shared constraints of acoustics and vision on the temporal structure of phonemes during the generation stage, enabling the singing speech and lip movements to maintain stable cross-modal consistency under continuous narrative conditions.

III. RESULTS ANALYSIS

A. ANALYSIS OF THE EFFECTIVENESS OF CONTINUOUS EMOTION TRAJECTORY MODELING

The experimental dataset is built upon the publicly available AISHELL-1 corpus containing 178 hours of Mandarin Chinese speech from 400 speakers. For this study, 2,400 narrative segments totaling 20 hours were selected and restructured to follow traditional Chinese folk tale structures, with annotations provided by 15 professional storytellers. The dataset was partitioned into training, validation, and test subsets with proportions of 80%, 10%, and 10% respectively while preserving consistent emotional category distribution across all partitions. The ground truth emotional trajectories for valence and arousal dimensions were manually annotated by the 15 professional storytellers using a continuous rating method during narrative delivery, with each segment independently evaluated three times and discrepancies resolved through consensus discussion to ensure annotation reliability.

This study investigates the ability of an emotion-driven generative neural architecture to model continuously evolving emotional vocal narratives in folk tales through a quantitative comparison of different strategies for responding to semantic shifts. Four models were employed in this comparative study: Model 1, the Emotion-Driven Generative Model; Model 2, Decoding Enhanced BERT with Disentangled Attention (DeBERTa); Model 3, Emoformer; and Model 4, the Bidirectional Long Short-Term Memory (Bi-LSTM) Network. Model 1 integrates narrative semantic parsing and continuous emotional space mapping into one complete emotion-driven generative model. Model 2 uses a regression network to map the semantic features extracted from a large pre-trained language model, Decoding Enhanced BERT with Disentangled Attention, into a continuous emotional space. Model 3 combines local features derived from attention and convolutional layers with long-range dependency representation techniques. Model 4 represents a traditional recurrent architecture using recurrent layers to form a complete temporal sequence of emotion-driven models within a sequence modeling framework. By mapping the predicted outputs of each model with the ground truth labeled trajectories on a standard time axis, the comparison results shown in Figure 2 are obtained.

As shown in Figure 2(a), the proposed model demonstrates high smoothness and close agreement with the ground-truth trajectory in the valence dimension. The model can accurately capture shifts in emotional polarity at plot turning points after a narrative semantic parsing mechanism is added, and the prediction curve continues to be consistent with the numerical fluctuations brought on by narrative conflict. The DeBERTa model, by contrast, struggles to capture the continuous emotional flow required for vocalization because its static representation produces substantial discrete noise in the output trajectory. In the arousal dimension (Figure 2(b)), the proposed model also maintains stable temporal consistency, effectively characterizing the changing trend of emotional energy during narrative progression. Emoformer and Bi-LSTM are prone to cumulative bias and response lag in long sequences, leading to a misalignment between the predicted peak and the actual emotional high point, making it difficult to maintain overall coordination in cross-modal generation. Experimental results show that the proposed model effectively alleviates the problems of sluggish response and unstable fitting in traditional architectures for nonlinear emotional evolution modeling, achieving a high degree of coupling between narrative logic and acoustic generation.

B. TEMPORAL MODELING AND ANALYSIS OF SINGULARIZED SPEECH

In the generation of singable speech, phoneme duration deviation can characterize the model's overall control over rhythmic structure. This paper compares four generation models based on phoneme-level temporal characteristics. Model 1 is a basic text-to-speech model without emotion or singable mechanisms. Model 2 introduces emotion-conditional encoding. Model 3 jointly models pitch and duration for singable expressions but does not explicitly constrain emotional continuity. The proposed model jointly predicts pitch, duration, and energy at the phoneme level while strengthening temporal consistency constraints. Furthermore, the distribution of phoneme duration deviation is statistically analyzed from three dimensions—phoneme category, true duration interval, and relative position within the sequence—as illustrated in Figure 3. Four models are compared: Model 1, basic text-to-speech model without emotion or singable mechanisms; Model 2, an emotion-conditional encoding model; Model 3, a joint pitch-duration modeling for singable expressions; and Model 4, the proposed emotion-driven singable model.

Figure 3 illustrates the distribution of phoneme-duration deviation in the vocalization models across all analysis dimensions. In the phoneme category analysis shown in Figure 3(a), the baseline model produced median deviation of 34.8 ms for monophthongs and 38.5 ms for diphthongs, whereas the proposed model reduced these values to 19.4 ms and 22.8 ms, respectively, indicating improved temporal control. In addition, all other models produced an increase in median deviation with longer phonemes as indicated by the true duration interval in Figure 3 (b). The median deviation of the vocalization model for all phonemes with actual durations greater than 300 ms was 58.2 ms, whereas the proposed model had a median deviation of 36.9 ms. Figure 3(c) shows that, across sequence positions, the median deviation of the baseline vocalization model increases from 27.6 ms in the first segment to 51.4 ms in the second segment, whereas the proposed model yields 17.3 ms and 32.9 ms, respectively, indicating that temporal error growth is suppressed. Overall, the proposed model demonstrates a more stable ability to model phoneme duration under different rhythmic conditions.

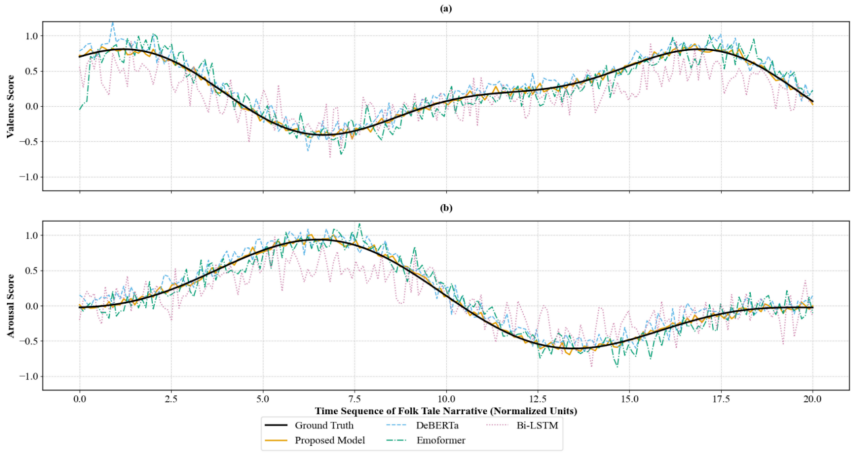

FIGURE 2. Comparison of Emotional Trajectory Fitting in the Narrative Process of Traditional Folk Tales: (a) Valence Dimension Mapping Distribution; (b) Arousal Dimension Mapping Distribution

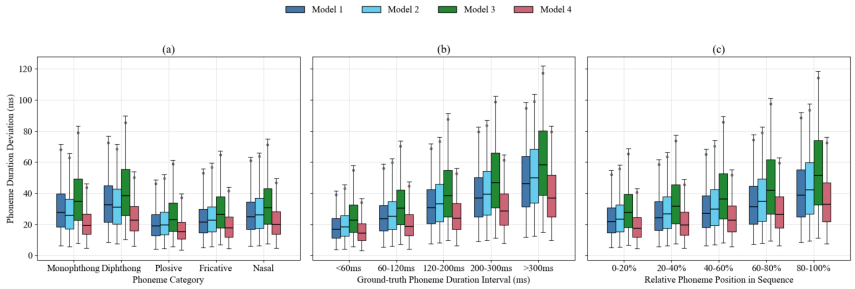

FIGURE 3. Analysis of the Distribution Characteristics of Phoneme Duration Deviation in the Generation of Sing-song Speech: (a) Different Phoneme Categories, (b) Different Phoneme True Duration Intervals, (c) Phoneme Sequence Position Segmentation

C. ANALYSIS OF LIP SYNCHRONIZATION ACCURACY

Lip-Sync Error is defined as the normalized Euclidean distance between predicted and ground-truth lip keypoint coordinates measured in pixels and normalized by face bounding box width to ensure scale invariance. In lip-sync modeling, the differences in temporal modeling granularity and constraint methods among generation paradigms are further amplified by the phoneme duration variations introduced by vocalization. Focusing on phoneme temporal structure, this paper compares several lip-sync generation models under a unified evaluation condition: Model 1 generates lip-sync based on phoneme-lip-sync rule mapping and fixed interpolation; Model 2, represented by Wav2Lip, directly regresses lip-sync from acoustic features but does not explicitly model phoneme boundaries and duration; Model 3 employs audiovisual joint generation with phoneme conditions, constraining lip-sync prediction with discrete phonemes but lacking cross-modal temporal joint optimization; the model in this paper introduces emotional conditions into phoneme-level acoustic modeling and constrains the consistency between phoneme duration and lip-sync movement through a differentiable temporal alignment mechanism. Based on the above models, the test set phonemes are divided into five categories according to duration: Very Short, Short, Medium, Long, and Very Long. The lip-sync error performance of each model under different phoneme duration conditions is compared, and the results are shown in Table 1.

As shown in Table 1, the lip-sync error rates of the comparative models increase monotonically with phoneme duration: the rule-driven model rises from 0.042 to 0.081, Wav2Lip from 0.038 to 0.071, and the phoneme-conditional audiovisual model from 0.035 to 0.060. The proposed model maintains error rates of 0.031-0.034 across all five duration conditions with standard deviation $\leq$ 0.009, indicating that the joint time alignment mechanism effectively suppresses synchronization drift caused by phoneme stretching, preserving audiovisual consistency.

D. CROSS-MODAL CONSISTENCY ANALYSIS IN LONG NARRATIVE PASSAGES

This study selects typical folktale segments with dramatic emotional fluctuations to verify the cross-modal synchronicity between vocalization features and visual representation in long narrative scenarios. Spectrograms and fundamental frequency curves were used to map the evolution of continuous emotional trajectories. The experiment demonstrated the real-time modulation effect of emotional intensity on phonetic features and geometric opening/closing by integrating narrative semantic stages, acoustic generation results, phoneme alignment boundaries, and lip-sync wireframes corresponding to key moments along a single timeline. Figure 4 displays pertinent generation details.

TABLE 1. Comparison of Lip-sync Errors under Different Phoneme Time Conditions in a Narrative Singing Scenario

Phoneme Temporal Condition	Proportion (%)	Phoneme-based Rule Alignment	Wav2Lip	Phoneme-conditioned AV	Proposed model
Very Short	19.2	0.042 ± 0.011	0.038 ± 0.010	0.035 ± 0.009	0.032 ± 0.007
Short	21.0	0.048 ± 0.013	0.043 ± 0.012	0.039 ± 0.010	0.031 ± 0.007
Medium	22.4	0.056 ± 0.016	0.049 ± 0.014	0.044 ± 0.012	0.031 ± 0.006
Long	19.1	0.067 ± 0.021	0.059 ± 0.019	0.051 ± 0.015	0.032 ± 0.007
Very Long	18.3	0.081 ± 0.028	0.071 ± 0.025	0.060 ± 0.020	0.034 ± 0.009

Note: Proportion (%) represents the percentage of all phonemes in the test set that fall under the corresponding phoneme time condition.

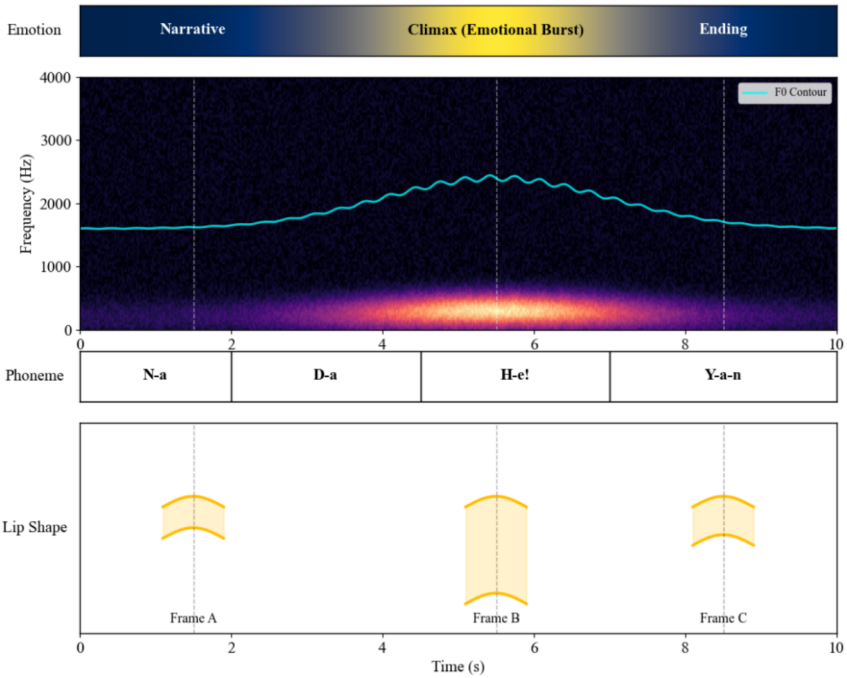

FIGURE 4. Alignment Analysis of Spectroscopic Features and Lip-Sync Sequences Driven by Continuous Emotions in Long Narrative Passages

Figure 4 illustrates how the fundamental frequency curve shows a notable upward trend accompanied by frequency oscillations as the story reaches its climax, or the peak of emotional arousal. High-emotion-load phonemes exhibit more exaggerated mouth openings within the temporal boundaries due to the dramatic expansion of the spectrogram energy distribution, which causes greater geometric deformation in the lip-sync generation network. This results from the differentiable alignment mechanism on the audiovisual representation and the synergistic constraint of the model's internal emotion modulation signal. A high level of consistency in cross-modal performance during the vocalization process is ensured by this interconnected evolution, which is fueled by the depth of narrative semantics.

IV. CONCLUSION

This paper addresses the unified challenge of integrating continuous emotional development and acoustic rhythm with lip-sync in the intelligent interpretation of traditional folk tales. This study proposes a method for generating sound-visual outputs guided by semantically rich narratives. The method optimizes both vocalization and visual synchronizations, produces phoneme-level acoustic representations based on a continuous emotional trajectory, and completes the lip-sync modeling within the same temporal structure. The research finds that the proposed system maintains a stable acoustic temporal structure and lip-sync consistency despite complex narrative rhythms, much like a high-precision loom ensuring structural integrity amidst intricate pattern shifts. This integration of emotional modeling and technical control enables high-realism generation of traditional folk tales for virtual presentation and dissemination.

AUTHOR CONTRIBUTIONS

Sha Li designed, collected and analyzed the data, and drafted the manuscript. Sha Li conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Sha Li participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research was supported by:

1) Research on the Vocalization of Traditional Folk Stories within the Scope of Vision Protection.
2) Research on the Cultivation of Self-accompaniment and Singing Skills among College Music Majors in Henan Province.
3) The Influence of Music on Shen Congwen's Thought and Creation.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] K. Liu, J. Wei, and R. Yuan, "Text-guided dynamic mouth motion capturing for person-generic talking face generation," Knowledge-Based Systems, vol. 316, no. 1, pp. 113354-113380, 2025, doi: 10.1016/j.knosys.2025.113354.

[2] X. Wang, Y. Huo, and Y. Liu, "Multimodal Feature-Guided Audio-Driven Emotional Talking Face Generation," Electronics, vol. 14, no. 13, pp. 2684-2704, 2025, doi: 10.3390/electronics14132684.

[3] J. Wang, O. Zhang, and Y. Jiang, "Multimodal diffusion framework for collaborative text image audio generation and applications," Scientific Reports, vol. 15, no. 1, pp. 20604-20618, 2025, doi: 10.1038/s41598-025-05794-4.

[4] J. Xu, X. Wu, and Z. Zhang, "MARS: Multimodal-Assisted Refined Semantic Alignment," Information Processing & Management, vol. 63, no. 1, 104292, 2026, doi: 10.1016/j.ipm.2025.104292.

[5] D. Jiang, J. Chang, and L. You, "Audio-Driven Facial Animation with Deep Learning: A Survey," Information, vol. 15, no. 11, pp. 675-698, 2024, doi: 10.3390/info15110675.

[6] X. Ji, Z. Liao, and L. Dong, "3D facial animation driven by speech-video dual-modal signals," Complex & Intelligent Systems, vol. 10, no. 5, pp. 5951-5964, 2024, doi: 10.1007/s40747-024-01481-5.

[7] L. Yu, H. Xie, and Y. Zhang, "Multimodal learning for temporally coherent talking face generation with articulator synergy," IEEE Transactions on Multimedia, vol. 24, no. 1, pp. 2950-2962, 2021.

[8] J. Salas-Caceres, J. Lorenzo-Navarro, and D. Freire-Obregon, "Multimodal emotion recognition based on a fusion of audio visual information with temporal dynamics," Multimedia tools and applications, vol. 84, no. 23, pp. 27327-27343, 2025, doi: 10.1007/s11042-024-20227-6.

[9] S. Liang, R. Zhou, and Q. Yuan, "ECE-TTS: A Zero-Shot Emotion Text-to-Speech Model with Simplified and Precise Control," Applied Sciences, vol. 15, no. 9, pp. 5108-5129, 2025, doi: 10.3390/app15095108.

[10] S. Hwang J, H. Lee S, and W. Lee S, "HiddenSinger: High-quality singing voice synthesis via neural audio co-dec and latent diffusion models," Neural Networks, vol. 181, no. 1, pp. 106762-106771, 2025, doi: 10.1016/j.neunet.2024.106762.

[11] R. Wu, Y. Yu, and F. Zhan, "Audio-driven talking face generation with diverse yet realistic facial animations," Pattern Recognition, vol. 144, no. 1, pp. 109865-109891, 2023, doi: 10.1016/j.patcog.2023.109865.

[12] X. Luo, S. Takamichi, and Y. Saito, "Emotion-controllable Speech Synthesis Using Emotion Soft Label, Utterance-level Prosodic Factors, and Word-level Prominence," APSIPA Transactions on Signal and Information Processing, vol. 13, no. 1, pp. 1-30, 2024, doi: 10.1561/116.00000242.

[13] W. Zhao and Z. Yang, "An emotion speech synthesis method based on vits," Applied Sciences, vol. 13, no. 4, pp. 2225-2236, 2023, doi: 10.3390/app13042225.

[14] Y. Hono, K. Hashimoto, and K. Oura, "Sinsy: A deep neural network-based singing voice synthesis system," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, no. 1, pp. 2803-2815, 2021, doi: 10.1109/TASLP.2021.3104165.

[15] Y. Zhao, M. Yang, and Y. Lin, "AI-enabled text-to-music generation: A comprehensive review of methods, frameworks, and future directions," Electronics, vol. 14, no. 6, pp. 1197-1249, 2025, doi: 10.3390/electronics14061197.

[16] L. Liu, J. Wang, and S. Chen, "VividWav2Lip: high-fidelity facial animation generation based on speech-driven lip synchronization," Electronics, vol. 13, no. 18, pp. 3657-3675, 2024, doi: 10.3390/electronics13183657.

[17] Y. Fan, Z. Lin, and J. Saito, "Joint audio-text model for expressive speech-driven 3d facial animation," Proceedings of the ACM on Computer Graphics and Interactive Techniques, vol. 5, no. 1, pp. 1-15, 2022, doi: 10.1145/3522615.

[18] X. Liu, X. Liu, and P. Yang, "An approach to optimizing semantic consistency for text-to-digital human generation," Engineering Applications of Artificial Intelligence, vol. 160, no. 1, 111909, 2025, doi: 10.1016/j.engappai.2025.111909.

[19] G. Chen, X. Liang, and W. He, "Enhancing Human-Computer Interaction Through Decoupling Motion and Camera Control in Human-Centric Video Generation," International Journal of Human-Computer Interaction, vol. 1, no. 1, pp. 1-15, 2025, doi: 10.1080/10447318.2025.2487877.

[20] J. Yang, J. Liu, and K. Huang, "Single-and cross-lingual speech emotion recognition based on WavLM domain emotion embedding," Electronics, vol. 13, no. 7, pp. 1380-1395, 2024, doi: 10.3390/electronics13071380.

[21] T. Lodewyckx, F. Tuerlinckx, and P. Kuppens, "A hierarchical state space approach to affective dynamics," Journal of mathematical psychology, vol. 55, no. 1, pp. 68-83, 2011, doi: 10.1016/j.jmp.2010.08.004.

[22] S. Sadok, S. Leglaive, and L. Girin, "A multimodal dynamical variational autoencoder for audiovisual speech representation learning," Neural Networks, vol. 172, no. 1, pp. 106120-106135, 2024, doi: 10.1016/j.neunet.2024.106120.

[23] 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, vol. 2024, no. 1, pp. 11-44, 2024, doi: 10.1186/s13636-024-00329-7.

[24] C. Shen, L. Zhao, and C. Fu, "Transinger: Cross-Lingual Singing Voice Synthesis via IPA-Based Phonetic Alignment," Sensors, vol. 25, no. 13, pp. 3973-3991, 2025, doi: 10.3390/s25133973.

[25] Y. Yao, T. Liang, and R. Feng, "SR-TTS: a rhyme-based end-to-end speech synthesis system," Frontiers in Neurorobotics, vol. 18, no. 1, pp. 1322312-1322322, 2024, doi: 10.3389/fnbot.2024.1322312.

[26] F. Khanam, A. Munmun F, and A. Ritu N, "Text to speech synthesis: a systematic review, deep learning based architecture and future research direction," Journal of Advances in Information Technology, vol. 13, no. 5, pp. 1-15, 2022, doi: 10.12720/jait.13.5.398-412.

[27] Z. Chen, Z. Ai, and Y. Ma, "Optimizing feature fusion for improved zero-shot adaptation in text-to-speech synthesis," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2024, no. 1, pp. 28-45, 2024, doi: 10.1186/s13636-024-00351-9.

[28] C. Galdino J, N. Matos A, and F. Svartman F R, "The evaluation of prosody in speech synthesis: a systematic review," Journal of the Brazilian Computer Society, vol. 31, no. 1, pp. 465-486, 2025, doi: 10.5753/jbcs.2025.5468.

[29] K. Zhou, B. Sisman, and R. Rana, "Speech synthesis with mixed emotions," IEEE Transactions on Affective Computing, vol. 14, no. 4, pp. 3120-3134, 2022, doi: 10.1109/TAFFC.2022.3233324.

[30] W. Byun S and P. Lee S, "Design of a multi-condition emotional speech synthesizer," Applied Sciences, vol. 11, no. 3, pp. 1144-1154, 2021, doi: 10.3390/app11031144.

[31] D. Pawar, P. Borde, and P. Yannawar, "Generating dynamic lip-syncing using target audio in a multimedia environment," Natural Language Processing Journal, vol. 8, no. 1, pp. 100084-100094, 2024, doi: 10.1016/j.nlp.2024.100084.

[32] Y. Tang, Y. Liu, and W. Li, "Continuous Talking Face Generation Based on Gaussian Blur and Dynamic Convolution," Sensors, vol. 25, no. 6, pp. 1885-1899, 2025, doi: 10.3390/s25061885.

[33] Z. Wang, W. He, and Y. Wei, "Flow2Flow: Audio-visual cross-modality generation for talking face videos with rhythmic head," Displays, vol. 80, no. 1, 102552, 2023, doi: 10.1016/j.displa.2023.102552.

[34] Z. Sheng, L. Nie, and M. Liu, "Toward fine-grained talking face generation," IEEE Transactions on Image Processing, vol. 32, no. 1, pp. 5794-5807, 2023, doi: 10.1109/TIP.2023.3323452.

[35] Z. Liu, X. Liu, and S. Chen, "Multimodal fusion for talking face generation utilizing speech-related facial action units," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 20, no. 9, pp. 1-24, 2024.

[36] Z. Sheng, L. Nie, and M. Zhang, "Stochastic latent talking face generation toward emotional expressions and head poses," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 4, pp. 2734-2748, 2023, doi: 10.1109/TCSVT.2023.3311039.