Professional English 2025 Presentation

From Speaker to Dubber: Movie Dubbing with Prosody and Duration Consistency Learning

WangGuokang YunXin YangYihui

School of Computer Science and Technology in USTC

2025 年 4 月 25 日

Outline

1. Background
2. Paper Overview
3. MTSP
4. PCL
5. DCR
6. Experiments
7. Summary

Outline

1 Background

2 Paper Overview

3 MTSP

4 PCL

5 DCR

6 Experiments

7 Summary

Uniqueness of Movie Dubbing Task

Movie dubbing: Convert scripts to speeches, aligning with clip in timing and emotion, while preserving reference audio's timbre.[5]

- **Traditional VC/TTS**: Rely on the input text for modeling.
- **Movie dubbing**: Align with clip in emotion, rate, lip movements; preserve vocal timbre.
- **Transformed to one-to-one mapping.**

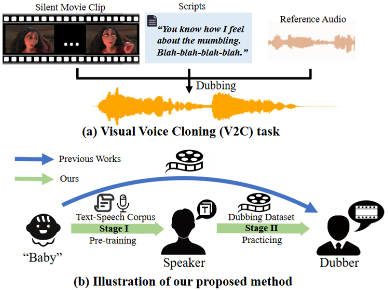

Figure 1: Movie dubbing vs. traditional VC/TTS

Objectives of Movie Dubbing Task

The model needs to adapt to changes in emotion and speech speed to keep pronunciation accurate.

- **Timbre**: Match dubbed speech to reference audio.
- **Emotion & Rate**: Align with movie characters' performance.
- **Lip Sync**: Match phoneme duration to lip movements.

Related Work

The Visual Voice Cloning (V2C) task [1] requires the generated dubbing to align with the video content in terms of lip movements, emotions, and duration, which makes traditional speech synthesis methods inapplicable.

- **Speech Synthesis:** FastSpeech[3] series excel in speech synthesis but lack video content alignment, making them unsuitable for V2C.
- **Visual Voice Cloning (V2C) :** Requires lip sync, emotion, and duration alignment with video, increasing task complexity.[1]
- **Pre-training in TTS:** Strategies like MP-BERT [4] and PLBERT [2] enhance speech naturalness via phoneme-level modeling, adopted in V2C to improve dubbing quality.

Outline

1. Background
2. Paper Overview
3. MTSP
4. PCL
5. DCR
6. Experiments
7. Summary

Task Objective Description

Task Goal: Generate dubbing that matches reference audio's timbre and aligns with movie clip's emotions and lip movements.

- **Timbre Matching:** Align with reference audio.
- **Emotion & Lip Sync:** Match movie clip's emotions and lip movements.
- **Accurate Pronunciation:** Adapt to emotional and pacing variations.

Task Description with Formula for Movie Dubbing Task :

$$\tilde{A}_{Dub} = Model(A_{Ref}, T_s, V_{Ref}) \tag{1}$$

Overall Method Framework Overview

Main Framework:

- **MTSP:** Multi-Task Speaker Pre-training.Includes TTS task and MLM task
- **Dubbing Training:** Includes PCL (Prosody Consistency Learning) and DCR (Duration Consistency Reasoning).

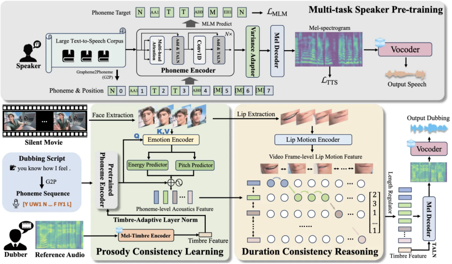

Figure 2: Main architecture of the two-stage dubbing method.

Detailed Description of the Overall Method Framework

- **Stage 1: Multi-Task Speaker Pre-training (MTSP)**
  - **TTS Task:** Learn accurate pronunciation using FastSpeech2[3] to predict target speech mel-spectrogram.
  - **MLM Task:** Predict masked phonemes to capture phoneme context for unseen text.
- **Stage 2: Dubbing Training**
  - **PCL:** Enhance audiovisual consistency by aligning movie clip emotions with phoneme-level prosody.
  - **DCR:** Ensure dubbing duration matches video content by reasoning phoneme-lip alignment.

Outline

1. Background
2. Paper Overview
3. MTSP
4. PCL
5. DCR
6. Experiments
7. Summary

Multi-task Speaker Pre-training (MTSP)

Goal: Improve pronunciation clarity and naturalness via multi-task learning.

- **TTS Task:** Learn accurate pronunciation from text-to-speech corpus.
- **MLM Task:** Capture phoneme context to handle unseen text.

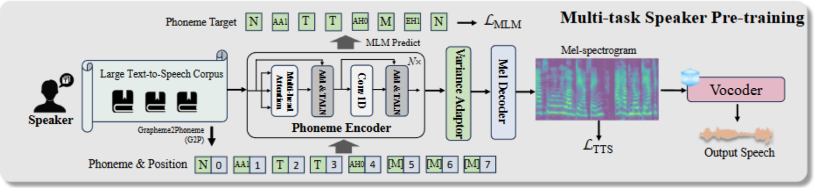

Figure 3: Illustration of MTSP

TTS Task in MTSP

Objective: Learn accurate pronunciation from text-to-speech corpus using FastSpeech2-like architecture.

- Convert text to phoneme sequence:
  $T_p = \text{G2P}(T_o)$
- Extract phoneme embeddings:
  $T_e = \text{PhonemeEncoder}(T_p)$
- Model prosody attributes:
  $T_{\text{mel}} = \text{VarianceAdaptor}(E_{\text{ph}}, D_{\text{ph}}, P_{\text{ph}}, E_{\text{ph}})$

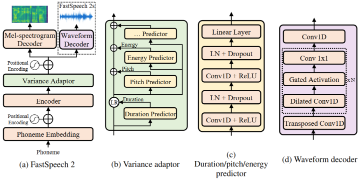

Figure 4: FastSpeech2 Architecture

Then predict the mel-spectrogram of the target speech and calculate the loss for the TTS task.

MLM Prediction Task in MTSP

The MLM prediction task helps the model learn contextual relationships between phonemes.

- **Input:** Randomly masked phoneme sequence.
- **Processing:** Input to the phoneme encoder to predict the masked phonemes.
- **Output:** Predict the masked phonemes using linear projection and softmax function.

MLM Prediction Task:

- Predict the masked phonemes:

$$L_{\mathrm{MLM}} = \mathrm{CE}(\mathrm{PhonemeEncoder}(T_{\mathrm{masked}}), T_{\mathrm{target}}) \tag{2}$$

Summary of MTSP

- **MTSP** combines **TTS** and **MLM** tasks to enhance pronunciation quality and expressiveness.
- Total loss function integrates losses of both tasks:

$$\mathcal{L}_{MTSP} = \alpha_1 \cdot \mathcal{L}_{TTS} + \alpha_2 \cdot \mathcal{L}_{MLM}, \tag{3}$$

where α_1 and α_2 are hyperparameters adjusting task weights.

Outline

1. Background
2. Paper Overview
3. MTSP
4. PCL
5. DCR
6. Experiments
7. Summary

Overview of Prosody Consistency Learning (PCL)

Objective: Enhance audiovisual consistency of dubbing.

- **Method:** Model phoneme-level prosody using emotional facial expressions. Key components:

- **Emotion-Prosody Alignment:** Align character emotions with phoneme-level prosody via cross-modal attention.
- **Timbre Consistency:** Maintain reference audio's timbre using TALN.

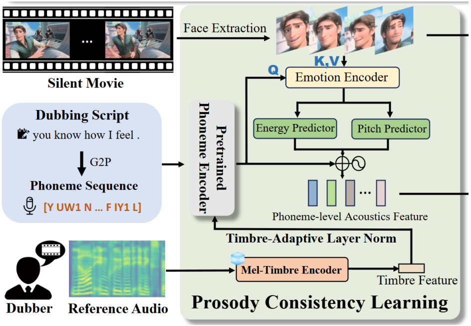

Figure 5: PCL module illustration

Prosody Consistency Learning (PCL)

(1) Alignment of Emotion and Prosody Attributes

- **Objective:** Align character emotions with phoneme-level prosody (pitch and energy) while maintaining reference audio's timbre.
- **Steps:**
  1. **Extract Emotion Features:** Detect faces with S3FD, encode emotions with EmoFAN.
  2. **Cross-modal Attention:** Align emotion features with phoneme prosody.
  3. **Predict Phoneme Attributes:** Embed predicted pitch and energy into phoneme sequence.
- **Step 1: Emotion Feature Extraction**
  - Detect facial regions using S3FD:

$$V_{\text{face}} = S^3FD(V_{\text{Ref}}) \in \mathbb{R}^{L_v \times H_{\text{face}} \times W_{\text{face}} \times C} \tag{4}$$

  - Encode emotions using EmoFAN:

$$V_{\text{emo}} = \text{EmoFAN}(V_{\text{face}}) \in \mathbb{R}^{L_v \times d_m} \tag{5}$$

Prosody Consistency Learning (PCL)

(1) Alignment of Emotion and Prosody Attributes

Step 2: Cross-modal Attention Mechanism

- Use multi-head cross-modal attention to align the emotional features of the character with the prosody attributes of each phoneme:

$$\xi^k_{\text{pho,pitch}} = \text{softmax}\left(\frac{Q^\top K_p}{\sqrt{d_m}}\right) V_p \in \mathbb{R}^{L_p \times \frac{d_m}{n_{\text{head}}}} \tag{6}$$

$$\xi^k_{\text{pho,energy}} = \text{softmax}\left(\frac{Q^\top K_e}{\sqrt{d_m}}\right) V_e \in \mathbb{R}^{L_p \times \frac{d_m}{n_{\text{head}}}} \tag{7}$$

- Where:

$$Q = W_j^Q T_e^\top, \quad K_p = W_j^{K_p} V_{\text{emo}}^\top, \quad V_p = W_j^{V_p} V_{\text{emo}}^\top \tag{8}$$

$$K_e = W_j^{K_e} V_e^\top, \quad V_e = W_j^{V_e} V_{\text{emo}}^\top \tag{9}$$

Prosody Consistency Learning (PCL)

(1) Alignment of Emotion and Prosody Attributes

Step 3: Prediction of Phoneme Pitch and Energy

- Use pitch and energy predictors to predict the pitch and energy of each phoneme, convert them into pitch and energy embeddings, and then add them to the phoneme sequence:

$$\tilde{P}_{\text{pho}}, \tilde{E}_{\text{pho}} = \text{Predictor}(\xi_{\text{pho,pitch}}, \xi_{\text{pho,energy}}) \in \mathbb{N}^{L_p} \tag{10}$$

- Add pitch and energy embeddings to the phoneme sequence:

$$T_a = T_e + \text{PitchEmb}(\tilde{P}_{\text{pho}}) + \text{EnergyEmb}(\tilde{E}_{\text{pho}}) \tag{11}$$

Prosody Consistency Learning (PCL)

(2) Timbre Consistency

- **Objective:** Replicate reference audio's timbre accurately.
- **Method:** Use TALN to integrate timbre features into phoneme encoding and mel-spectrogram generation.
- **Formula:**

$$\mathrm{TALN}(x, E_{\mathrm{timbre}}) = \mathrm{gain}(E_{\mathrm{timbre}})\frac{x-\mu}{\sigma} + \mathrm{bias}(E_{\mathrm{timbre}}) \tag{12}$$

- **Parameters:**
  - x: Input sequence.
  - E_{timbre}: Timbre feature.
  - μ, σ: Mean and variance of x.
  - $\mathrm{gain}(\cdot)$, $\mathrm{bias}(\cdot)$: Predicted gain and bias.
- **Application:** Apply TALN to each FFT block of phoneme encoder and mel-decoder.

Outline

1. Background
2. Paper Overview
3. MTSP
4. PCL
5. DCR
6. Experiments
7. Summary

Overview of Duration Consistency Alignment

Objective: Enhance temporal consistency of dubbed audio.

Module Composition:

- **Lip Motion-Phoneme Alignment:** Extract lip motion features and align with phoneme features.
- **Phoneme Duration Expansion:** Use dynamic programming to optimize phoneme durations.
- **Mel-Spectrogram Duration Expansion:** Adjust mel-spectrogram length based on audio duration.

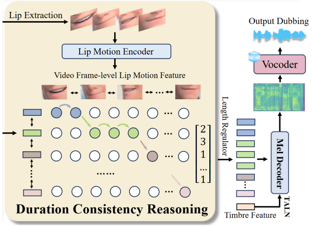

Figure 6: DCR module illustration

Duration Consistency Alignment (DCR)

(1) Lip Motion-Phoneme Alignment

Step 1: Lip Motion-Phoneme Alignment

- **Objective:** Extract lip motion features from the reference video and align them with phoneme features.
- **Steps:**
  1. Extract the lip motion region from the video:

$$V_{\text{lip}} \in \mathbb{R}^{L_v \times H_{\text{lip}} \times W_{\text{lip}} \times C} \tag{13}$$

  2. Obtain the lip motion representation using a lip motion encoder:

$$E_{\text{lip}} \in \mathbb{R}^{L_v \times d_{\text{model}}} \tag{14}$$

Duration Consistency Alignment (DCR)

(2) Phoneme Duration Expansion

Step 2: Phoneme Duration Expansion

- **Objective:** Expand phoneme durations to match lip motion features.
- **Steps:**
  1. **Similarity Matrix Calculation:** Compute the similarity matrix between phoneme-level acoustic features and lip motion features.

$$S_{\text{pho,lip}} = \text{Similarity}(T_a^i, E_{\text{lip}}^j) \tag{15}$$

  2. **Dynamic Programming Alignment:** Use dynamic programming to find the optimal alignment.

$$A_{i,j} = \begin{cases} \text{None}, & \text{if } i > j \text{ or } j - i < L_p - L_{\text{p}} \\ \max(A_{i-1,j}, A_{i-1,j-1}) + s_{i,j}, & \text{otherwise} \end{cases} \tag{16}$$

Duration Consistency Alignment (DCR)

(3) Mel-Spectrogram Duration Expansion

➢ Step 3: Mel-Spectrogram Duration Expansion

- **Objective:** Expand the length of the mel-spectrogram to match the audio duration.
- **Steps:**
  1. **Fixed Ratio Relationship:** Expand the length of the mel-spectrogram based on audio duration.

$$n = \frac{L_{\text{mel}}}{L_0} = \frac{sr/hs}{FPS} \in \mathbb{N}^* \tag{17}$$

  2. **Mel-Spectrogram Generation:** Use a length regulator to generate the mel-spectrogram of the desired length:

$$\tilde{A}_{\text{Dub}} = \text{Vocoder}(\text{MelDecoder}(LR(T_a, A^* \times n), E_{\text{imbr}})) \tag{18}$$

Outline

1. Background
2. Paper Overview
3. MTSP
4. PCL
5. DCR
6. Experiments
7. Summary

Experimental Results: Comparison with SOTA

➤ Key metrics maintain high levels across multiple datasets

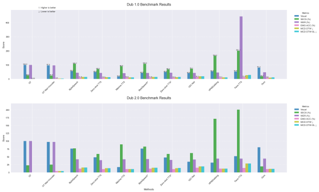

Figure 7: V2C-A Dataset Test Results

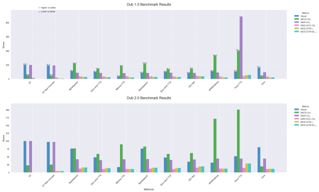

Figure 8: GRID Dataset Test Results

Dataset	V2C-Animation			GRID		
Methods	NMOS ↑	SMOS ↑	CMOS ↑	NMOS ↑	SMOS ↑	CMOS ↑
GT	4.52±0.13	-	+0.23	4.69±0.07	-	+0.14
GT Mel + Vocoder	4.39±0.16	4.41±0.18	+0.21	4.66±0.08	4.53±0.10	+0.16
StyleSpeech [34]	3.34±0.13	3.37±0.14	-0.22	3.56±0.14	3.60±0.19	-0.25
Zero-shot TTS [62]	3.38±0.14	3.50±0.19	-0.26	3.57±0.12	3.54±0.13	-0.23
StyleSpeech* [34]	3.31±0.21	3.35±0.12	-0.20	3.50±0.10	3.58±0.11	-0.24
Zero-shot TTS* [62]	3.40±0.12	3.47±0.18	-0.24	3.58±0.21	3.52±0.15	-0.21
V2C-Net [4]	3.54±0.16	3.51±0.18	-0.21	3.62±0.06	3.67±0.11	-0.19
HPMDubbing [6]	3.57±0.17	3.54±0.12	-0.18	3.77±0.20	3.74±0.13	-0.14
Face-TTS [26]	3.18±0.13	3.24±0.16	-0.37	3.39±0.21	3.32±0.17	-0.32
Ours	**3.92±0.19**	**3.87±0.14**	**0.00**	**4.03±0.09**	**4.05±0.11**	**0.00**

Figure 9: Zero-shot Test

Method	Visual	SECS ↑	WER ↓	NMOS ↑	SMOS ↑	CMOS ↑
StyleSpeech [34]	X	55.81	93.40	3.49±0.17	3.52±0.21	-0.19
Zero-shot TTS [62]	X	57.23	31.47	3.53±0.16	3.56±0.11	-0.18
StyleSpeech* [34]	✓	58.71	105.64	3.51±0.12	3.52±0.23	-0.21
Zero-shot TTS* [62]	✓	61.12	35.10	3.54±0.21	3.57±0.12	-0.16
V2C-Net [4]	✓	39.43	143.54	3.61±0.22	3.64±0.17	-0.14
HPMDubbing [6]	✓	49.31	106.45	3.62±0.16	3.61±0.23	-0.11
FaceTTS [26]	✓	33.80	231.63	3.46±0.09	3.51±0.17	-0.29
Ours	✓	**73.44**	**16.05**	**3.85±0.12**	**3.87±0.09**	**0.0**

Figure 10: Subjective Evaluation Results

Experimental Results: Qualitative Analysis of Mel Spectrograms

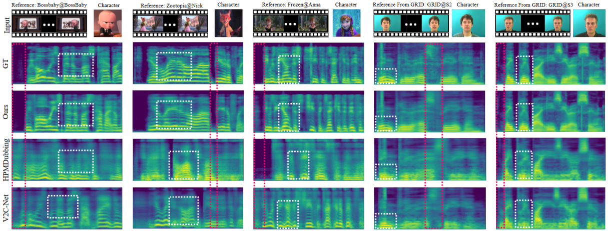

Figure 11: Comparison of Mel Spectrograms of Audio Generated by Different Models

- **Analysis Results:**
  - Red Box: the model better maintains phoneme and pause durations, especially in V2C-Animation benchmark.
  - White Box: the model shows clearer and more natural pronunciation details.

Experimental Results: Ablation Study

The ablation study results fully demonstrate the necessity of each module.

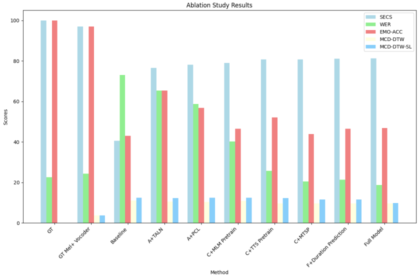

Figure 12: Comparison of Ablation Study Results

- **PCL Module:** Enhances timbre cloning and emotional consistency.
- **MTSP Module:** Reduces WER, improves pronunciation quality.
- **DCR Module:** Reduces MCD-DTW-SL, enhances duration consistency.

Outline

1 Background
2 Paper Overview
3 MTSP
4 PCL
5 DCR
6 Experiments
7 Summary

Summary of the Paper

- **Research Contributions**:
  - Proposed a two-stage movie dubbing method enhancing timbre, emotion, and duration alignment via MTSP and dubbing training.
  - PCL module improves emotional consistency by aligning emotions with phoneme-level prosody.
  - DCR module achieves precise duration matching by aligning lip motion with phoneme features.
- **Technical Innovations**:
  - MTSP integrates TTS and MLM tasks, improving pronunciation quality and contextual understanding.
  - TALN integrates timbre features into phoneme encoding and mel-spectrogram generation, ensuring timbre consistency.
- **Experimental Results**:
  - Method outperforms state-of-the-art methods on V2C-Animation and GRID datasets.

Reflections on the Paper

- **Learning Method from 'Simple' to 'Complex'**:
  - Uses a two-stage framework: pre-trains on TTS data for clear pronunciation, then fine-tunes on dubbing data for emotion and lip-sync alignment.
  - Addresses data scarcity and complexity, improving dubbing quality.
- **Implications for Practical Applications**:
  - Pre-training on general data followed by fine-tuning on specialized data enhances model performance.
  - Provides insights for multimodal alignment tasks like virtual anchors and customer service.
- **Future Research Directions**:
  - Optimize two-stage learning by adding more modalities (e.g., gestures, facial expressions) during pre-training.
  - Introduce user feedback mechanisms to further optimize dubbing effects.

References:

[1] Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu. V2c: Visual voice cloning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21242–21251, 2022.

[2] Yinghao Aaron Li, Cong Han, Xilin Jiang, and Nima Mesgarani. Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[3] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

[4] Guangyan Zhang, Kaitao Song, Xu Tan, Daxin Tan, Yuzi Yan, Yanqing Liu, Gang Wang, Wei Zhou, Tao Qin, Tan Lee, et al. Mixed-phoneme bert: Improving bert with mixed phoneme and sup-phoneme representations for text to speech. *arXiv preprint arXiv:2203.17190*, 2022.

[5] Zhedong Zhang, Liang Li, Gaoxiang Cong, Haibing Yin, Yuhan Gao, Chenggang Yan, Anton van den Hengel, and Yuankai Qi. From speaker to dubber: movie dubbing with prosody and duration consistency learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7523–7532, 2024.

Thanks!