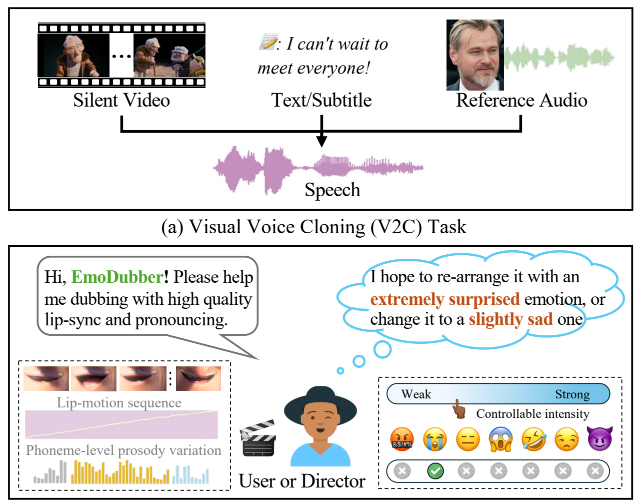
EmoDubber: Towards High Quality and Emotion Controllable Movie Dubbing

Gaoxiang Cong[1*] Jiadong Pan[1*] Liang Li[1†] Yuankai Qi[2] Yuxin Peng[3]
Anton van den Hengel[4] Jian Yang[2] Qingming Huang [1]
[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS
[2]Macquarie University [3]Peking University [4]University of Adelaide
Project page with demo: https://galaxycong.github.io/EmoDub

Abstract

Given a piece of text, a video clip, and a reference audio, the movie dubbing task aims to generate speech that aligns with the video while cloning the desired voice. The existing methods have two primary deficiencies: (1) They struggle to simultaneously hold audio-visual sync and achieve clear pronunciation; (2) They lack the capacity to express user-defined emotions. To address these problems, we propose EmoDubber, an emotion-controllable dubbing architecture that allows users to specify emotion type and emotional intensity while satisfying high-quality lip sync and pronunciation. Specifically, we first design Lip-related Prosody Aligning (LPA), which focuses on learning the inherent consistency between lip motion and prosody variation by duration level contrastive learning to incorporate reasonable alignment. Then, we design Pronunciation Enhancing (PE) strategy to fuse the video-level phoneme sequences by efficient conformer to improve speech intelligibility. Next, the speaker identity adapting module aims to decode acoustics prior and inject the speaker style embedding. After that, the proposed Flow-based User Emotion Controlling (FUEC) is used to synthesize waveform by flow matching prediction network conditioned on acoustics prior. In this process, the FUEC determines the gradient direction and guidance scale based on the user's emotion instructions by the positive and negative guidance mechanism, which focuses on amplifying the desired emotion while suppressing others. Extensive experimental results on three benchmark datasets demonstrate favorable performance compared to several state-of-the-art methods.



(a) Visual Voice Cloning (V2C) Task

(b) Illustration of our proposed method (EmoDubber)

Figure 1. (a) Illustration of the V2C tasks. (b) EmoDubber can help users achieve audio-visual sync and maintain clear pronunciation (left), while controlling the intensity of emotions according to the user's intentions (right).

1. Introduction

Movie Dubbing, also known as Visual Voice Cloning (V2C), aims to convert a script into speech with the voice characteristics specified by the reference audio, while aligning lip-sync with the silent video (see Figure 1 (a)). V2C is far more challenging than conventional text-to-speech (TTS), but has a host of applications, not least in repurposing the vast volumes of existing video to be reproduced by movie creators or enthusiasts.

Existing dubbing methods broadly fall into two groups. The first focuses primarily on learning and applying effective speaker style representations. For example, V2C-Net [3] and VDTTS [15] utilize a pre-trained GE2E [47] to obtain unique and normalized utterance embedding. StyleDubber [6] applies multi-scale learning to apply style cues to the embeddings extracted by GE2E, while Speak2Dub [50] adopts pre-training strategy on TTS corpus to improve expressiveness. Because these methods rely on simple MSE-based loss [3, 50] and overall scaling [6],

*Equal contribution.
†Corresponding author.

their capacity to establish accurate correspondence between lip motion and speech is limited. The second group of methods aim to generate proper prosody by incorporating visual information from the provided video [6, 16, 24, 51]. For example, HPMDubbing [6] applies visual information to speech prosody through hierarchical modeling of lips, face, and scene. Recently, MCDubber [51] operates at the level of context sequence with previous and following facial information to enhance generated speech prosody. These methods work at the video-frame and mel-spectrogram levels, however, which means they ignore the role of phoneme-level pronunciations, and thus often generate seemingly mumbled articulation.

Except for the above shorts, most current dubbing models suffer from rigid emotional expression due to lacking controllability. Some studies [18] suggest that the emotional intensity in dubbing can affect the listener's emotions and psychological perception of the film. Since the film's post-production can make up for the deficiencies of previous recordings, especially in emotional expression, the actors need to re-record in studio according to the director's instructions until they meet the requirements. Unlike previous methods, our model not only satisfies the basic function (lip-sync and clear pronunciation), but also learns to control the attribute and intensity of emotions to meet customized needs, as shown in Figure 1 (b).

In this paper, we propose a novel movie dubbing architecture named EmoDubber, which achieves emotion synthesis with controllable intensity while maintaining audio-visual alignment and clear pronunciation. Specifically, we first design a Lip-related Prosody Aligning (LPA) module that controls speech speed by duration level contrastive learning between lip motion and phoneme prosody sequence, which helps the model to reason the correct audio-visual aligning. Second, we propose the Pronunciation Enhancing (PE) strategy, which focuses on expanding video-level phoneme sequences by monotonic alignment search and fusing it with the output of LPA by an efficient conformer to improve pronunciation. Next, the speaker identity adapting module is used to absorbs the fused sequence from PE and injects style embedding from the reference speaker to target acoustics prior information. Finally, the proposed Flow-based User Emotion Controlling (FUEC) aims to generate waveform by optimal-transport conditional flow matching based on acoustic priors while rendering the user-specified emotion. It is worth noting that we propose positive and negative guidance mechanisms (PNGM) in FUEC to allow user control emotional intensity flexibly, which determines the gradient direction of emotion generation and adjusts the dual guidance scale based on the user's emotion prompt, amplifying the target emotion and suppressing others.

The contributions of this paper are summarized below:

- We propose EmoDubber, a controllable emotion dubbing architecture to help users specify the emotion they need while satisfying high-quality lip sync and pronunciation.
- We design a FUEC with positive and negative guidance to dynamically adjust the flow-matching vector field prediction process to achieve intensity control flexibly.
- We simultaneously achieve high-quality lip sync and clear pronunciation by aligning duration-level contrastive learning and phoneme-enhancing strategy.
- Extensive experimental results demonstrate the proposed Emodubber performs favourablly against state-of-the-art models on three benchmark datasets.

2. Related Work

2.1. Visual Voice Cloning

The V2C requires generating a waveform representing how a text might be said, but in step with the lip movements portrayed by a character, and in vocal style exemplified by reference audio. Some works focus on improving speaker identity to handle multi-speaker scenes [3, 7, 15, 28, 50]. For example, Speak2Dub [50] introduces speaker embedding extracted by pre-trained GE2E to phoneme encoder and mel-spectrogram decoder by learnable style affine transform, while StyleDubber [7] propose multi-scale style adaptor with phoneme and utterance level to strengthen speaker's characteristics. Besides, some works attempt to combine visual representation to enhance prosody expressiver [6, 16, 24, 51]. For example, HPMDubbing [6] is a hierarchical dubbing method by bridging acoustic details with visual information: lip motion, face region, and scene. To improve the contextual prosody, MCDubber [51] enlarges the modeling object from single sentence to previous and following sentences, which incorporates more contextual video scenes. Although the speaker identity and prosody modeling have received attention, existing works still suffer from poor lip-sync and lifeless emotional expression, which is unacceptable in dubbing. In this work, we propose EmoDubber, a controllable emotional dubbing architecture to help users specify emotion they need in video, while bringing high quality lip-sync and pronunciation.

2.2. Flow Matching and Classifier Guidance

Flow Matching [27] is a simulation-free method to train Continuous Normalizing Flows (CNFs) [4] models, which model arbitrary probability path and capture the probability trajectories represented by diffusion processes [43]. It has demonstrated exceptional performance in image generation and geometric domains, such as Stable Diffusion 3 [10], Lumina-T2X [11], and EQUIFM [44]. Due to its advantages of high sampling speed and generation quality, flow matching has attracted significant attention in audio generation [14, 23, 34]. Recently, Matcha-TTS [34] introduces optimal-transport conditional flow matching (OT-

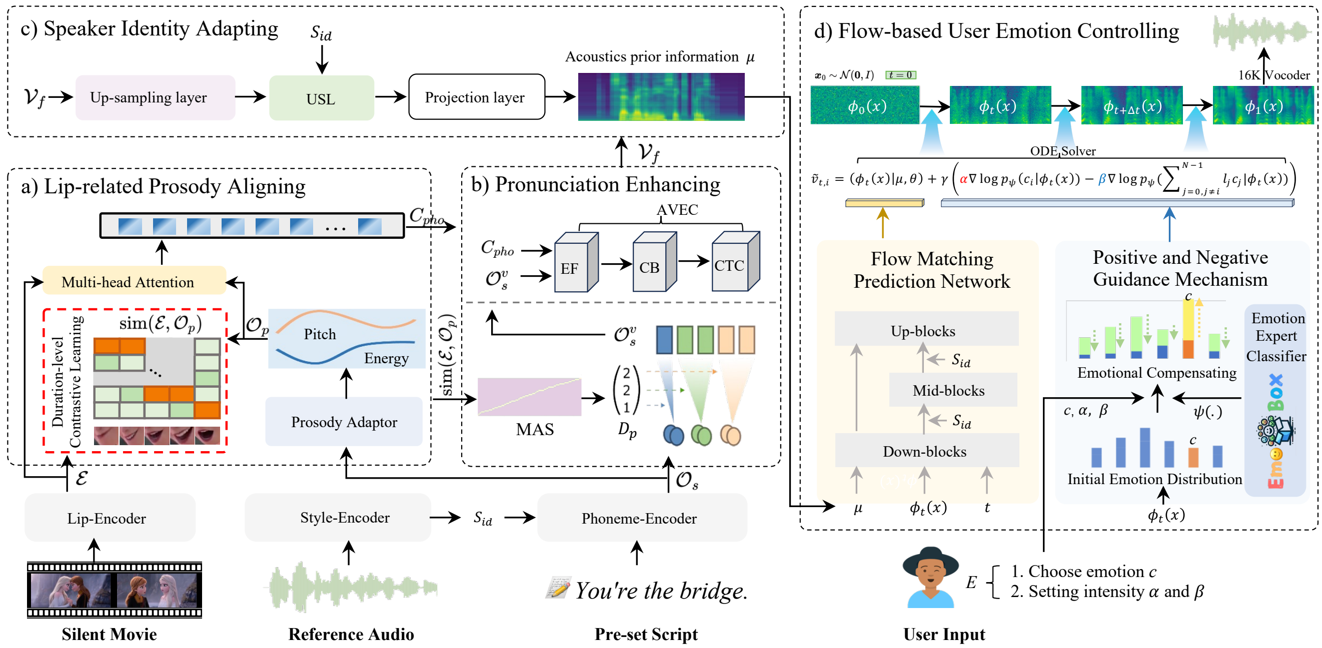


Figure 2. Architecture of the proposed EmoDubber, which consists of four main components: Lip-related Prosody Aligning (LPA) focuses on learning inherent consistency between lip motion and phoneme prosody by duration level contrastive learning; Pronunciation Enhancing (PE) fuses the output of LPA with expanding phoneme sequence by efficient conformer; Speaker Identity Adapting (SIA) aims to generate acoustics prior information μ while injecting speaker style; and Flow-based User Emotion Controlling (FUEC) renders user-specified emotion and intensity E in the flow-matching prediction process using positive and negative guidance.

CFM) for training, which yields an ODE-based decoder to improve the mel-spectrograms fidelity. However, these works are limited in the field of TTS and cannot be applied to V2C task. The Classifier Guidance [9] (CG) has been widely adopted for controlling specific attributes, *e.g.*, text-to-image and emotional TTS [13, 26, 45]. However, the existing emotional TTS using CG only enhances the needed emotion, which struggles to control complex speech containing mixed emotions. In this work, we introduce a flow-based user emotion controlling with positive and negative guidance mechanisms, which allows users to manipulate desired emotions and intensity more freely and promotes the development of artificial intelligence in movie dubbing.

3. Proposed Method

3.1. Overview

Given a silent video clip V_l, a reference audio R_a, a piece of text T_p, and user emotion guidance E, the goal of EmoDubber is to generate an audio clip $\hat{Y}$ that ensures precise lip-sync and clear pronunciation, while allowing users to control the intensity of emotion by adjusting E:

$$\hat{Y} = \mathrm{EmoDubber}(R_a, T_p, V_l, E), \tag{1}$$

specifically, $E = \{c, \alpha, \beta\}$, where c, α, and β are emotion label, positive weight, and negative weight, respectively. The main architecture of the proposed model is shown in Figure 2. First, the Lip-related Prosody Aligning (LPA) module absorbs the input information (*i.e.*, T_p, R_a, and V_l) to generate the lip-prosody context sequences with consistent duration cues, like basic pause and speech rate, which are guided by the proposed duration level contrastive learning. Next, the Pronunciation Enhancing (PE) strategy focuses on expanding phoneme sequences to video level by monotonic alignment search (MAS) and fusing it with the output of LPA by efficient conformer. Then, the Speaker Identity Adapting (SIA) module further decodes the output of the PE to acoustics prior information while introducing style from speaker identity. Finally, the Flow-based User Emotion Controlling (FUEC) renders emotions in the flow-matching vector field prediction process using Positive and Negative Guidance Mechanisms (PNGM) according to the user's instructions E to generate desired emotional speech. We detail each module below.

3.2. Lip-related Prosody Aligning

The proposed Lip-related Prosody Aligner (LPA) takes the script, silent video clips, and reference audio as input, and outputs the lip-prosody context sequences, which learn the pause and speed to align video by duration-level contrastive learning.

Extracting Phoneme-level Prosody and Lip-motion Embedding. Firstly, the open-source grapheme-to-phoneme (G2P) is used to obtain the textual phoneme sequence from raw scripts. Then, the phoneme encoder with affine transform [7] is used to extract style phoneme embeddings $\mathcal{O}_s$:

$$\mathcal{O}_s = \mathrm{PhoEncoder}(T_r \in \mathbb{R}^P, S_{id}), \tag{2}$$

where $\mathcal{O}_s \in \mathbb{R}^{P\times d_m}$, and the d_m and P represent the hidden dimension and length of the phoneme sequence, respectively. S_{id} is speaker style embedding, which is extracted by pre-trained speaker encoder from reference audio R_a, following [7, 50]. Next, we adopt prosody adaptor [6, 50] to generate phoneme-level prosody variations: pitch embedding P_{pho} and energy embedding E_{pho}. Thus, the phoneme-level prosody sequences $\mathcal{O}_p = \mathcal{O}_s \oplus P_{pho} \oplus E_{pho}$ is calculated by combining style phoneme embedding and prosody variations. To obtain the lip-motion embedding from the input silent video clip V_l, we adopt the same extracting pipeline as [6]:

$$\mathcal{E} = \mathrm{LipEncoder}(M_{roi} \in \mathbb{R}^{F\times D_w \times D_h \times D_c}), \tag{3}$$

where M_{roi} indicates the mouth Region of Interest (ROI) frame sequence cropped by face landmarks from V_l, following [6]. D_w, D_h, and D_c indicate the number of width, height, and channels of images in the mouth ROI frame. F denotes the total length of mouth ROI frame. $\mathcal{E} \in \mathbb{R}^{F\times d_m}$ denotes the output lip motion embedding.

Duration level Contrastive Learning. Inspired by [6, 7], we use multi-head attention to capture lip-prosody context sequences by serving lip motion embedding $\mathcal{E}$ as Query and prosody phonemes embedding $\mathcal{O}_p$ as Key and Value:

$$C_{pho} = \mathrm{softmax}(\frac{\mathcal{E}^\top \mathcal{O}_p}{\sqrt{d_m}}){\mathcal{O}_p}^\top, \tag{4}$$

where $C_{pho} \in \mathbb{R}^{F\times d_m}$ denotes the lip-prosody context sequences with the same length with video clips. Instead of non-constraint [6, 22] or simple diagonal-constraint [7, 16], our LPA module focuses on duration-level contrastive learning (DLCL) to achieve corrected alignment to guarantee monotonicity and subjectivity of weight matrix between prosody and lip-motion sequences:

$$\mathcal{L}_{cl} = -\log \frac{\sum \exp\left((\mathrm{sim}^+(\mathcal{E}, \mathcal{O}_p))/\tau\right)}{\sum \exp((\mathrm{sim}(\mathcal{E}, \mathcal{O}_p)))}, \tag{5}$$

where $\mathrm{sim}(\mathcal{E}, \mathcal{O}_p)$ indicates the attention weight matrix between $\mathcal{O}_p$ and $\mathcal{E}$. The positive pair $\mathrm{sim}^+(\mathcal{E}, \mathcal{O}_p)$ is calculated by multiplying $\mathrm{sim}(\mathcal{E}, \mathcal{O}_p)$ with correct duration-level correspondence $M^{gt}_{lip,pho}$:

$$\mathrm{sim}^+(\mathcal{E}, \mathcal{O}_p) = \mathrm{sim}(\mathcal{E}, \mathcal{O}_p) \times M^{gt}_{lip,pho}, \tag{6}$$

where $M^{gt}_{lip,pho}$ is a "0-1" matrix with P-th row and F-th column and satisfies the monotonicity and surjectivity. The value "1" represents correct correspondence between textual phoneme and lip motion through Montreal Forced Aligner (MFA) [33] model and coefficient between Frames per Second of the video (FPS) and sampling rate (SR). In this case, the DLCL encourages the positive pair to have a higher similarity to ensure the phoneme prosody unit focuses on the strongly related part in lip motion sequence.

3.3. Pronunciation Enhancing

The proposed Pronunciation Enhancing (PE) strategy aims to generate video-level phoneme enhancement sequences and fuse it with lip-prosody context sequences.

Explicit duration based Expanding. To obtain the duration of each phoneme unit directly from the learnable attention weight matrix $\mathrm{sim}(\mathcal{E}, \mathcal{O}_p)$, we use monotonic alignment search (MAS) [20] for explicit alignment. Specifically, MAS implements dynamic programming algorithms to find the optimal alignment path on matrix $\mathrm{sim}(\mathcal{E}, \mathcal{O}_p)$:

$$\mathcal{O}^v_s = \mathrm{LR}(D_p, \mathcal{O}_s), \tag{7}$$

specifically, $D_p = \mathrm{MAS}(\mathcal{E}, \mathcal{O}_p) \in \mathbb{R}^{P\times 1}$ denotes the explicit duration for each phoneme unit, which records the integer multiple alignment between phonemes and video frames. The $\mathrm{LR}(\cdot)$ is a length regulator, which aims to expand style phoneme sequences $\mathcal{O}_s \in \mathbb{R}^{P\times d_m}$ to video-level phoneme enhancement sequences $\mathcal{O}^v_s \in \mathbb{R}^{F\times d_m}$.

Efficient Conformer based Fusing. Inspired by [2], we use audio-visual efficient conformer (AVEC) to model both local and global dependencies using convolution and attention to reach better fusing performance on two kinds of feature: lip-prosody context sequences C_{pho} (from Eq. 4) and phoneme enhancement sequences $\mathcal{O}^v_s$ (from Eq. 7):

$$\mathcal{V}_f = \mathrm{Conformer}(C_{pho}, \mathcal{O}^v_s), \tag{8}$$

where $\mathcal{V}_f \in \mathbb{R}^{F\times d_m}$ indicates the fused intermediate feature. The $\mathrm{Conformer}(\cdot)$ represents AVEC, which consists of an early fusion (EF) strategy [30] to reduce model complexity, 5 Conformer blocks (CB) without downsampling, and connectionist temporal classification (CTC) [12, 21] layer to maximize the sum of probabilities of correct target phoneme to ensure pronunciation.

3.4. Speaker Identity Adapting

The Speaker Identity Adapting (SIA) aims to generate acoustics prior information μ in mel-spectrogram level from $\mathcal{V}_f$ (Eq. 8) and speaker style embedding S_{id}:

$$\mu = \mathrm{Proj}(\mathrm{USL}(\mathrm{Up}(\mathcal{V}_f), S_{id})), \tag{9}$$

where $\mu \in \mathbb{R}^{M_l\times d_a}$, and the M_l and d_a represent the length and hidden dimension of desired mel-spectrogram sequence. The SIA consists of up-sampling layer, utterance-level style learning (USL) module [7], and projection layer [7]. Firstly, we upsample the time axis of $\mathcal{V}_f$ to mel-spectrogram level by applying two layers of 2D convolutions [49]. Then, we use USL to inject the style information from style embedding S_{id}. Finally, the projection layer is used to project the output feature to target dimension.

3.5. Flow-based User Emotion Controlling

For given user's emotion instructions $E = \{c, \alpha, \beta\}$, the proposed Flow-based User Emotion Controlling (FUEC) focuses on determining the gradient direction and guidance scale based on E while iteratively converting noise into the mel-spectrogram to inject emotion expressiveness. It consists of Flow Matching Prediction Network (FMPN) and Positive and Negative Guidance Mechanisms (PNGM).

Flow Matching Prediction Network. Given the mel-spectrograms data space with data point M, where $M \sim q(M)$ and $q(M)$ is an unknown data distribution of mel-spectrograms, a possible approach to sample M from $q(M)$ is to give a probability density path defined as $p_t(x)$ where $t \in [0, 1]$, $p_0(x) = \mathcal{N}(x; \mathbf{0}, \boldsymbol{I})$ and $p_1(x) \approx q(x)$. Flow matching model can estimate the probability density path, gradually transforming noise $x_0 \sim p_0(x)$ into mel spectrogram $M \sim q(M)$. Here, we train Flow Matching Prediction Network (FMPN) based on optimal-transport conditional flow matching (OT-CFM) with a linear interpolation flow $\phi_t(x) = (1 - (1 - \sigma_{\min})t)x_0 + tM$, gradually transform noise x_0 to mel-spectrogram M from $t = 0$ to $t = 1$. Its gradient vector field is $u_t(\phi_t(x)|M) = M - (1 - \sigma_{min})x_0$, facilitating fast training and inference from noise to mel spectrogram due to its linear and time-invariant properties. The training objective is to train FMPN donated by θ to predict the gradient vector field of $\phi_t(x)$:

$$\mathcal{L}_\theta = \mathbb{E}_{t,q(M),p_t(x|\mu,M)}||v_t(\phi_t(x)|\mu, \theta) - u_t(\phi_t(x)|M)||^2, \tag{10}$$

where $v_t(\phi_t(x)|\mu, \theta)$ is the predicted gradient vector field of $\phi_t(x)$ according to the acoustics prior information μ. Then, we can solve the ODE $d\phi_t(x) = v_t(\phi_t(x)|\mu, \theta)dt$ from $t = 0$ to $t = 1$ to generate the target mel-spectrogram $\hat{M}$ from noise x_0. Instead of Matcha-TTS [34], we carefully design two kinds of style affine learning in FMPN (intra-blocks and inter-blocks) to adapt multi-speaker scenarios effectively.

Positive and Negative Guidance Mechanisms. Inspired by the emotions in human speech that are often blended rather than single [52], where multiple emotions can naturally overlap or co-exist, we propose Positive and Negative Guidance Mechanisms (PNGM) based on Classifier Guidance [9] to guide $v_t(\phi_t(x)|\mu, \theta)$ with specific emotion c. Unlike previous emotion synthesis methods [13, 52], PNGM allows users to perform more flexible manipulations in flow-matching prediction process by introducing positive and negative guidance to enhance the desired emotion and suppress others.

Suppose we have a well-trained FMPN θ predicting $v_t(\phi_t(x)|\mu, \theta)$ and an emotional expert classifier ψ which can predict the probability $p_\psi(c|\phi_t(x))$ that $\phi_t(x)$ belongs to emotion c. Note that our emotional expert classifier is pre-trained on multiple large-scale emotion datasets recorded in Emobox [31], enabling a better determination of real human emotions and improving the performance of FMPN-guided emotion. We can set emotion classes as $\{c_0, \cdots, c_{N-1}\}$ one-hot vector with N kinds of emotions on mel-spectrogram data M. The emotional softmax logit of $\phi_t(x)$ predicted by the emotional classifier ψ is $l_\psi(\phi_t(x)) = [l_0, \cdots, l_{N-1}]$, which can be seen as the emotional mix ratio of $\phi_t(x)$, and the mix emotion is $c_M = \sum_{i=0}^{N-1} l_i c_i$. To enhance emotion $c_i, i \in \{0, \cdots, N-1\}$, we can use positive guidance to guide M toward the direction of c_i and use negative guidance to suppress others, which can be formulated as:

$$\begin{aligned}\tilde{v}_{t,i} &= v_t(\phi_t(x)|\mu, \theta) \\ &+ \gamma\Big(\alpha\nabla \log p_\psi(c_i|\phi_t(x)) - \beta\nabla \log p_\psi(\sum_{j=0,j\neq i}^{j=N-1} l_j c_j|\phi_t(x))\Big),\end{aligned} \tag{11}$$

where γ controls the total degree of PNGM, α is the positive guidance scale controlling the emotion of c_i and β is the negative guidance scale controlling the degree weakening other emotions. This way, users can change α and β to control the emotion expressiveness of synthesized $\hat{M}$. Finally, the generated emotional mel-spectrograms $\hat{M}$ are converted to time-domain wave $\hat{Y}$ via the powerful vocoder.

4. Experimental Results

4.1. Implementation Details

Video frames are sampled at 25 FPS and all audios are resampled to 16kHz. The lip region is resized to 96 × 96 and pre-trained on ResNet-18, following [29, 32]. The window length, frame size, and hop length in STFT are 640, 1,024, and 160, respectively. The flow-prediction network is pre-trained on LibriSpeech [38] using 2 downsampling blocks, 2 midblocks, and 2 upsampling blocks. Each block had one Transformer layer with hidden dimensionality 256 and style affine layers [7], 2 heads, attention dimensionality 64, and snakebeta [25] activations. We use 8 heads for multi-head attention in LPA with 256 hidden sizes. The temperature coefficient τ of $\mathcal{L}_{cl}$ as 0.1. There are 5 conformer blocks in AVEC. We trained V2C, Chem, and GRID with batch sizes 64, 32, and 64, respectively. The Emotion Expert Classifier ψ is trained on 13 emotional datasets with more than 50,000 emotional audio recordings in Emobox [31], which collects large-scale Speech Emotion Recognition (SER) benchmarks. During the inference process, γ is set to 15, while $\alpha \in [0, 5]$ and $\beta \in [0, 2]$. Both training and inference are implemented with PyTorch on a GeForce RTX 4090 GPU. More details about the model and

Setting	Dubbing Setting 1.0						Dubbing Setting 2.0					
Methods	LSE-C ↑	LSE-D ↓	WER ↓	SECS ↑	MCD ↓	MCD-SL ↓	LSE-C ↑	LSE-D ↓	WER ↓	SECS ↑	MCD ↓	MCD-SL ↓
GT	8.12	6.59	03.85	100.00	0.0	0.0	8.12	6.59	03.85	100.00	0.0	0.0
Fastspeech2* [42]	3.34	11.60	15.33	79.26	5.88	6.73	3.34	11.60	15.33	79.26	5.88	6.73
StyleSpeech* [35]	2.06	12.27	79.14	63.00	7.64	9.87	2.12	12.14	80.01	60.06	8.31	10.28
Face-TTS [24]	1.98	12.50	62.24	59.56	7.51	12.59	1.96	12.53	68.13	53.44	7.64	12.79
V2C-Net [3]	1.97	12.17	90.47	51.52	6.25	8.31	1.82	12.09	94.59	44.19	6.74	9.04
HPMDubbing [6]	7.85	7.19	16.05	85.09	6.12	7.25	3.98	9.50	29.82	73.55	6.91	8.56
Speaker2Dub [50]	3.76	10.56	16.98	74.73	7.67	7.89	3.45	11.17	18.10	69.28	8.06	8.21
StyleDubber [7]	3.87	10.92	13.14	87.72	**5.41**	**5.73**	3.74	11.00	14.18	82.07	**6.01**	**6.36**
Ours	**8.11**	**6.92**	**11.72**	**90.62**	5.87	5.87	**8.09**	**6.96**	**12.81**	**85.06**	6.51	6.51

Table 1. Results on Chem benchmark. The method with "*" refers to a variant taking video embedding as an additional input following.

Methods	LSE-C ↑	LSE-D ↓	WER ↓	MOS-S ↑	MOS-N ↑
StyleDubber [6]	6.17	9.11	15.10	4.03±0.10	3.85±0.15
Speaker2Dub [6]	4.83	10.39	15.91	3.98±0.09	4.01±0.13
Ours	**7.40**	**6.65**	**14.03**	**4.07±0.09**	**4.05±0.06**

Table 2. The zero shot results under Dub 3.0 setting, which use unseen speaker as refernce audio.

#	Methods	LSE-C ↑	LSE-D ↓	WER ↓	SECS ↑	MCD ↓
1	w/o PE	**8.10**	**6.81**	53.36	84.92	7.19
2	w/o SIA	8.05	6.99	13.07	82.04	6.72
3	w/o LPA	4.47	10.45	36.31	85.03	7.41
4	Full model	8.09	6.96	**12.81**	**85.06**	**6.51**

Table 3. Ablation study of the proposed EmoDubber on the Chem benchmark dataset with 2.0 setting.

training are given in the supplementary material.

4.2. Datasets

Chem is a popular dubbing dataset recording a chemistry teacher speaking in the class [39]. It is collected from YouTube, with a total video length of approximately nine hours. For complete dubbing, each video has clip to sentence-level [16]. The number of train, validation, and test data are 6,082, 50, and 196, respectively.

GRID is a dubbing benchmark for multi-speaker dubbing [8]. The whole dataset has 33 speakers, each with 1,000 short English samples. All participants are recorded in studio with unified background. The number of train and test data are 32,670 and 3,280, respectively.

V2C-Animation is a multi-speaker dataset for animation movie dubbing with identity and emotion annotations [3]. It is collected from 26 Disney cartoon movies and covers 153 diverse characters. The whole dataset has 10,217 video clips with paired audio and subtitles. The training/validation/test size are 60%, 10%, 30%.

4.3. Evaluation Metrics

Audio-visual Sync Evaluation. To evaluate the synchronization between the generated speech and the video quantitatively, we adopt Lip Sync Error Distance (LSE-D) and Lip Sync Error Confidence (LSE-C) as our metrics, which are widely used to lip reading [49], talking face [19, 48], and video dubbing task [16, 28]. These metrics are based on the pre-trained SyncNet [5], which can explicitly test for lip synchronization in unconstrained videos in the wild [5, 40]. Compared to the length metric MCD-SL [3], we believe that LSE-C and LSE-D can more accurately measure the synchronization of vision and audio. The discussion of the two kinds of metrics is in Appendix E.

Speech Quality Evaluation. The Word Error Rate (WER) [36] is used to measure pronunciation accuracy by using Whisper-V3 [41] as the ASR model. To evaluate the timbre consistency between the generated dubbing and the reference audio, we employ the speaker encoder cosine similarity (SECS) following [7, 50] to compute the similarity of speaker identity. Besides, we adopt the Mel Cepstral Distortion Dynamic Time Warping (MCD) and speech length variant (MCD-SL) [1, 3, 37] to measure the difference between generated speech and real speech.

Emotional Evaluation. We use the Intensity Score, the average softmax logit of the target emotion, which ranges from 0 to 1, to measure the emotional intensity of generated audio. While previous works [13, 52] use average classification probability, they fail to distinguish varying intensities within the same emotion class. The average softmax logit provides a finer-grained measure, allowing for a more effective evaluation of emotion intensity differences in audio.

Subjective Evaluation. To further evaluate the quality of generated speech, we conduct a human study using a subjective evaluation metric, following the settings in [3]. Specifically, we adopt the MOS-naturalness (MOS-N) and MOS-similarity (MOS-S) to assess the naturalness of the generated speech and the recognization of the desired voice.

4.4. Comparison with SOTA Dubbing Methods

To compare with SOTA dubbing model without the function of emotion control, we remove the PNGM, *i.e.*, maintaining Figure 2 (a)-(c) and flow matching prediction network to generate waveform. We compare with the recent dubbing baselines to comprehensively analyze. More details about baselines and V2C results are in Appendix.

Results on the Chem Dataset. As shown in Table 1, our method achieves the best performance on almost all

Setting	Dubbing Setting 1.0						Dubbing Setting 2.0					
Methods	LSE-C ↑	LSE-D ↓	WER ↓	SECS ↑	MCD ↓	MCD-SL ↓	LSE-C ↑	LSE-D ↓	WER ↓	SECS ↑	MCD ↓	MCD-SL ↓
GT	7.134	6.786	22.41	100.00	0.00	0.00	7.134	6.786	22.41	100.00	0.00	0.00
Fastspeech2* [42]	5.01	9.79	19.61	11.35	7.24	7.95	5.01	9.79	19.61	11.35	7.24	7.95
StyleSpeech* [35]	5.90	9.24	22.62	90.04	5.74	5.88	4.79	10.28	19.82	59.58	7.01	7.82
Zero-shot TTS* [53]	5.03	10.02	20.05	85.93	5.75	6.40	4.48	10.54	21.05	81.34	6.27	7.29
Face-TTS [24]	4.69	10.14	44.37	82.97	7.44	8.16	4.55	10.27	39.05	34.14	7.77	8.59
V2C-Net [3]	5.59	9.52	47.82	80.98	6.79	7.23	5.34	9.76	49.09	71.51	7.29	7.86
HPMDubbing [6]	5.76	9.13	45.51	85.11	6.49	6.78	5.82	9.10	44.15	71.99	6.79	7.09
StyleDubber [7]	6.12	9.03	18.88	93.79	5.61	5.69	6.09	9.08	19.58	**86.67**	6.33	6.42
Speak2Dub [50]	5.27	9.84	**17.07**	**94.50**	5.34	5.45	5.19	9.93	**17.42**	85.76	6.17	6.43
Ours	**7.12**	**6.82**	18.53	92.22	**3.13**	**3.13**	**7.10**	**6.89**	19.75	86.02	**3.92**	**3.92**

Table 4. Results on GRID benchmark. The method with "*" refers to a variant taking video embedding as an additional input following.

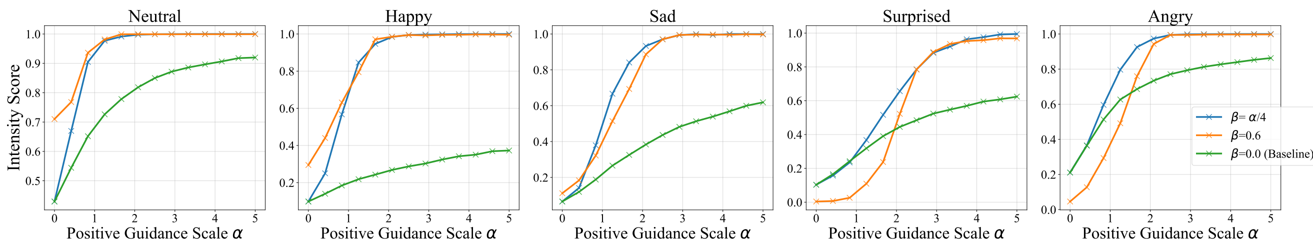


Figure 3. Intensity performance of EmoDubber on Chem. The horizontal axis shows the positive guidance α, and vertical axis displays the Intensity Score (IS), with different curves for various negative guidance β. Higher IS indicate stronger emotional intensity in audio.

metrics on Chem benchmark. Although StyleDubber can achieve good MCD and MCD-SL, they performed poorly on LSE-D and LSE-C, which reflects that they did not achieve true lip-sync. In contrast, our method achieves the best LSE-C and LSE-D, with absolute improvements of 4.24% and 4.0%, as well as the best WER with an improvement of 10.80%, demonstrating the effectiveness of proposed method to maintain high-quality audio-visual alignment and clear pronunciation simultaneously. Regarding speaker similarity (see SECS), the proposed method outperforms SOTA baseline StyleDubber with an absolute margin of 2.9%. Please note that setting 2 is more challenging than setting 1, which requires the model to have strong robustness. Despite challenging, EmoDubber still has a lead in lip-sync, pronunciation, and speaker identity.

Results on the GRID Dataset. We report the GRID result in Table 4. Our method is currently the only one that achieves the best performance in terms of both lip-sync (see LSE-C and LSE-D) and pronunciation clarity (see WER), whether in setting or setting 2. Specifically, our method significantly improves 19.46% LSE-C and 24.12% LSE-D on challenging setting 2, which indicates the effectiveness of the proposed approach in achieving accurate lip sync even in multi-speaker dubbing scenes. In addition, our method also achieves competitive results in WER, only slightly lower than the best pre-trained model Speak2Dub. But it turns out that our WER result (18.53%) exceeds the ground truth WER result (22.41%), which means that the intelligibility has reached the acceptable range for humans. Finally, our method achieves lowest MCD and MCD-SL compared to all baselines, which indicates our method achieves minimal acoustic difference in challenging setting 2.0.

Results on the Speaker Zero-shot test. This setting uses the audio of unseen characters (from another dataset) as reference audio to measure the generalizability of the dubbing model. Here, we use the audio from the Chem dataset as reference audio to measure the GRID dataset. Since there is no target audio at this setting, we only compare LSE-C/D and WER, and make subjective evaluations. As shown in Table 2, We were surprised to find that our method outperforms the SOTA dubbing methods (StyleDubber and Spk2Dub) on all metrics. In particular, our method surpasses the current best dubbing method Spk2Dub in WER, which reflects that our model is more robust in maintaining clear pronunciation in unseen speaker scenes. Furthermore, the proposed method still maintains the leading position in audio-visual synchronization (see LSE-C and LSE-D), which other SOTA dubbing methods cannot achieve.

Ablation Studies. The ablation results are presented in Table 3. It shows that all three modules contribute significantly to the overall performance, and each module has a different focus. After removing the PE, the WER severely drop. This reflects that the PE achieves better pronunciation by fusing video-level phoneme enhancement sequences with explicit duration. In contrast, the SECS is most affected by SIA, which indicates decoding mel-spectrograms by introducing global style is beneficial to identity recognition. Finally, the performance of LSE-C and LSE-D drops the most when removing LPA. It proves the effectiveness of modeling the relevance between lip motion and phoneme

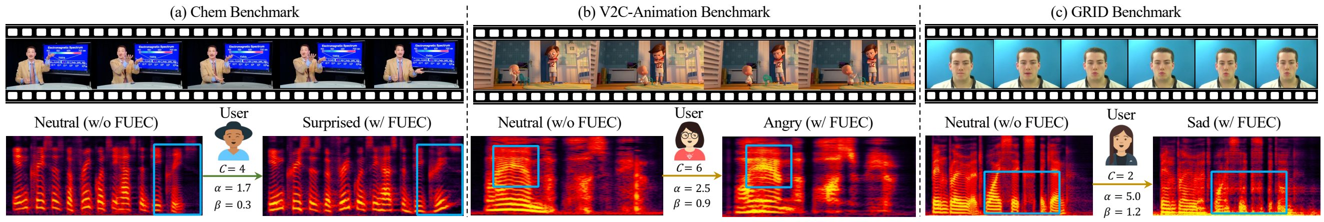


Figure 4. Visualization of audio samples generated by EmoDubber: one uses the proposed FUEC to guide emotions by users, and the other does not (Neutral). The green rectangles highlight key regions that have significant differences in emotional expressiveness.

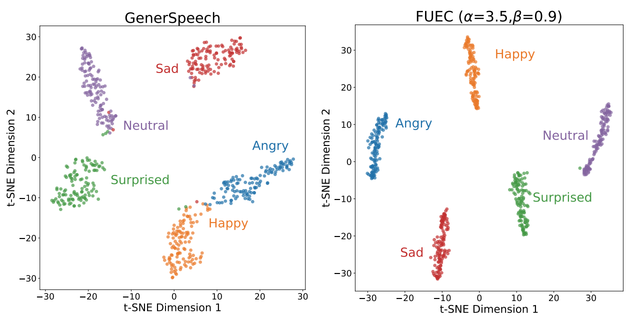


Figure 5. Visual results of emotional audio features by t-SNE, the TTS baseline is shown on the left and EmoDubber on the right.

#	Emotion	LSE-C ↑	LSE-D ↓	WER ↓	SECS ↑
0	Neutral	8.11	6.92	**11.32**	89.05
1	Happy	8.10	6.93	11.83	89.19
2	Sad	8.06	6.92	12.12	89.48
3	Fearful	8.09	6.90	11.91	89.00
4	Surprise	8.09	**6.89**	11.96	87.91
5	Disgusted	8.11	6.90	11.98	88.28
6	Angry	8.10	6.94	11.76	88.63
7	Original (w/o FUEC)	**8.11**	6.92	11.72	**90.62**

Table 5. Emotional speech quality study of EmoDubber.

prosody by duration-level contrastive learning, which is beneficial to reason correct audio-video aligning.

4.5. Emotional Controlling Evaluation

Intensity Controlling Results. Figure 3 shows the intensity-controlling results of EmoDubber on the Chem benchmark. The higher Intensity Score means a stronger emotional intensity of speech. We vary α from 0.0 to 5.0 and β from 0.0 to 2.0 and present results for two cases: one with a fixed β value and another where β varies with α to compare with the traditional baseline with no dual-scale guidance. Compared to the baseline, EmoDubber offers a wider range of intensity control, evident from our broader Intensity Score range from $\alpha = 0.0$ to $\alpha = 5.0$. Additionally, our method enables stronger emotional modulation, significantly achieving a higher Intensity Score than the baseline when $\alpha = 5.0$. This demonstrates that our EmoDubber supports a broader range of emotional control and enables the generation of audio with stronger emotional intensity by FUEC. We also note that the combination values of α and β can be more diverse, supporting a wider range of emotional intensity controls. More results about GRID and V2C are in Appendix B.

Emotion Zero-shot Conversion. To verify the emotion generalization on the emotionless dubbing dataset, we re-synthesis speech with five kinds of emotion on the Chem dataset (no emotion label). We use the publicly available GenerSpeech [17] as our baseline, which is a SOTA TTS model for emotional transfer in out-of-domain. As shown in Figure 5, we visualize the audio features using t-SNE [46]. Compared to the TTS baseline, EmoDubber demonstrates better emotion differentiation, highlighting its ability to generate audio with distinct emotions and generalize to other emotionless videos. Besides, it is worth noting that our model can free to adjust emotion intensity control in lip-sync dubbing, which is not possible with current emotional TTS baselines.

Emotion Speech Quality Results. To verify speech quality after emotion control, we test 7 kinds of emotional speech with each of 196 samples (*i.e.*, whole test set) by user turning on Chem dataset. Note that we set $\alpha = 3.5$ because this is sufficient to achieve strong emotion conversions, as shown in Figure 3. The results are shown in Table 5. We find that the emotional controlling did not have much effect on the audio-visual alignment, and the lip sync of the seven emotions was still well expressed (see LSE-D and LSE-C). In addition, the WER of the six emotions (except "sad") is lower than 12.0%, indicating the pronunciation clarity was not affected. As for "sad", it may be because the model does blur the pronunciation when imitating strong sadness. Finally, in terms of speaker similarity, the seven emotions are still very close to the original speech, which shows that emotion control does not affect the timbre of information.

Qualitative Results. To demonstrate the effect of EmoDubber's emotion synthesis, we invite three volunteers to express their expected emotions and intensities on three videos from dubbing datasets, respectively. We visualize the mel-spectrograms and provide comparison using FUEC or not. For instance, in (a), the user selects "surprise" (c=4) with an intensity of $\alpha = 3.1, \beta = 0.3$, which results in a noticeable rise trend at the end of the spectrograms (see blue box). In (b), the user selects "angry" (c=6) with an intensity of $\alpha = 4.5, \beta = 0.9$, which brings a remarkable increase in energy. It indicates that a strong tone is being expressed to render angry. Finally, in (c), the user selects "sad" (c=2) with an intensity of $\alpha = 5.0, \beta = 1.7$. This relatively high-intensity setting produces a spectrum with diminished high-frequency energy and softened transitions, typical of frustrated expressions associated with sadness. Appendix A and D provide more visualization results.

5. Conclusion

In this work, we propose EmoDubber, a controllable emotion dubbing architecture to help users specify the emotion they need while satisfying high-quality lip sync and clear pronunciation. The lip-related prosody aligning learns the inherent consistency by duration-level contrastive learning to reason the correct audio-visual alignment. Building on contextual lip motion and prosody information, the proposed pronunciation enhancing strategy fuses the video-level phoneme sequence to improve intelligibility. Besides, the Flow-based User Emotion Controlling (FUEC) with positive and negative guidance dynamically adjusts the flow-matching prediction process to control emotion intensity flexibly. Extensive experiments on three widely adopted benchmarks show favorable performance.

References

[1] Eric Battenberg, R. J. Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, and Tom Bagby. Location-relative attention mechanisms for robust long-form speech synthesis. In *ICASSP*, pages 6194–6198, 2020. 6

[2] Maxime Burchi and Radu Timofte. Audio-visual efficient conformer for robust speech recognition. In *WACV*, pages 2257–2266, 2023. 4

[3] Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu. V2C: visual voice cloning. In *CVPR*, pages 21210–21219, 2022. 1, 2, 6, 7

[4] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018. 2

[5] Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *ACCV Workshop*, pages 251–263, 2016. 6

[6] Gaoxiang Cong, Liang Li, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Wenyu Wang, Bin Jiang, Ming-Hsuan Yang, and Qingming Huang. Learning to dub movies via hierarchical prosody models. In *CVPR*, pages 14687–14697, 2023. 1, 2, 4, 6, 7

[7] Gaoxiang Cong, Yuankai Qi, Liang Li, Amin Beheshti, Zhedong Zhang, Anton van den Hengel, Ming-Hsuan Yang, Chenggang Yan, and Qingming Huang. Styledubber: Towards multi-scale style learning for movie dubbing. In *Findings of ACL*, pages 6767–6779, 2024. 2, 4, 5, 6, 7

[8] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006. 6

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3, 5

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 2

[11] Peng Gao, Le Zhuo, Ziyi Lin, Chris Liu, Junsong Chen, Ruoyi Du, Enze Xie, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv*, 2024. 2

[12] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pages 369–376, 2006. 4

[13] Yiwei Guo, Chenpeng Du, Xie Chen, and Kai Yu. Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3, 5, 6

[14] Yiwei Guo, Chenpeng Du, Ziyang Ma, Xie Chen, and Kai Yu. Voiceflow: Efficient text-to-speech with rectified flow matching. In *ICASSP*, pages 11121–11125, 2024. 2

[15] Michael Hassid, Michelle Tadmor Ramanovich, Brendan Shillingford, Miaosen Wang, Ye Jia, and Tal Remez. More than words: In-the-wild visually-driven prosody for text-to-speech. In *CVPR*, pages 10577–10587, 2022. 1, 2

[16] Chenxu Hu, Qiao Tian, Tingle Li, Yuping Wang, Yuxuan Wang, and Hang Zhao. Neural dubber: Dubbing for videos according to scripts. In *NeurIPS*, pages 16582–16595, 2021. 2, 4, 6

[17] Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech. In *NeurIPS*, 2022. 8

[18] Gonzalo Iturregui-Gallardo and Anna Matamala. Audio subtitling: Dubbing and voice-over effects and their impact on user experience. *Perspectives*, 29(1):64–83, 2021. 2

[19] Youngjoon Jang, Ji-Hoon Kim, Junseok Ahn, Doyeop Kwak, Hongsun Yang, Yooncheol Ju, Ilhwan Kim, Byeong-Yeol Kim, and Joon Son Chung. Faces that speak: Jointly synthesising talking face and speech from text. In *CVPR*, pages 8818–8828, 2024. 6

[20] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In *NeurIPS*, 2020. 4

[21] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip-to-speech synthesis in the wild with multi-task learning. In *ICASSP*, pages 1–5, 2023. 4

[22] Minsu Kim, Chae Won Kim, and Yong Man Ro. Deep visual forced alignment: Learning to align transcription with talking face video. In *AAAI*, pages 8273–8281, 2023. 4

[23] Sungwon Kim, Kevin J. Shih, Rohan Badlani, João Felipe Santos, Evelina Bakhturina, Mikyas Desta, Rafael Valle, Sungroh Yoon, and Bryan Catanzaro. P-flow: A fast and data-efficient zero-shot TTS through speech prompting. In *NeurIPS*, 2023. 2

[24] Jiyoung Lee, Joon Son Chung, and Soo-Whan Chung. Imaginary voice: Face-styled diffusion model for text-to-speech. In *ICASSP*, pages 1–5, 2023. 2, 6, 7

[25] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *ICLR*, 2023. 5

[26] Xiang Li, Zhi-Qi Cheng, Jun-Yan He, Xiaojiang Peng, and Alexander G Hauptmann. Mm-tts: A unified framework for multimodal, prompt-induced emotional text-to-speech synthesis. *arXiv preprint arXiv:2404.18398*, 2024. 3

[27] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2

[28] Junchen Lu, Berrak Sisman, Rui Liu, Mingyang Zhang, and Haizhou Li. Visualtts: TTS with accurate lip-speech synchronization for automatic voice over. In *ICASSP*, pages 8032–8036, 2022. 2, 6

[29] Pingchuan Ma, Brais Martinez, Stavros Petridis, and Maja Pantic. Towards practical lipreading with distilled and efficient models. In *ICASSP*, pages 7608–7612, 2021. 5

[30] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP*, pages 7613–7617, 2021. 4

[31] Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, Jiaxin Ye, Xie Chen, and Thomas Hain. Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. *arXiv preprint arXiv:2406.07162*, 2024. 5

[32] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Lipreading using temporal convolutional networks. In *ICASSP*, pages 6319–6323, 2020. 5

[33] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, pages 498–502, 2017. 4

[34] Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP*, pages 11341–11345, 2024. 2, 5

[35] Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. Meta-stylespeech : Multi-speaker adaptive text-to-speech generation. In *ICML*, pages 7748–7759, 2021. 6, 7

[36] Andrew Cameron Morris, Viktoria Maier, and Phil D. Green. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech*, pages 2765–2768, 2004. 6

[37] Meinard Müller. Dynamic time warping. *Information Retrieval for Music and Motion*, pages 69–84, 2007. 6

[38] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP*, pages 5206–5210, 2015. 5

[39] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *CVPR*, pages 13793–13802, 2020. 6

[40] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM MM*, pages 484–492, 2020. 6

[41] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, pages 28492–28518, 2023. 6

[42] 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. In *ICLR*, 2021. 6, 7

[43] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34: 1415–1428, 2021. 2

[44] Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3d molecule generation. In *NeurIPS*, 2023. 2

[45] Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. Emomix: Emotion mixing via diffusion models for emotional speech synthesis. *arXiv preprint arXiv:2306.00648*, 2023. 3

[46] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[47] Li Wan, Quan Wang, Alan Papir, and Ignacio López-Moreno. Generalized end-to-end loss for speaker verification. In *ICASSP*, pages 4879–4883, 2018. 1

[48] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T. Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *CVPR*, pages 14653–14662, 2023. 6

[49] Yochai Yemini, Aviv Shamsian, Lior Bracha, Sharon Gannot, and Ethan Fetaya. Lipvoicer: Generating speech from silent videos guided by lip reading. In *ICLR*, 2024. 5, 6

[50] Zhedong Zhang, Liang Li, Gaoxiang Cong, YIN Haibing, Yuhan Gao, Chenggang Yan, Anton van den Hengel, and Yuankai Qi. From speaker to dubber: Movie dubbing with prosody and duration consistency learning. In *ACM MM*, 2024. 1, 2, 4, 6, 7

[51] Yuan Zhao, Zhenqi Jia, Rui Liu, De Hu, Feilong Bao, and Guanglai Gao. Mcdubber: Multimodal context-aware expressive video dubbing. *arXiv preprint arXiv:2408.11593*, 2024. 2

[52] Kun Zhou, Berrak Sisman, Rajib Rana, Björn W. Schuller, and Haizhou Li. Speech synthesis with mixed emotions. *IEEE Trans. Affect. Comput.*, 14(4):3120–3134, 2023. 5, 6

[53] Yixuan Zhou, Changhe Song, Xiang Li, Luwen Zhang, Zhiyong Wu, Yanyao Bian, Dan Su, and Helen Meng. Content-dependent fine-grained speaker embedding for zero-shot speaker adaptation in text-to-speech synthesis. In *Interspeech*, pages 2573–2577, 2022. 7