Learning to Dub Movies via Hierarchical Prosody Models

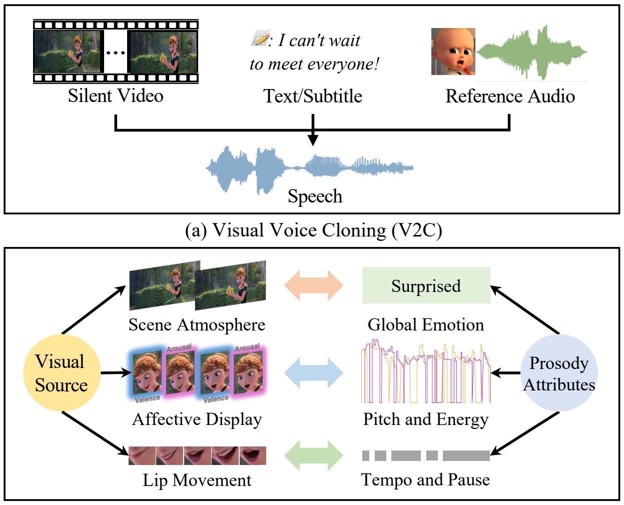
Gaoxiang Cong[1] Liang Li[2,3†] Yuankai Qi[4] Zheng-Jun Zha[5] Qi Wu[4] Wenyu Wang[1]
Bin Jiang[1] Ming-Hsuan Yang[6,7] Qingming Huang[2]
[1]Shandong University [2]Institute of Computing Technology, Chinese Academy of Sciences
[3]Lishui Institute of Hangzhou Dianzi University
[4]Australian Institute for Machine Learning, University of Adelaide
[5]University of Science and Technology of China [6]University of California [7]Yonsei University

Abstract

Given a piece of text, a video clip and a reference audio, the movie dubbing (also known as visual voice clone, V2C) task aims to generate speeches that match the speaker's emotion presented in the video using the desired speaker voice as reference. V2C is more challenging than conventional text-to-speech tasks as it additionally requires the generated speech to exactly match the varying emotions and speaking speed presented in the video. Unlike previous works, we propose a novel movie dubbing architecture to tackle these problems via hierarchical prosody modeling, which bridges the visual information to corresponding speech prosody from three aspects: lip, face, and scene. Specifically, we align lip movement to the speech duration, and convey facial expression to speech energy and pitch via attention mechanism based on valence and arousal representations inspired by the psychology findings. Moreover, we design an emotion booster to capture the atmosphere from global video scenes. All these embeddings are used together to generate mel-spectrogram, which is then converted into speech waves by an existing vocoder. Extensive experimental results on the V2C and Chem benchmark datasets demonstrate the favourable performance of the proposed method. The code and trained models will be made available at https://github.com/GalaxyCong/HPMDubbing.

1. Introduction

Movie dubbing, also known as visual voice clone (V2C) [9], aims to convert a paragraph of text to a speech with both desired voice specified by reference audio and desired emotion and speed presented in the reference video as shown in the top panel of Figure 1. V2C is more challenging than other speech synthesis tasks in two aspects: first, it requires synchronization between lip motion and generated speech; second, it requires proper prosodic variations of the generated speech to reflect the speaker's emotion in the video (*i.e.*, the movie's plot). These pose significant challenges to existing voice cloning methods.



(a) Visual Voice Cloning (V2C)

(b) Hierarchical Prosody Modeling

Figure 1. (a) Illustration of the V2C tasks. (b) To generate natural speech with proper emotions, we align the phonemes with lip motion, estimate pitch and energy based on facial expression's arousal and valence, and predict global emotion from video scenes.

Although significant progress has been made, existing methods do not handle the challenges in V2C well. Specifically, text-based dubbing methods [46–48, 54] construct speeches from given text conditioned on the different speaker embedding but do not consider audio-visual synchronization. On the other hand, lip-referred dubbing schemes [18,32,55] predict mel-spectrograms directly from a sequence of lip movements typically by encoder-decoder models. Due to high error rates in generated words, these methods can hardly guarantee high-quality results. Further-

[†] Corresponding author.

more, video-and-text based dubbing methods [17, 20, 32] focus on inferring speaker characters (*e.g.*, age and gender). However, these visual references usually do not convey targeted emotion well as intended in V2C.

An ideal dub should align well with the target character so that the audiences feel it is the character speaking instead of the dubber [7]. Thus, a professional dubber usually has a keen sense of observing the unique characteristics of the subject and acts on voice accordingly. In this work, we address these issues with a hierarchical dubbing architecture to synthesize speech. Unlike previous methods, our model connects video representations to speech counterparts at three levels: lip, face, and scene, as shown in Figure 1.

In this paper, we propose a hierarchical prosody modeling for movie dubbing, which could keep the audio-visual sync and synthesis speech with proper prosody following the movie's plot. Specifically, we first design a duration alignment module that controls speech speed by learning temporal correspondence via multi-head attention over phonemes and lip motion. Second, we propose an affective-display based Prosody Adaptor (PA), which learns affective psychology computing conditioned on facial expression and is supervised by corresponding energy and pitch in the target voice. In particular, we introduce arousal and valence features extracted from facial regions as emotion representations. This is inspired by the affective computing method [51], which analyses the facial affect relying on dimensional measures, namely valence (how positive the emotional display is) and arousal (how calming or exciting the expression looks). Third, we exploit a scene-atmosphere based emotion booster, which fuses the global video representation with the above adapted hidden sequence and is supervised by the emotive state of the whole voice. The outputs of these three modules are fed into a transformer-based decoder, which converts the speech-related representations into mel-spectrogram. Finally, we output the target speech waves from the mel-spectrogram via a powerful vocoder.

The contributions of this paper are summarized below:

- We propose a novel hierarchical movie dubbing architecture to better synthesize speech with proper prosody by associating them with visual counterparts: lips, facial expressions, and surrounding scenes.
- We design an affective display-based prosody adaptor to predict the energy and pitch of speech from the arousal and valence fluctuations of facial regions in videos, which provides a fine-grained alignment with speakers' emotions.
- Extensive experimental results demonstrate the proposed method performs well against state-of-the-art models on two benchmark datasets.

2. Related Work

Text to Speech Synthesis. Over the recent years, numerous TTS models [2, 29, 40, 41, 47, 48, 54] have been proposed for generating high-quality natural speech conditioned on given text. Tacotron [54] is an end-to-end generative TTS model that synthesizes speech directly from characters. Then, Tacotron2 [29] replaces the RNN structures by introducing the attention mechanism to improve training efficiency and solve the long dependency issue. Furthermore, FastSpeech [47] and Fastspeech2 [46] exploit the Feed-Forward Transformer (FFT) to generate mel-spectrogram from phoneme sequences. Despite the impressive voice generated, these methods cannot provide the audio with desired emotion and audio-visual sync for movie dubbing.

Lip to Speech Synthesis. This task aims to reconstruct speech based on the lip motions alone [3,25]. Lip2Wav [42] is a sequence-to-sequence architecture focusing on learning mappings between lip and speech for individual speakers. Recently, [15,18,49,55] improve the architecture and training methods, and provide the possibility of unconstrained speech synthesis in the wild. However, lip-to-speech is incompetent for movie dubbing because the word error rate is still high [1, 3, 11, 13, 16]. In this work, we focus on reconstructing accurate speech from lip motions and generating the desired emotion and identity with proper prosody.

Talking Heads. Numerous methods have been developed for audio-visual translation [58] or speaking style transfer [57] by reconstructing the visual content in video [8, 30, 31, 50, 53, 59, 61–63]. Wav2Lip [43] uses an expert lip-syncs discriminator to morph lip movements of arbitrary identities. Recently, Papantoniou *et al.* [38] develop a Neural Emotion Director (NED) to manipulate emotions while preserving speech-related lip movements. However, these methods cannot adapt to the movie dubbing task because they emphasize using generative models to readjust the facial regions instead of reconstructing the desired speech.

Visual Voice Cloning. Movie dubbing, also known as visual voice clone, aims to convert scripts to speech with both desired voice identity and emotion based on the reference audio and video. To control the speed of generated speech, Neural Dubber [20] exploits a text-video aligner by using scaled dot-product attention mechanism. VTTS [17] uses multi-source attention to fuse the triplets feature and outputs the mel-spectrogram via an RNN-based decoder. Since explicit emotion categories [28] do not exist in these methods, Chen *et al.* [9] develops a V2C model on a more challenging Densiny Animation dataset, which concentrates on emotional dubbing for movie characters. Although the V2C considers emotion labels, the adopted global video representation negatively affects the fine-grained emotional expression and makes it challenging to render correct prosody corresponding to plot developments. To solve this issue, we

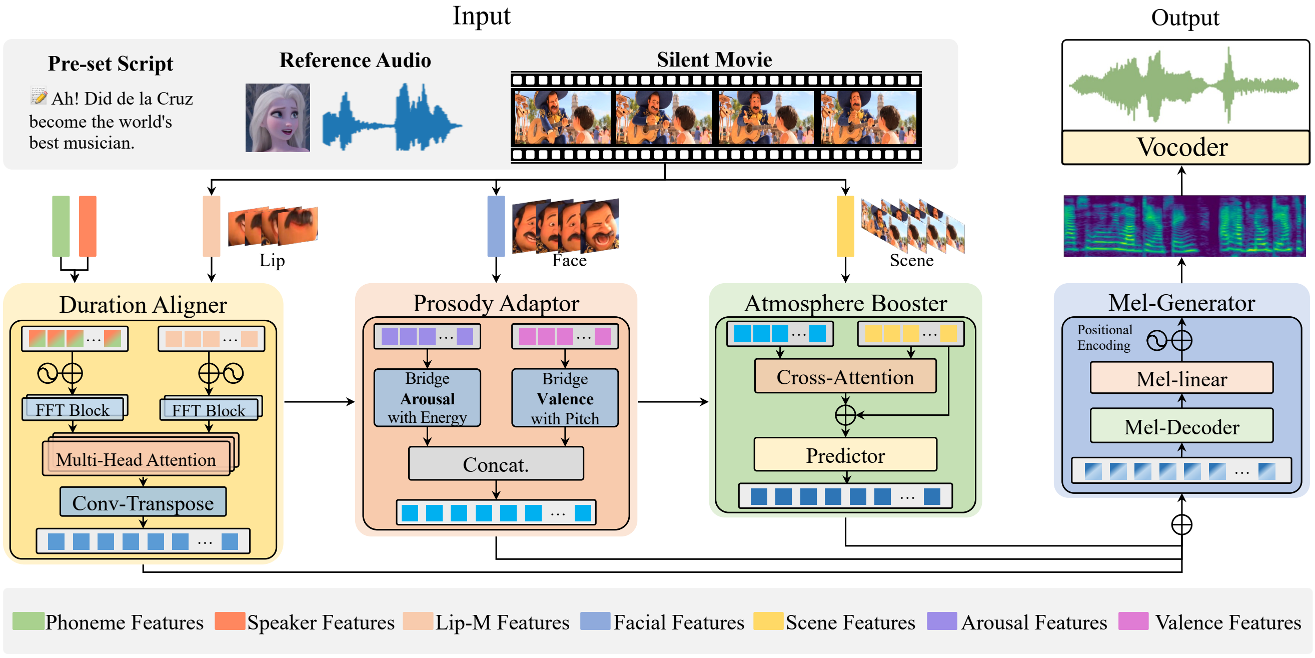


Figure 2. Architecture of the proposed hierarchical modular network for movie dubbing, which consists of four main components: Duration Aligner (Sec. 3.1), which learns to predict speech duration based on aligning lip movement and text phoneme; Prosody Adaptor (Sec. 3.2), which predicts energy and pitch from facial arousal and valence, respectively; Atmosphere Booster (Sec. 3.3), which learns a global emotion embedding from a video scene level; and Mel-Generator (Sec. 3.4), which generates mel-spectrograms from embeddings obtained by the aforementioned three modules. The mel-spectrograms are finally converted to audio by a widely adopted vocoder.

propose a hierarchical movie dubbing architecture to better synthesize speech with proper prosody and emotion.

3. Method

The main architecture of the proposed model is shown in Fig. 2. First, we use a phoneme encoder [9] to convert the input text Z_{text} to a series of phoneme embeddings $\mathcal{O} = \{\mathbf{o}_1, ..., \mathbf{o}_L\}$ and use a speaker encoder F_{spk} [9] to capture the voice characteristics $\mathcal{U}$ from different speakers. Then, taking phonemes and lip regions as input, the duration aligner module uses a multi-head attention mechanism to learn to associate phonemes with related lip movements. Next, the affective display-based prosody adaptor module learns to predict the energy and pitch of the desired speech based on arousal and valence features extracted from facial expressions, respectively. And then, the scene atmosphere booster encodes a global representation of emotion of the entire video content. All the outputs of the above three modules are combined to generate mel-spectrograms, which are finally transformed to a waveform Y_{voice} using a adopted vocoder. We detail each module below.

3.1. Duration Aligner

The duration aligner contains three steps: (1) extracting the lip features from movie; (2) aligning the phonemes of text with the lips; (3) expanding the fused phoneme-lip representation to the desired mel-spectrogram length.

Extracting lip feature. Let D_w, D_h and D_c be the width, height and number of channels of the video frames, respectively. We first extract lip regions $\mathbf{x_m} \in \mathbb{R}^{T_v \times D_w \times D_h \times D_c}$ from the given video using mouth region pre-processing from [20, 33–36]. Then we exploit the LipEncoder to obtain the lip movement representation:

$$\mathbf{E_{lip}} = \text{LipEncoder}(\mathbf{x_m}) \in \mathbb{R}^{T_v \times D_m}, \tag{1}$$

where T_v denotes the number of video frames, and D_m is the hidden dimension of the dynamic lip feature. The LipEncoder consists of several feed-forward transformer blocks that are suitable for capturing both long-term and short-term dynamics lip movement features.

Aligning text with lips. Inspired by the success of attention mechanism for cross-modality alignment [10, 14, 19, 26, 27, 44, 45, 64], we adopt multi-head attention to learn the alignment between the text phoneme and the lip movement sequence. We use lip embedding as a query to compute the attention on text phonemes. The larger the attention, the more related between a lip embedding and a text phoneme. Due to variations of mouth shapes and pronunciations, the multi-head attention mechanism is suitable for learning their alignments from different aspects. The text-video context sequence $\mathbf{E_{lip,txt}} = [\alpha^1_{lip,txt}, ..., \alpha^n_{lip,txt}] \in$

$\mathbb{R}^{T_v \times D_m}$ is a concatenation of outputs of n attention heads. Concretely, the k-th head's output $\alpha^k_{lip,txt}$ is obtained by:

$$\begin{aligned} \alpha^k_{lip,txt} &= \mathrm{softmax}(\frac{\mathrm{Q}^\top \mathrm{K}}{\sqrt{d_k}} + \mathrm{M_t})\mathrm{V}^\top, \\ \mathrm{Q} = \mathrm{W}_j^Q {E_{lip}}^\top, \mathrm{K} &= \mathrm{W}_j^K \mathcal{O}^\top, \mathrm{V} = \mathrm{W}_j^V \mathcal{O}^\top, \end{aligned} \tag{2}$$

where $\mathbf{W}_j^*$ is a learnable parameter matrix, d_k is the embeeding dimension of $\alpha^k_{lip,txt}$, and $\mathbf{M_t}$ is a mask matrix indicating whether a token can be attended. The aligned representation $E_{lip,txt}$ is later expanded to the length of the desired mel-spectrogram.

Expanding to the desired length. According to the findings in [20], in an audio-visual clip, the length of a mel-spectrograms sequence is n times that of a video frame sequence because they are temporally synchronized in audio and visual modalities. The number n is computed as:

$$n = \frac{T_{mel}}{T_v} = \frac{sr/hs}{FPS} \in \mathbb{N}^+, \tag{3}$$

where FPS denotes the Frames per Second of the video, sr denotes the sampling rate of the audio, and hs denotes hop size when transforming the raw waveform into mel-spectrograms. Phoneme-lip feature $E_{lip,txt}$ is simply duplicated n times in [20] as the final phoneme-lip representation, which lacks flexibility. Instead, we propose to use transposed convolutions to learn the expansion of $E_{lip,txt}$, which can be formulated as:

$$\mathbf{M}_{pho,lip} = \text{Conv-Transpose}(n, E_{lip,txt}) \in \mathbb{R}^{T_y \times D_m}, \tag{4}$$

where T_y denotes the length of the desired mel-spectrogram, D_m is the dimension of the initial mel-spectrogram. The parameters (stride and kernel size) of the transposed convolution can be set under the guidance of n so that $T_y \approx n \times T_v$.

3.2. Affective-display based Prosody Adaptor

One of the critical issues in the V2C task is to describe the speaker's emotions in the given video. To solve this problem, we design an affective-display based Prosody Adaptor (PA), which uses the arousal and valence extracted from facial expressions to represent the emotion. The arousal and valence are then used to predict the energy and pitch of the desired speech model.

Valence and Arousal Feature. To accurately capture the valence and arousal information from facial expressions, we utilize an emotion face-alignment network (EmoFAN) [51] to encode the facial region into valence $\mathbf{V}$ and arousal $\mathbf{A}$. $\mathbf{V}, \mathbf{A} = \mathrm{EmoFAN}(x_f) \in \mathbb{R}^{T_v \times D_m}$, $x_f \in \mathbb{R}^{T_v \times D_w \times D_h \times D_c}$ is the face region extracted via S^3FD face detection [60]. The EmoFAN focuses on facial regions relevant to emotion estimation, which utilizes a face alignment network (FAN) for facial point detection to ensure robustness by jointly predicting categorical and continuous emotions.

Bridging Arousal with Energy. Arousal is the physiological and psychological state of being awoken or of sense organs stimulated to the point of perception. To bridge the vocal energy with arousal display, we compute an arousal context vector A_i^l for frame-level energy of the desired speech from phoneme-lip representation $\mathbf{M}_{pho,lip}$:

$$\begin{aligned} A_i^l &= \sum_{k=0}^{T_y-1} \xi_{i,k} M^k_{pho,lip}, \\ \xi_{i,k} &= \exp(\hat{\xi}_{i,k}) / \sum_{j=0}^{T_y-1} \exp(\hat{\xi}_{i,j}), \\ \hat{\xi}_{i,k} &= \mathbf{w_a}^\top \tanh(\mathbf{W_a}^\top A_i + \mathbf{U_a}^\top M^k_{pho,lip} + \mathbf{b_a}) \end{aligned} \tag{5}$$

where i is frame index, A_i is the i-th row of $\mathbf{A}$, $\xi_{i,k}$ is the attention weight on the k-th phoneme-lip feature $M^k_{pho,lip}$ (the k-th row of $\mathbf{M}_{pho,lip}$) regarding i-th arousal display; $\mathbf{w_a^T}$, $\mathbf{W_a^T}$, $\mathbf{U_a^T}$ and $\mathbf{b_a}$ are learnable parameters; T_y is the length of the desired mel-spectrogram.

Then, we use an energy predictor to project the arousal-related phoneme-lip embedding A_i^l to the energy of speech:

$$E_{aro} = \mathrm{Predictor}(\{A_i^l\}_{i=1}^{T_y}), \tag{6}$$

where $E_{aro} \in \mathbb{R}^{T_y}$ represents the predicted energy of speech, and the energy predictor consists of several fully-connected layers, Conv1D blocks and layer normalization.

Bridging Valence with Pitch. In prosody linguistics, speakers can speak with a wide pitch range (this is usually associated with excitement) while, at other times, with a narrow range. Valence is the affective quality referring to the intrinsic positiveness or negativeness of an event or situation. Similar to arousal to energy, we first compute a valence context vector for the frame-level pitch of the desired speech from phoneme-lip representation $\mathbf{M}_{pho,lip}$:

$$\begin{aligned} V_i^l &= \sum_{k=0}^{T_y-1} \psi_{i,k} M^k_{pho,lip}, \\ \psi_{i,k} &= \exp(\hat{\psi}_{i,k}) / \sum_{j=0}^{T_y-1} \exp(\hat{\psi}_{i,j}), \\ \hat{\psi}_{i,k} &= \mathbf{w_g}^\top \tanh(\mathbf{W_g}^\top V_i + \mathbf{U_g}^\top M^k_{pho,lip} + \mathbf{b_g}) \end{aligned} \tag{7}$$

where $\mathbf{w_g^T}$, $\mathbf{W_g^T}$, $\mathbf{U_g^T}$ and $\mathbf{b_g}$ are learnable parameters. Then, we exploit a pitch predictor to convert the valence-related phoneme-lip embedding V_i^l to the pitch of speech, which can be formulated as:

$$P_{val} = \mathrm{Predictor}(\{V_i^l\}_{i=1}^{T_y}). \tag{8}$$

Finally, we concatenate the two prosody-related features as the contextualized affect primitives:

$$\mathbf{M_p} = \{[A_i^l; V_i^l]\}_{i=1}^{T_y}. \tag{9}$$

3.3. Scene Atmosphere Booster

As a unique form of artistic expression, the scene layout and colours of the film convey an emotional atmosphere to evoke resonance with the audience [5, 56]. To reason the comprehensive emotion, we design a scene atmosphere booster to combine the global context information and generated prosody. First, we use the I3D model [6] to extract the scene representation $\mathcal{S}$ from the video. Then, we fuse the global contextual emotion of vision with prosody information of speech through a cross-model attention mechanism by:

$$\mathbf{M_e} = \mathrm{Softmax}(\frac{\mathbf{M_p} S^\top}{\sqrt{D_m}})\mathbf{M_p} \in \mathbb{R}^{T_y \times D_w}. \tag{10}$$

Finally, formulated as a maxpool and fully-connected layer, the emotional predictor projects prosody-scene context sequence $\mathbf{M_e}$ to emotional embedding:

$$G_{emo} = \mathrm{Maxpool}(\mathrm{Predictor}(\mathbf{M_e})). \tag{11}$$

3.4. Audio Generation

We fuse the three kinds of speech attribute information by concatenation operation in the spatial dimension. Then, we use transformer-based mel-spectrogram decoder to convert the adapted hidden sequence into mel-spectrogram sequence in parallel by:

$$\mathbf{f} = \mathrm{TransformerDecoder}(\mathbf{M}_{pho,lip} \oplus \mathbf{M_p} \oplus \mathbf{M_e}), \tag{12}$$

where $M_{pho,lip}$, M_p, and M_e denote the hidden representation of phoneme-lip feature, prosody variances, and emotional tone, respectively. Specifically, our mel-spectrogram decoder consists of a stack of self-attention layers [52] and 1D-convolution layers as in FastSpeech [47]. We then use the mel-linear layer and postnet [48] to refine the hidden states into final dimensional mel-spectrograms by:

$$y = \mathrm{PostNet}(\mathrm{FC}(f)). \tag{13}$$

Finally, to generate the time-domain waveform y_w from mel-spectrogram y, we use HiFi-GAN [23] as our vocoder, which mainly consists of a transposed convolution network and a multi-receptive field fusion module.

3.5. Loss Functions

Our model is trained in an end-to-end fashion via optimizing the sum of all losses:

$$\mathcal{L}_{pitch} = \frac{1}{T_y} \sum_{t=1}^{T_y-1} (P_t - \hat{P}_{val}^t)^2, \tag{14}$$

$$\mathcal{L}_{energy} = \frac{1}{T_y} \sum_{t=1}^{T_y-1} (E_t - \hat{E}_{aro}^t)^2, \tag{15}$$

$$\mathcal{L}_{emo} = -\sum_{i=1}^{C} G_i \mathrm{log}(\hat{G}_{emo}^i), \tag{16}$$

$$\mathcal{L}_{mel} = \frac{1}{T_y} \sum_{t=1}^{T_y-1} \left\| \mathbf{f}_t - \hat{\mathbf{f}}_t \right\|. \tag{17}$$

$$\mathcal{L}_S = \lambda_1 \mathcal{L}_{mel} + \lambda_2 \mathcal{L}_{pitch} + \lambda_3 \mathcal{L}_{energy} + \lambda_4 \mathcal{L}_{emo}, \tag{18}$$

where $\mathcal{L}_{mel}$, $\mathcal{L}_{pitch}$, $\mathcal{L}_{energy}$ and $\mathcal{L}_{emo}$, denote the losses of mel-spectrogram, pitch, energy and global emotion respectively. P_t, E_t, and $\mathbf{f}_t$ are ground-truth pitch, energy, and mel-spectrogram on frame-level, respectively. G is the ground-truth emotional label and C denotes all categories.

4. Experimental Results

In this section, we first briefly describe the datasets used for evaluation and the evaluation metric. Then we present the implementation details of the proposed method. Last, we show the results compared to state-of-the-art methods and the ablation study.

4.1. Datasets

V2C is a multi-speaker dataset for animation movie dubbing with identity and emotion annotations [9]. It is collected from 26 Disney cartoon movies and covers 153 diverse characters. V2C not only needs to generate voices with identity characteristics according to the reference audio but also capture emotional information based on the reference movie clips. The whole dataset has 10,217 video clips with paired audio and subtitles. The training/validation/test size are 60%, 10%, 30%.

Chem is a single-speaker dataset composed of 6,640 short video clips collected from the YouTube, with the total video length of approximately nine hours [20]. The Chem dataset is originally used for the unconstrained single-speaker lip-to-speech synthesis [42], which takes place in a chemistry lecture. For fluency and complete dubbing, each video clip has sentence-level text and audio based on the start and end timestamps. There are 6,240, 200, and 200 dubbing clips for training, validation, and testing, respectively.

4.2. Evaluation Metrics

Audio-visual synchronization. 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 can explicitly test for synchronization between lip motions and speech in unconstrained videos in the wild [12, 43].

Mel Cepstral Distortion and its variants. MCD [24], MCD-DTW [4] and MCD-DTW-SL [9] are adopted, which

Methods	LSE-D ↓	LSE-C ↑	MCD ↓	MCD-DTW ↓	MCD-DTW-SL ↓	Id. Acc. ↑	Emo. Acc. ↑	MOS-N ↑	MOS-S ↑
Ground Truth	6.734	7.813	00.00	00.00	00.00	90.62	84.38	4.61 ± 0.15	4.74 ± 0.12
SV2TTS [21]	13.733	2.725	21.08	12.87	49.56	33.62	37.19	2.03 ± 0.22	1.92 ± 0.15
SV2TTS* [21]	12.617	3.349	19.38	12.73	34.51	35.18	42.05	2.07 ± 0.07	2.15 ± 0.09
Tacotron* [54]	13.475	2.938	19.79	18.73	42.15	32.49	39.68	2.12 ± 0.17	2.06 ± 0.12
FastSpeech2 [46]	12.261	2.958	20.78	14.39	19.41	21.72	46.82	2.79 ± 0.10	2.63 ± 0.09
FastSpeech2* [46]	12.113	2.604	20.66	14.59	20.79	23.44	46.90	3.08 ± 0.06	2.89 ± 0.07
V2C-Net[1] [9]	11.784	3.026	20.61	14.23	19.15	26.84	48.41	3.19 ± 0.04	3.06 ± 0.06
Ours	**8.036**	**5.608**	**15.66**	**12.29**	**13.48**	**37.75**	**61.46**	**4.03 ± 0.08**	**3.89 ± 0.07**

Table 1. Results on the V2C dataset with comparisons against state-of-the-art methods. We provide the results using both objective metrics (*i.e.*, LSE-D, LSE-C, MCD, MCD-DTW and MCD-DTW-SL) and subjective metrics (*i.e.*, MOS-Naturalness and MOS-Similarity). "Id. Acc." and "Emo. Acc." are the identity and emotion accuracy of the generated speech, respectively. The method with "*" refers to a variant taking video (emotion) embedding as an additional input as in [9]. ↑ (↓) means that the higher (lower) value is better.

reflect the similarity of mel-spectrograms. MCD-DTW uses the Dynamic Time Warping (DTW) [37] algorithm to find the minimum MCD between two speeches, while MCD-DTW-SL introduces the duration measure coefficient to consider the length and the quality of generated speech [9].

Emotion and identity accuracy. To measure whether the generated speech carries proper emotion and speaker identity, we adopt an emotion accuracy (Emo. Acc.) and an identity accuracy (Id. Acc.) as our metrics as in [9].

Subjective evaluations. To further evaluate the quality of generated speech, we conduct a human study using a subjective evaluation metric, following the settings in [9]. 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.

4.3. Implementation Details

For the duration aligner, we use 4 Feed-Forward Transformer (FFT) blocks and 3 FFT blocks for the phoneme encoder and lip movement encoder, respectively. We set the dimension of the phoneme feature $\mathcal{O}$ and lip features E_{lip} to 256. We use 8 attention heads for alignment between lip and phoneme. For each movie clip, Our FPS set is 25, the sampling rate sr is $22050Hz$. We use short-time fourier transform (STFT) to obtain the mel-spectrum, and the number of points of the fourier transform is 1024. We use the Conv-Transpose1D module with 2 stride and 4 kernel sizes to obtain the duration features. In Valence and Arousal Feature Encoder, the EmoFAN consists of one 2D convolution with a kernel size of 7 × 7 and 3 convolution blocks (ConvBlock) with a kernel size of 3 × 3 and Average Pooling stride of 2 × 2. Similarly, we set the dimension of the valence and arousal feature to 256. For the mel-spectrograms generator, the mel-spectrogram decoder consists of 6 FFT blocks and the hidden state of mel-linear layer is of size 80.

For training, we use Adam [22] with $\beta_1 = 0.9$, $\beta_2 = 0.98$, ϵ=10^{-9} to optimize our model. For V2C and Chem dataset, we set the learning rate schedule to 0.00001 and 0.00005, respectively. In this work, we use pretrained HiFiGAN [23] as the vocoder to transform the generated mel-spectrograms into audio samples. We set the batch size to 16 on two datasets. Our model is implemented in PyTorch [39]. All the models are performed on a single NVIDIA GTX3090Ti GPU. We train the model with 600 epochs on the V2C dataset and 400 epochs on the Chem dataset.

4.4. Quantitative Evaluation

We compare with five related baselines of speech synthesizes. (1) SV2TTS [21] is a basic TTS model to generate speech with reference audio for multi-speakers; (2) Tacotron [54] is an end-to-end generative TTS model that synthesizes speech directly from textual characters; (3) FastSpeech2 [46] introduces the variance adaptor to convert the text to waveform by end-to-end; (4) Neural Dubber [20] synthesizes human speech for given video according to the corresponding text; (5) V2C-Net [9] is the first model for movie dubbing, which match the speaker's emotion presented in video. Note that we do not compare with Neural Dubber [20] on the V2C benchmark due to the unavailability of its code and missing implementation details.

Results on the V2C benchmark. The results are presented in Table 1. Our method achieves the best performance on all nine metrics. Specifically, in terms of audio-visual sync, our method achieves 8.036 of LSE-D and 5.608 of LSE-C, which significantly surpasses the previous best results and is much closer to human performance. In terms of MCD, MCD-DTW, and MCD-DTW-SL, our method achieves relative 24.02%, 13.63% and 29.61% improvements, respectively. This indicates our method can achieve a better mel-spectrogram than others. The above results together show that bridging specific attributes of speech with corresponding visual counterparts can make the generated speech present better prosodies and lip motion sync. Additionally, our method outperforms the previous method by a large margin in emotion accuracy and can gain better identification accuracy. This indicates the proposed method can better capture and convey emotions, which is of great importance for the movie dubbing task. Last, the human subjec-

Methods	AQ ↑	AV Sync ↑	LSE-D ↓	LSE-C ↑
Ground Truth	3.93 ± 0.08	4.13±0.07	6.926	7.711
FastSpeech2 [46]	3.71±0.08	3.29±0.09	11.86	2.805
Tacotron* [54]	3.55±0.09	3.03±0.10	11.79	2.231
V2C-Net [9]	3.48±0.14	3.25±0.11	11.26	2.907
Neural Dubber [20]	3.74±0.08	3.91±0.07	7.212	7.037
Ours	**3.84±0.11**	**3.97±0.08**	**6.975**	**7.176**

Table 2. Results on the Chem dataset with comparisons against state-of-the-art methods. AQ (Audio Quality) and AV Sync (audio-visual synchronization) are subjective metrics. Note that the Chem dataset is a single-speaker non-movie dataset, and thus there is no identity accuracy and emotion accuracy.

#	Methods	LSE-D ↓	LSE-C ↑	MCD ↓	Id. Acc. ↑	Emo. Acc. ↑
1	w/o DA	11.835	3.716	19.53	18.75	38.33
2	w/o PA	8.514	5.274	16.34	10.42	22.08
3	w/o AB	8.261	5.408	15.90	33.33	53.92
4	w/o Valence	9.215	4.935	16.31	29.81	35.31
5	w/o Arousal	8.793	5.216	15.79	32.74	46.38
6	VA $v.s.$ FF	9.160	5.011	19.87	23.67	39.58
7	w/o multi-head	10.948	3.894	18.65	24.85	40.25
8	Duplication	11.475	3.814	18.63	21.78	37.92
9	Full model	**8.036**	**5.608**	**15.66**	**37.75**	**61.46**

Table 3. Ablation study of the proposed method on the V2C benchmark dataset.

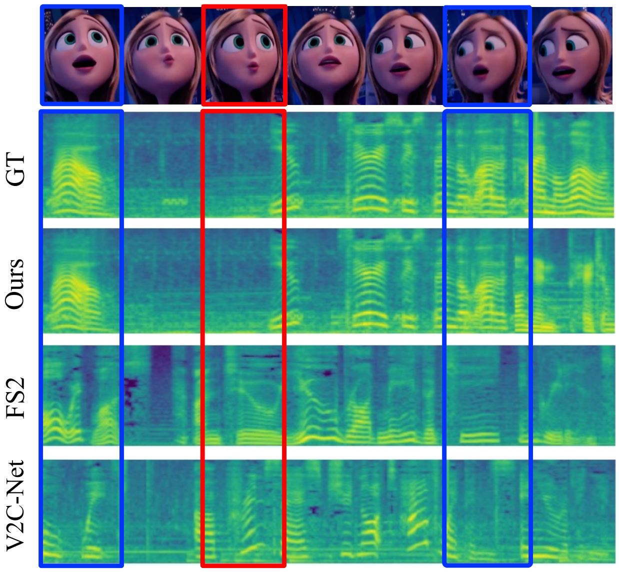


Figure 3. Audiovisual consistency visualization on V2C dataset: Ground Truth (GT), our model, FastSpeech 2 (FS2) and V2C-Net.

tive evaluation results (see MOS-N and MOS-S) also show that our method can generate speeches that are closer to realistic speech according to naturalness and similarity.

Results on the Chem benchmark. As shown in Table 2, our model is ahead of the state-of-the-art methods in all metrics on the Chem benchmark. In terms of the audio-visual sync, our method achieves 6.975 LSE-D and 7.176 LSE-C. Furthermore, in the subjective evaluations, our method improves 10.34% on AQ and 22.15% on AV Sync. The results show that our performance is much closer to the ground truth recording, which indicates that our model synthesizes high-quality natural speech by controlling the prosody from hierarchical visual representation.

4.5. Ablation Studies

Effectiveness of Duration Aligner, Prosody Adaptor, and Atmosphere Booster. We evaluate the effectiveness of these three modules by removing them separately and retraining the model. The results are shown in Row 1∼3 of Table 3. The result shows that all the proposed modules contribute significantly to the overall performance, and each module has a different focus. Specifically, the performance on the audio-visual metrics (LSE-D, LSE-C, and MCD) drops the most when removing the Duration Aligner (DA). This reflects the DA module indeed helps the model learn a better temporal synchronization. By contrast, the performance on identity accuracy and emotion accuracy drop the most when removing the Prosody Adaptor (PA). This can be attributed to the predicted pitch and energy representing a speaker's identity and his/her emotion. When removing the Atmosphere Booster (AB), the performance drops compared to the full model (Row 9) but does not drop as much as when removing the other two modules. This indicates the AB module also contributes the overall performance improvement but contributes the least in the three modules.

Effectiveness of valence and arousal. In our model, we bridge arousal with energy and valence with pitch by attention mechanism. To evaluate their effectiveness, we cut off the connection and predict energy and pitch directly from the phoneme-lip representation. The results are presented in Row 4∼5 of Table 3. It shows that the performance drops significantly when removing either of them and drops more when removing valence. This indicates valence contributes more than arousal on the V2C task.

Valence-and-Arousal v.s. Facial Expression. To compare the role of affective display and facial expressions on prosody inference, we replace the input of the APA module with original embeddings of facial features from CNN. As row "VA $v.s.$ FF" of Table 3 shows, the model performance drops significantly. This is likely caused by facial features that are still far from the information-containing emotion, which is not enough to guide prosody generation.

Multi-head Attention in Duration Adaptor. In our duration aligner, we exploit multi-head attention to learn the relation between the phoneme sequences and the lip motion. To verify its effectiveness, we conduct experiments using conventional dot-product attention. The results are

[1] V2C republish official results on https://github.com/chenqi008/V2C.

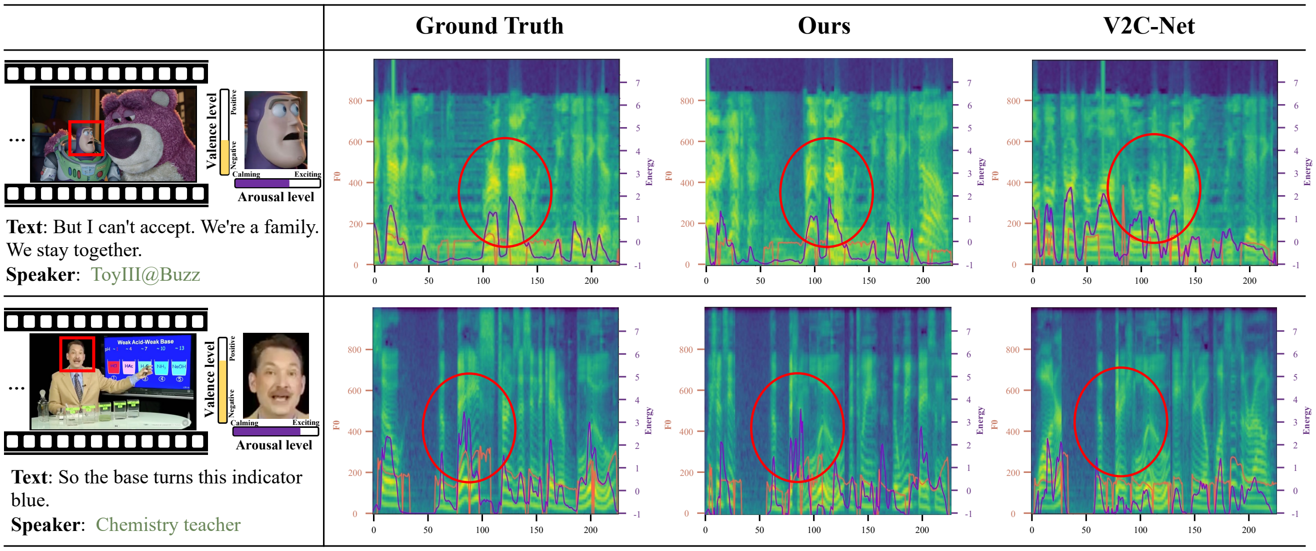


Figure 4. Visualization of audio on V2C dataset (top) and Chem dataset (bottom). Orange curves are F_0 contours, where F_0 is the fundamental frequency of audio. Purple curves refer to the energy (volume) of audio. The horizontal axis is the duration of the audio. The red circles highlight the mel-spectrograms at the same moment as the frame shown on the left side.

presented in Row 7 of Table 3. The performance drops significantly on all metrics, such as 26.59% and 44.02% decrease on LSE-D and LSE-C compared to Row 9, respectively. This indicates multi-head attention learns a much better correlation between phonemes with lip movement.

Conv-Transpose v.s. Duplication. In our duration aligner, we propose to use Conv-Transpose to learn the expansion of the fused phoneme-lip representation to its desired length. To verify its effectiveness, we replace it with simple make n duplications as in [20]. The results are shown in Row 8 of Table 3. It shows that the performance drops significantly. For example, it falls 15.94% on MCD metric. This demonstrates the superiority of using transposed convolution to learn the upsampling than simply copying.

4.6. Qualitative Results

Audiovisual consistency visualization. Figure 3 presents the mel-spectrograms of generated audios along with its frames. Blue and red bounding boxes denote whether the character is speaking or not, respectively. Compared with other methods, the mel-spectrogram generated by our model is closer to the ground truth, indicating better audio-visual synchronization. This can be attributed to our duration aligner, which exploits multi-head alignment between the text phoneme sequence and the lip movement sequence. By controlling the lip movements explicitly, we obtain the desired length of mel-spectrograms, which makes the speech well synchronized with the input video.

Arousal and valence with prosody visualization. We selected two examples from the test set of the V2C dataset and Chem dataset to demonstrate the alignment between energy and arousal as well as pitch and valence. The valence (positive or negative) and arousal (calming or exciting) of facial expressions are shown in the first column. The main pitch and energy are shown in orange and blue curves in the right column. We use the red circle to highlight the pitch and energy that correspond to the video frame shown in the left column. The result shows that our method achieves energy and pitch closer to the ground truth speech. When the chemistry teacher becomes excited and positive, our model successfully leverages the affective display to synthesize a similar pitch and energy as the ground-truth speech.

5. Conclusion

In this work, we propose a hierarchical prosody modeling network for movie dubbing, which bridges video representations and speech attributes from three levels: lip, facial expression, and scene. By associating these visual representations with their voice counterparts, we obtain more powerful representations for dubbing. Furthermore, we design an affective-display based prosody adaptor, which effectively learns to align the valence and arousal to the pitch and energy of speeches. Our proposed model sets new state-of-the-art on both Chem and V2C-Animation benchmarks.

Acknowledgement. This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102000, and in part by National Natural Science Foundation of China under Grants 62225207 and U19B2038, the Key R&D Plan Project of Zhejiang Province (No. 2023C01004), Shandong Provincial Natural Science Foundation (ZR2020MA064), Youth Innovation Promotion Association of CAS under Grant 2020108.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Deep lip reading: A comparison of models and an online application. In *Interspeech*, pages 3514–3518, 2018. 2

[2] Sercan Ömer Arik, Mike Chrzanowski, Adam Coates, Gregory Frederick Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Y. Ng, Jonathan Raiman, Shubho Sengupta, and Mohammad Shoeybi. Deep voice: Real-time neural text-to-speech. In *ICML*, pages 195–204, 2017. 2

[3] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016. 2

[4] 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. 5

[5] Manuel Burghardt, Michael Kao, and Christian Wolff. Beyond shot lengths - using language data and color information as additional parameters for quantitative movie analysis. In *ADHO*, pages 753–755, 2016. 5

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 5

[7] Frederic Chaume. Dubbing practices in europe: localisation beats globalisation. *LANS–TTS*, 6, 2007. 2

[8] Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, pages 7832–7841, 2019. 2

[9] 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, 3, 5, 6, 7

[10] Weidong Chen, Dexiang Hong, Yuankai Qi, Zhenjun Han, Shuhui Wang, Laiyun Qing, Qingming Huang, and Guorong Li. Multi-attention network for compressed video referring object segmentation. In *ACM MM*, pages 4416–4425, 2022. 3

[11] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *CVPR*, pages 3444–3453, 2017. 2

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

[13] Joon Son Chung and Andrew Zisserman. Lip reading in profile. In *BMVC*, 2017. 2

[14] Gaoxiang Cong, Liang Li, Zhenhuan Liu, Yunbin Tu, Weijun Qin, Shenyuan Zhang, Chengang Yan, Wenyu Wang, and Bin Jiang. LS-GAN: iterative language-based image manipulation via long and short term consistency reasoning. In *ACM MM*, pages 4496–4504, 2022. 3

[15] Rodrigo Schoburg Carrillo de Mira, Alexandros Haliassos, Stavros Petridis, Björn W. Schuller, and Maja Pantic. SVTS: scalable video-to-speech synthesis. In *INTERSPEECH*, pages 1836–1840, 2022. 2

[16] Ariel Ephrat and Shmuel Peleg. Vid2speech: Speech reconstruction from silent video. In *ICASSP*, pages 5095–5099, 2017. 2

[17] 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. 2

[18] Sindhu B. Hegde, K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. Lip-to-speech synthesis for arbitrary speakers in the wild. In *ACM MM*, pages 6250–6258, 2022. 1, 2

[19] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez Opazo, and Stephen Gould. A recurrent vision-and-language BERT for navigation. *CoRR*, abs/2011.13922, 2020. 3

[20] 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, 3, 4, 5, 6, 7, 8

[21] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez-Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *NeurIPS*, pages 4485–4495, 2018. 6

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015. 6

[23] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *NeurIPS*, 2020. 5, 6

[24] R. Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *PACRIM*, pages 125–128, 1993. 5

[25] Yaman Kumar, Rohit Jain, Khwaja Mohd. Salik, Rajiv Ratn Shah, Yifang Yin, and Roger Zimmermann. Lipper: Synthesizing thy speech using multi-view lipreading. In *AAAI*, pages 2588–2595, 2019. 2

[26] Liang Li, Xingyu Gao, Jincan Deng, Yunbin Tu, Zheng-Jun Zha, and Qingming Huang. Long short-term relation transformer with global gating for video captioning. *IEEE Trans. Image Process.*, pages 2726–2738, 2022. 3

[27] Liang Li, Chenggang Clarence Yan, Xing Chen, Chunjie Zhang, Jian Yin, Baochen Jiang, and Qingming Huang. Distributed image understanding with semantic dictionary and semantic expansion. *Neurocomputing*, pages 384–392, 2016. 3

[28] Liang Li, Xinge Zhu, Yiming Hao, Shuhui Wang, Xingyu Gao, and Qingming Huang. A hierarchical cnn-rnn approach for visual emotion classification. *ACM Trans. Multim. Comput. Commun. Appl.*, pages 1–17, 2019. 2

[29] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *AAAI*, pages 6706–6713, 2019. 2

[30] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *CVPR*, pages 3377–3386, 2022. 2

[31] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *CVPR*, pages 10452–10462, 2022. 2

[32] 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. 1, 2

[33] Pingchuan Ma, Brais Martinez, Stavros Petridis, and Maja Pantic. Towards practical lipreading with distilled and efficient models. In *ICASSP*, pages 7608–7612, 2021. 3

[34] Pingchuan Ma, Yujiang Wang, Stavros Petridis, Jie Shen, and Maja Pantic. Training strategies for improved lip-reading. In *ICASSP*, pages 8472–8476, 2022. 3

[35] Pingchuan Ma, Yujiang Wang, Jie Shen, Stavros Petridis, and Maja Pantic. Lip-reading with densely connected temporal convolutional networks. In *WACV*, pages 2857–2866, 2021. 3

[36] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Lipreading using temporal convolutional networks. In *ICASSP*, pages 6319–6323, 2020. 3

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

[38] Foivos Paraperas Papantoniou, Panagiotis Paraskevas Filntisis, Petros Maragos, and Anastasios Roussos. Neural emotion director: Speech-preserving semantic control of facial expressions in ”in-the-wild” videos. In *CVPR*, pages 18759–18768, 2022. 2

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019. 6

[40] Wei Ping, Kainan Peng, and Jitong Chen. Clarinet: Parallel wave generation in end-to-end text-to-speech. In *ICLR*, 2019. 2

[41] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In *ICLR*, 2018. 2

[42] 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. 2, 5

[43] 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. 2, 5

[44] Yanyuan Qiao, Qi Chen, Chaorui Deng, Ning Ding, Yuankai Qi, Mingkui Tan, Xincheng Ren, and Qi Wu. R-GAN: exploring human-like way for reasonable text-to-image synthesis via generative adversarial networks. In *ACM MM*, pages 2085–2093, 2021. 3

[45] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. Hop+: History-enhanced and order-aware pre-training for vision-and-language navigation. *TPAMI*, 2023. 3

[46] 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. 1, 2, 6, 7

[47] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. In *NeurIPS*, pages 3165–3174, 2019. 1, 2, 5

[48] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ-Skerrv Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *ICASSP*, pages 4779–4783, 2018. 1, 2, 5

[49] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *ICLR*, 2022. 2

[50] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *IEEE Trans. Inf. Forensics Secur.*, 17:585–598, 2022. 2

[51] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nat. Mach. Intell.*, 3(1):42–50, 2021. 2, 4

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 5

[53] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *AAAI*, pages 2531–2539, 2022. 2

[54] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. In *Interspeech*, pages 4006–4010, 2017. 1, 2, 6, 7

[55] Yongqi Wang and Zhou Zhao. Fastlts: Non-autoregressive end-to-end unconstrained lip-to-speech synthesis. In *ACM MM*, pages 5678–5687, 2022. 1, 2

[56] Jonatas Wehrmann and Rodrigo C. Barros. Movie genre classification: A multi-label approach based on convolutions through time. *Appl. Soft Comput.*, 61:973–982, 2017. 5

[57] Tianyi Xie, Liucheng Liao, Cheng Bi, Benlai Tang, Xiang Yin, Jianfei Yang, Mingjie Wang, Jiali Yao, Yang Zhang, and Zejun Ma. Towards realistic visual dubbing with heterogeneous sources. In *ACM MM*, pages 1739–1747, 2021. 2

[58] Yi Yang, Brendan Shillingford, Yannis M. Assael, Miaosen Wang, Wendi Liu, Yutian Chen, Yu Zhang, Eren Sezener, Luis C. Cobo, Misha Denil, Yusuf Aytar, and Nando de Freitas. Large-scale multilingual audio visual dubbing. *CoRR*, abs/2011.03530, 2020. 2

[59] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pretrained stylegan. *arXiv preprint arXiv:2203.04036*, 2022. 2

[60] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *CVPR*, pages 192–201, 2017. 4

[61] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, pages 9299–9306, 2019. 2

[62] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Trans. Graph.*, 39(6):221:1–221:15, 2020. 2

[63] Yang Zhou, Jimei Yang, Dingzeyu Li, Jun Saito, Deepali Aneja, and Evangelos Kalogerakis. Audio-driven neural gesture reenactment with video motion graphs. In *CVPR*, pages 3408–3418, 2022. 2

[64] Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Eric Wang, Qi Wu, Miguel P. Eckstein, and William Yang Wang. Diagnosing vision-and-language navigation: What really matters. *CoRR*, abs/2103.16561, 2021. 3