INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Prosody Modeling with 3D Visual Information for Expressive Video Dubbing

Zhihan Yang[1†], *Shansong Liu*[2*], *Xu Li*[2*], *Haozhe Wu*[3], *Zhiyong Wu*[1*], *Ying Shan*[2], *Jia Jia*[3]

[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]ARC Lab, Tencent PCG
[3]Department of Computer Science and Technology, Tsinghua University, Beijing, China

zhihan-y21@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn, shansongliu@tencent.com, nelsonxli@tencent.com

Abstract

The automatic video dubbing task is proposed to meet personal and industrial demands for dubbing. Current methods mostly focus on duration matching and overlook the synchronization of prosody, and thus lack expressiveness. In this paper, we introduce visual prosody modeling to promote expressiveness for video dubbing, defined as the expression and head pose in 3D space, which has the advantages of 1) high relevance to the tone and stress of utterances; 2) more accurate than 2D images; 3) disentanglement from irrelevant factors such as speaker identity. We propose a 3D-VD (3D Video Dubber) system to incorporate visual prosody, utilizing a visual-text step-wise aligner to control the generated prosody. Experiments demonstrate that the proposed method outperforms previous methods that only consider 2D face images in terms of naturalness, lip-speech alignment, and synchronization of visual and auditory prosody. The case study demonstrates the correlation between expression and pitch.

Index Terms: visual text-to-speech, speech synthesis, multi-modality generation

1. Introduction

Dubbing, in video production, describes the process of adding new dialogue of actors to the soundtrack, commonly required by both commercial film companies and self-published video bloggers. With the explosion of short videos and the development of the film industry, there is a huge demand for video dubbing techniques. However, video dubbing is costly, requiring a specific recording studio and professional voice actors, as well as a certain recording period.

The Automatic Video Dubbing (AVD) task [1], proposed to meet the demand of automatic dubbing, is to synthesize speech synchronized with a given silent video according to the corresponding script. Current deep learning methods mainly focus on two aspects: duration control and lip synchronization. The former usually involves machine translation (MT) techniques to improve the duration matching [2, 3, 4]. The latter mostly focuses on text-to-speech (TTS) methods to promote synchronization between the lip and the speech. For example, visual TTS [5] establishes a text-video aligner and a vision fusion module to introduce lip motion to control the generating process of Tacotron2 [6]. However, all of these methods overlook the prosodic relationship between video and speech, resulting in the lack of expressiveness of generated speech, hardly meeting the needs of voice dubbing for expressiveness.

One of the approaches to obtain expressiveness is to use video information to conduct prosody modeling. The expression and movement of the speaker, described as "visual prosody", always convey concordant information consistent with the speech prosody, such as tone and emphasis [7]. The head movement correlates strongly with the pitch (fundamental frequency) and amplitude of the talker's voice [8]. For example, shaking heads and staring may reflect a high-level activation of emotion. Due to the significance of visual prosody, it is practical and necessary to explore the manipulation of visual prosody in video dubbing, which can enhance the naturalness and expressiveness of the synthesized speech. There exist two main challenges to applying visual prosody: 1) How to capture and model the visual dynamic characteristics of the speaker; 2) How to align this visual feature with text in a typical TTS architecture.

Previous works explored various types of representations of visual prosody. For example, facial gestures and emotions using the pleasure-arousability-dominance (PAD) model are related to pitch changes [9, 10, 11]. Recent studies, such as Neural Dubber [1], have proposed a text-video aligner adapting multi-head attention [12] and upsampling to align text with video for speech synthesis, generating fluent and natural dubbed videos. VDTTS [13] introduces a multi-source attention mechanism that concatenates the context vectors of features from two modalities. However, the diagonal constraint overlooks the different lengths of pauses and does not strictly guarantee monotonicity. To summarize, these works are meaningful but lack relevance and accuracy in terms of prosody modeling and alignment.

In this paper, we define visual prosody as expression and head pose in 3D space. It has three advantages: 1) **Relevance**. Expression and pose of the speaker are highly related to the pitch and stress of the utterance [7, 14], such as the correlation between excited expression and high tone. 2) **Accuracy**. The expression and pose information obtained from a three-dimensional (3D) face is more accurate than that from the two-dimensional (2D) method, containing semantic information that can restore the face. 3) **Disentanglement**. Compared with the face, face representation decouples the noise information such as speaker identity and illumination. We incorporate the visual prosody and mouth region embedding coherently, not only predicting pronunciation and duration information but also improving the ability of pitch and energy prediction that are highly correlated with the expressiveness of the generated speech. In addition, we adopt stepwise attention [15] for better alignment between visual and audio modalities. The method keeps strict monotonicity in TTS and constraints on hard attention, ensuring that the alignment between the visual and audio sequences is monotonic without skipping visual frames.

We propose a multi-modal framework, named 3D-VD (3D

† Work done when Z. Yang was an intern at Tencent ARC Lab.
∗ Corresponding authors: Z. Wu , S. Liu, X. Li

10.21437/Interspeech.2023-289

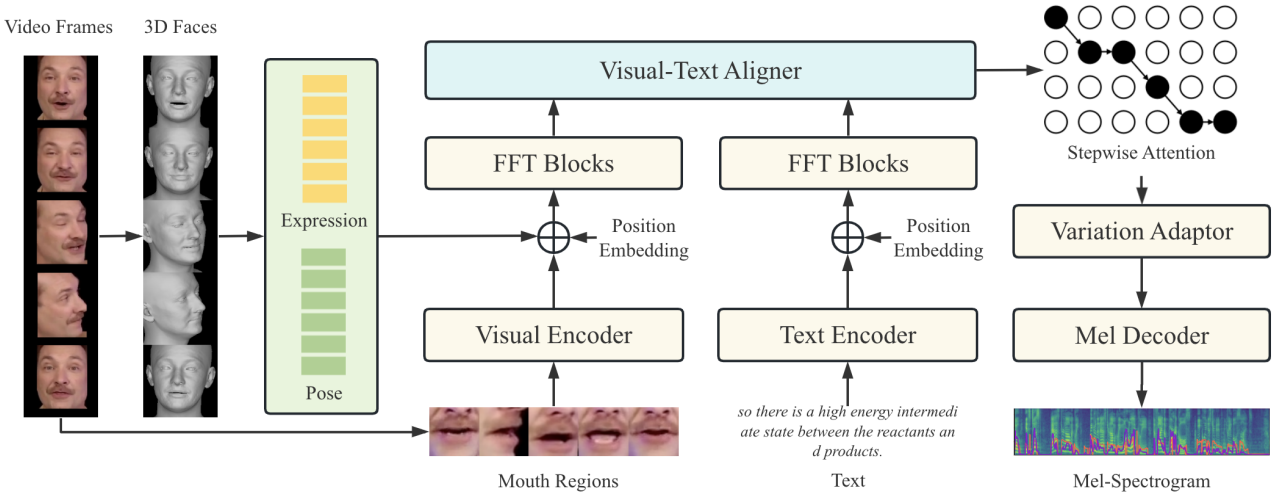

Figure 1: *The pipeline of our model. We first reconstruct the 3D face and its parameters from video frames. Then we incorporate expression and pose with mouth features, aligning the prosody with the visual information, to generate the speech from the given text.*

Video Dubber), to incorporate visual prosody. First, we employ a 3D face constructor to restore the 3D face shape and parameters of its expression and pose from a video. We concatenate the visual prosody with mouth features extracted from a visual encoder. Then we utilize a visual-text stepwise aligner to align the visual and text features and upsample the aligned context features to the length of mel-spectrogram features. The aligned features are utilized to predict the pitch and energy of the speech because it contains the information from the visual prosody. We implement our framework on a dataset from Lip2Wav [16] and conduct experiments. Objective comparison and subjective evaluations suggest that the speech produced by our model qualifies for matching the visual prosody of the speaker, along with high audio quality and synchronization, outperforming the method considering 2D face image merely.

We summarize our contributions as follows:

- We introduce visual prosody matching in video dubbing tasks, that is, the prosody of synthetic speech is matched with the visual prosody (pose, expression) of the speaker;
- We model the visual prosody information in the form of 3D face parameters, and build a complete framework for synthesizing synchronous and expressive voice from silent video;
- The experimental results show that the proposed model is better than baselines in naturalness, synchronization, and expressiveness, and demonstrates the correlation between visual prosody and speech prosody.

2. Methodology

In this section, we introduce the framework of our 3D-VD as shown in Fig 1. First, we employ a 3D face constructor to restore 3D face shape and parameters of its expression and pose from a video. We concatenate the visual prosody with mouth features extracted from a visual encoder. Then we utilize a visual-text stepwise aligner to align the visual and text features. The aligned features are utilized to predict the pitch and energy of the speech, and finally converted to mel-spectrogram.

2.1. Visual Prosody and 3D Face Reconstruction

Given a video clip $S_v = [V_1, V_2, ...V_{T_v}]$, we restore the 3D face model and extract the parameter vectors of expressions and poses, as $e = [e_1, e_2, ..., e_{T_v}]$ and $p = [p_1, p_2, ..., p_{T_v}]$ respectively. By this approach, we only reserve the related factors, discarding the parameters of speaker identity, lighting, and other parameters which are irrelevant to visual prosody. We adopt the FLAME [17] statistical face model to represent each face and infer the reconstructed parameters by DECA [18] model.

The face model M is controlled by three groups of parameters: facial identity $i \in \mathbb{R}^{|i|}$, pose $p \in \mathbb{R}^{3k+3}$ (where k is additional joints of neck, jaw, and eyeballs) and expression $e \in \mathbb{R}^{|e|}$.

$$M(i, p, e) = W(T_P(i, p, e), J(i), p, \mathcal{W}) \quad (1)$$

where $W(T, J, p, \mathcal{W})$ is a blend skinning function, rotating the vertices of the face in $T \in \mathbb{R}^{3n}$ around the joints $J \in \mathbb{R}^{3k}$. $\mathcal{W}$ is blendweights for linearly smoothing. J provides joint locations by identity i.

Finally, we introduce the visual prosody $v = [v_1, v_2, ..., v_T]$, defined as the concatenation of sequences of expression e and pose p.

$$v_t = [e_t; p_t] \quad (2)$$

2.2. 3D Video Dubber

We incorporate the visual prosody and utilize it to control the speech synthesis by 3D-VD. Given a silent video frame sequence S_v and a phoneme sequence $S_p = [P_1, P_2, ..., P_N]$, the 3D-VD predicts a dubbing mel-spectrogram sequence $S_m = [M_1, M_2, ..., M_{T_m}]$. The overall model architecture of 3D-VD is shown in Figure 1.

The text encoder turns the phoneme sequence S_p into hidden representations $H_p = f_p(S_p) \in \mathbb{R}^{N \times d}$, where d is the dimension of hidden space. In the visual module, video frames are processed by a 3D reconstruction f_r and a visual encoder f_v respectively. The face image sequences are cropped from the video and sent to the 3D reconstruction to extract the visual prosody $v = f_r(S_v) \in \mathbb{R}^{T_v \times (3k+3+|e|)}$. The visual encoder converts the images only containing the mouth region of the

speaker into mouth features $m = f_v(S_V) \in \mathbb{R}^{T\times|m|}$. The visual prosody v is concatenated to the mouth features m, and turned into visual hidden sequences $H_v = [v; m] \in \mathbb{R}^{T_v\times d}$.

Then we feed H_v and H_p to the visual-text stepwise aligner, which is introduced in detail in the next subsection, to get the aligned mel-spectrogram hidden sequence $H_m \in \mathbb{R}^{T_m\times d}$, which contains both content information (phoneme and viseme) and prosody information.

After the alignment, a variance adaptor utilizes H_m to predict the pitch and energy by different predictors. Following Neural Dubber [1], our variance adaptor contains no duration predictor, because the video and audio are naturally aligned, despite the mismatch between the length of the text and audio sequence.

2.3. Visual-Text Stepwise Aligner

We employ the visual-text stepwise aligner to align the video and text sequence to the same length as the speech. It not only requires the alignment of the content, which is between the lip motion and phoneme, but also the alignment of the visual and auditory prosody, which is between the expression and head pose of the speaker and the pitch and energy in the utterance.

Stepwise attention [15] is an expanded variation of monotonic attention. Monotonic attention [19] is effective to ensure the monotonicity and locality of alignment. The mechanism generates attention scores step by step: Given the visual features H_v as query entry and text feature H_p as key and value, an energy value is calculated as the attention score of multi-head scale-dot attention [12]:

$$E = Attention(Q, K) = \frac{H_v H_p^T}{\sqrt{d}} \tag{3}$$

Instead of directly calculating the attention scores, the energy value produces a select probability to decide the attention score step by step:

$$P = Softmax(E) \tag{4}$$

At each time step, the method decides whether to inspect the next entry by sampling from a Bernoulli function, as $z_{i,j} \sim Bernoulli(p_{i,j})$. Instead of a "soft" attention with continuous weights, the "stepwise" property means that it only decides to move forward by one step or stay at the temporary step.

$$a_{i,j} = a_{i-1,j-1}(1 - p_{i,j}) + a_{i-1,j}p_{ij} \tag{5}$$

where $a_{i,j}$ is one element of the attention score matrix A. Finally, we get the visual-text context features by applying the attention scores to the value matrix H_p:

$$H_c = AH_p \tag{6}$$

The context vector has the same length as H_v, which is naturally aligned with the mel vector because video and audio are synchronized in time sequence. Therefore, we can align the context feature to the mel feature by applying the ratio of respective sampling rates r.

$$H_m = Upsample(H_c, r) \tag{7}$$

Practically, $r = \frac{sr/hs}{FPS}$, where sr is the sampling rate of the waveform, hs is the hop size of the mel-spectrogram and FPS is the frame rate of the video.

Table 1: *The results of subjective tests. MOS Scores are presented with 95% confidence intervals.*

Method	MOS		
	Naturalness	AV Sync	Prosody Sync
GT (Mel+vocoder)	4.82 ± 0.01	4.81 ± 0.01	4.75 ± 0.01
FastSpeech2	2.83 ± 0.09	2.19 ± 0.13	2.42 ± 0.11
Neural Dubber	3.50 ± 0.04	4.24 ± 0.02	4.04 ± 0.03
3D-VD (ours)	$\mathbf{3.68 \pm 0.04}$	$\mathbf{4.30 \pm 0.03}$	$\mathbf{4.13 \pm 0.03}$

3. Experiment

3.1. Dataset

We utilize a part of Lip2Wav [16] dataset, the chemistry lecture sub-dataset, described as Chem below, with diverse information on speaker expressions and head poses, and also a rich diversity of speech prosody. It contains 346 video clips of several minutes individually from YouTube.

We first collected and segmented the scripts, and then filtered the clips without the face of the speaker. Finally, we construct our dataset consisting of 7,216 clips with scripts, having a total duration of around 9 hours. We randomly split Chem as 6,873 samples for training, and 343 samples for evaluation.

3.2. Implementation Details

Data Preparation. We resample the audio with the sample rate 19.2kHz, corresponding to the 30 FPS of videos. The window size and hop size are set as 640 samples and 160 samples. Mouth regions are marked by facial landmarks and cropped by a window whose size is 1.25 times the mouth landmarks. Cropped images are resized as 96×96. The dimension of the expression e and pose p is 50 and 6.

Model Configuration. The visual encoder has an architecture following Lipreading [20], starting with a 3D convolution layer and consisting of a ResNet18 as its main part [21]. The configurations of the FFT block, the mel spectrogram decoder, and the variation adaptor (including pitch and energy predictors)are the same as those in FastSpeech 2 [22].

Training. We train our model on 8 NVIDIA V100 GPU for $450k$ epochs, with a batch size of 16. We adopt the Adam optimizer with a learning rate of 0.0625. We use HiFiGAN [23] as our vocoder, and train on 1 NVIDIA V100 GPU for $1000k$ steps.

3.3. User Study

To examine the generation quality, we carried out the subjective test to assess the *Naturalness*, *AV Sync*, and *Prosody Sync*.

Comparison Systems. We compare the performance of our model with some alternative systems to synthesize speech. **GT (Mel+Vocoder)**. We convert the ground truth audio to mel (mel-spectrogram) and then reconstruct the mel to the waveform. We utilize this process to avoid the disturbance involved by the vocoder. **FastSpeech2** [22]. We adopt an open-source implementation [1], which generates speech only from the text. **Neural Dubber** [1]. It aligns the mouth and text features. We force its multi-head attention with stepwise constraints, as the reason discussed in Section 3.4.

[1] https://github.com/ming024/FastSpeech2

We use the mean opinion score (MOS) to evaluate the degree of satisfaction of users. The range of MOS is 1-5 with 1 point interval, the higher the better. We randomly sampled 15 video clips from the evaluation set, and we replaced the audio of these samples with the audio produced by the methods mentioned in Section 3.3. We provided the same text and video to all participants. 21 evaluators participated in the subjective tests.

Naturalness. We adapt this index to evaluate the noise level and pronunciation of the generated speech. As shown in Table 1, our model achieves a higher level than Neural Dubber and also surpasses FastSpeech2, indicating our system generates high-quality utterances. **AV Sync**, i.e. the synchronization between audio and lip motion. 3D-VD slightly outperforms Neural Dubber, indicating that the expressions can also help the alignment of lips because the lip motions also drive the facial muscles. **Prosody Sync**, which is the synchronization between speech prosody (tone, emphasis, etc.) and the expression and head pose of the speaker. Our model outperforms Neural Dubber, indicating that our model successfully captures the visual prosody within the expression and pose and matches it with the speech prosody.

All results listed above indicate that our model generates speech with better quality, synchronization, and expressiveness, outperforming baselines.

3.4. Ablation Study

We implement an ablation study to test the performance of our model when each component is absent to examine the contributions of each component to speech generation.

LSE-D (Lip Sync Error-Distance) describes the minimum offset to align the audio and video frames synchronously. The lower the LSE-D, the higher the synchronization level of the generated audio matching the video. LSE-C (Lip Syn Error-Confidence) denotes the confidence score of the SyncNet to decide the LSE-D. A higher score of LSE-C means a higher probability to align the audio with the video. Following the Lip2Wav [16], we use STOI and ESTOI for estimating intelligibility and PESQ for measuring speech quality. The higher these indicators are, the better the generation speech is.

We remove the 3D parameters and stepwise method from our 3D-VD to verify the effectiveness of these two components. As the results shown in Table 2, stepwise attention is significant for the synchronization and audio quality. Without stepwise enforcement, multi-head attention tends to predict vague attention scores and unclear pronunciations. So that we add the stepwise constraints to Neural Dubber in Section 3.3 for better audio quality. The 3D parameters does not help improve the LSE-D and LSE-C because these indexes mainly focus on the alignment between lip motion and speech. However, the 3D parameters promote intelligibility in terms of STOI and ESTOI, so that the generated speech is easier to understand. These results also confirm the results of subjective tests in Table 1.

3.5. Case Study

We carry out a case study to intuitively demonstrate the relationship between expression and pose coefficients and the generated speech. We first define and calculate PV as the short-time pitch variation of audio, which is the variation of the pitch in a sliding window. We set the window size as 100 samples and the hop size as 10 samples on the mel-spectrogram. Then we define *pose* and *exp* as the sum of the absolute value of the pose and expression parameters and we normalize them between 0 to 1. We visualize the PV of utterances generated by different methods and corresponding pose and exp. As the case shown in Fig 2, higher fundamental frequency usually corresponds to a more obvious posture or expression, such as staring and shaking head. As shown in the red box, our model captures the motion of looking up and staring and thus predicts a more accurate pitch. Besides, the overall contour of 3D-VD is more similar to GT rather than the one of w/o 3D, indicating that visual prosody can help predict the prosody of speech.

Table 2: *The results of ablation study. 'w/o 3D' means expression and pose vectors are not incorporated to 3D-VD. 'w/o stepwise' means attention scores are not forced to align step-wisely, equivalent to vanilla multi-head attention.*

Method	STOI	ESTOI	PESQ	LSE-D(↓)	LSE-C(↑)
3D-VD	0.542	0.354	1.12	6.93	8.13
w/o 3D	0.535	0.349	1.12	6.94	8.13
w/o stepwise	0.434	0.170	1.07	10.60	3.94

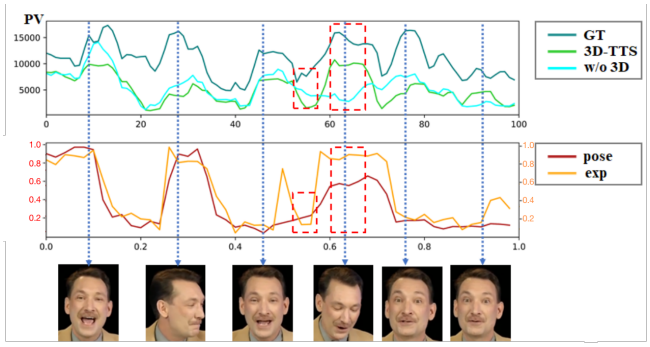

Figure 2: *The pitch variation of generated speech and the corresponding pose and expression parameters (normalized to [0,1]). 3D-VD captures the motion of looking up and staring and thus predicts a more accurate pitch.*

4. Conclusion

In this paper, we model the visual prosody information in the form of 3D expression and pose parameters, and build a complete framework for synthesizing synchronous and expressive voice from silent video. The 3D visual prosody takes the advantages of high relevance to speech prosody and disentanglement from noise information. The experimental results show that the proposed model is better than baseline in naturalness and synchronization, and demonstrate that the correlation between visual prosody and speech prosody.

5. Acknowledgement

This work is supported by National Key Research and Development Plan (2021QY1500), National Natural Science Foundation of China (62076144), Shenzhen Science and Technology Program (WDZC20220816140515001, JCYJ20220818101014030), Shenzhen Key Laboratory of next generation interactive media innovative technology (ZDSYS20210623092001004) and AMiner.Shenzhen SciBrain fund.

6. References

[1] C. Hu, Q. Tian, T. Li, W. Yuping, Y. Wang, and H. Zhao, "Neural dubber: dubbing for videos according to scripts," *Advances in neural information processing systems*, vol. 34, pp. 16 582–16 595, 2021.

[2] M. Federico, Y. Virkar, R. Enyedi, and R. Barra-Chicote, "Evaluating and optimizing prosodic alignment for automatic dubbing," 2020.

[3] Y. Wu, J. Guo, X. Tan, C. Zhang, B. Li, R. Song, L. He, S. Zhao, A. Menezes, and J. Bian, "Videodubber: Machine translation with speech-aware length control for video dubbing," *arXiv preprint arXiv:2211.16934*, 2022.

[4] Y. Virkar, M. Federico, R. Enyedi, and R. Barra-Chicote, "Prosodic alignment for off-screen automatic dubbing," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 496–500. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-11089

[5] J. Lu, B. Sisman, R. Liu, M. Zhang, and H. Li, "Visualtts: Tts with accurate lip-speech synchronization for automatic voice over," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8032–8036.

[6] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[7] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," in *Proceedings of fifth IEEE international conference on automatic face gesture recognition*. IEEE, 2002, pp. 396–401.

[8] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," *Psychological science*, vol. 15, no. 2, pp. 133–137, 2004.

[9] J. Jia, Z. Wu, S. Zhang, H. M. Meng, and L. Cai, "Head and facial gestures synthesis using pad model for an expressive talking avatar," *Multimedia Tools and Applications*, vol. 73, pp. 439–461, 2014.

[10] S. Zhang, Z. Wu, H. M. Meng, and L. Cai, "Head movement synthesis based on semantic and prosodic features for a chinese expressive avatar," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–837.

[11] ——, "Facial expression synthesis using pad emotional parameters for a chinese expressive avatar," in *Affective Computing and Intelligent Interaction*, A. C. R. Paiva, R. Prada, and R. W. Picard, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 24–35.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] M. Hassid, M. T. Ramanovich, B. Shillingford, M. Wang, Y. Jia, and T. Remez, "More than words: In-the-wild visually-driven prosody for text-to-speech," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 587–10 597.

[14] G. Cong, L. Li, Y. Qi, Z. Zha, Q. Wu, W. Wang, B. Jiang, M.-H. Yang, and Q. Huang, "Learning to dub movies via hierarchical prosody models," *arXiv e-prints*, pp. arXiv–2212, 2022.

[15] M. He, Y. Deng, and L. He, "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural TTS," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 1293–1297. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-1972

[16] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 796–13 805.

[17] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[18] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–13, 2021.

[19] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," in *International conference on machine learning*. PMLR, 2017, pp. 2837–2846.

[20] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6319–6323.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[23] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.