TRANSFORMER-S2A: ROBUST AND EFFICIENT SPEECH-TO-ANIMATION

Liyang Chen[1,3], Zhiyong Wu[1], Jun Ling[2,3], Runnan Li[3*], Xu Tan[3], Sheng Zhao[3]*

[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] Microsoft, China

cly21@mails.tsinghua.edu.cn, zywu@se.cuhk.edu.hk, lingjun@sjtu.edu.cn, {runnan.li, xu.tan, sheng.zhao}@microsoft.com

ABSTRACT

We propose a novel robust and efficient Speech-to-Animation (S2A) approach for synchronized facial animation generation in human-computer interaction. Compared with conventional approaches, the proposed approach utilizes phonetic posteriorgrams (PPGs) of spoken phonemes as input to ensure the cross-language and cross-speaker ability, and introduces corresponding prosody features (i.e. pitch and energy) to further enhance the expression of generated animation. Mixture-of-experts (MOE)-based Transformer is employed to better model contextual information while provide significant optimization on computation efficiency. Experiments demonstrate the effectiveness of the proposed approach on both objective and subjective evaluation with $17\times$ inference speedup compared with the state-of-the-art approach.

Index Terms— Speech-to-Animation, Transformer, Phonetic Posteriorgrams, Mixture-of-Experts

1. INTRODUCTION

With the development of human-computer interaction, the single modality of interaction can less and less satisfy the various interaction scenarios. To enhance user experience, especially in conversational scenarios, such as education and dialogue, multi-modal interaction is thus becoming more and more important. In these modalities, speech and vision, the two most natural communication channels, have drawn rising attraction. Since latest text-to-speech technology [1] can already synthesize human-like speech, how to produce high-quality synchronized facial animation becomes the key to provide high-performance multi-modal service. Speech-to-animation (S2A) technology is then developed to address this task. The generated animation parameters from S2A can be any predefined visual parameters, such as blendshapes or landmarks, to drive virtual avatars in different communication scenarios. S2A can also help in conventional content generation scenarios, such as movie production process, to minimize the expensive, time-consuming and non-scalable effort in producing synchronized facial animations.

S2A technology is developed to automatically estimate synchronized facial animation parameters from given speech, including lip, jaw and cheek-area movements. A rendering engine like Unreal Engine 4 (UE4) is then employed to generate the final animated avatar with these predicted parameters. One major research interest in S2A development is to find out appropriate speech representation for building a robust S2A system [2–6]. Compared with mel-frequency cepstral coefficients (MFCC) used in [2, 3], the latest proposed phonetic posteriorgrams (PPGs)-based approaches [4–6] can leverage the property of pre-trained large-scale automatic speech recognition (ASR) model [7], to provide speaker and language independent abilities for S2A systems. However, since PPGs are more about linguistic information, using PPGs only as input always fails in preserving prosody information and results in poor expressiveness in animation generation.

The model construction is another key component in S2A development. Early attempts about the unit selection [8, 9] were made to map acoustic phonemes to visual visemes. These methods suffer from poor synchronization between speech and animation, since they fail to embody continuous variance of animations in adjacent frames. With the development of deep learning, data-driven approaches including frame-level [3, 6] and sequence-level [2, 5, 10] methods are developed for S2A. In frame-level methods, T frames of features in a temporal window are stacked as input, and the animations are predicted with the window sliding along the given speech sequence. These methods can maintain temporal coherency in modeling but ignore the dependencies between suprasegmentals, leading to unexpected lip tremble since the preceding and following phonemes can significantly affect the animation of the current phoneme spoken. The sequence-level methods [2, 5, 10] directly process the whole speech sequence instead of batching it into multiple frames. A typical implementation is to employ bi-directional long short-term memory (BLSTM) to explore time-related features and generate plausible facial animations. These methods still have limitation in modeling suprasegmental contextual information and the auto-regressive manner could limit the inference efficiency.

---

*Corresponding authors.

ICASSP 2022

In this paper, we propose a robust and efficient S2A system to address aforementioned issues by using additional prosody features and mixture-of-experts (MOE) Transformer. For prosody features, we introduce pitch and energy as inputs to provide more variation information in speech. We further incorporate MOE-based Transformer to better model the contextual information and speed up the inference speed of sequential model. The contribution can be summarized as:

- Propose a robust and efficient S2A system to generate facial animation from speech. The proposed system is speaker and language independent with the use of PPG as input, and the expression of generated animation is better than the state-of-the-art (SOTA) with the use of prosody features as additional inputs.

- Propose MOE-based Transformer for S2A model construction. With the MOE layer, the system can better exploit the contextual information from the given input sequence by automatically selecting expert candidates.

2. METHODOLOGY

In this section, we firstly describe the optimized input features and then introduce MOE-based Transformer for modeling.

2.1. PPG Augmented with Prosody Features

Since collecting paired speech-animation data is a time-consuming and expensive work, it is challenging to develop a speaker-independent and cross-lingual S2A system. Inspired by previous works [4, 5, 11], we utilize PPGs as the basic speech representation to generate animations across speakers and languages with comparable synthesized quality.
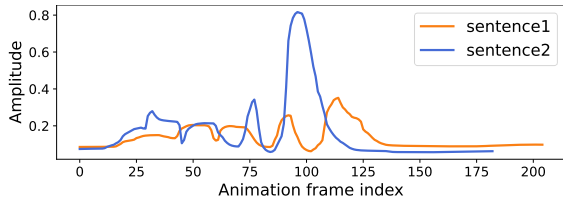

Fig. 1. The jawOpen curves of two sentences with the same context "millions of dollars".

PPGs conveys more linguistic information but less acoustic information, since prosody related information has been removed in the extraction of PPGs. Considering PPGs from the same speech transcription may be assigned with various animation presentations, the lack of prosody information may result in poor expression and variance in generated animations, which refers to the typical one-to-many mapping issue [1, 12]. As shown in Figure 1, the curves of jawOpen, one of the crucial parameters in animations, differ a lot from each other but share the same speech content. To prevent the
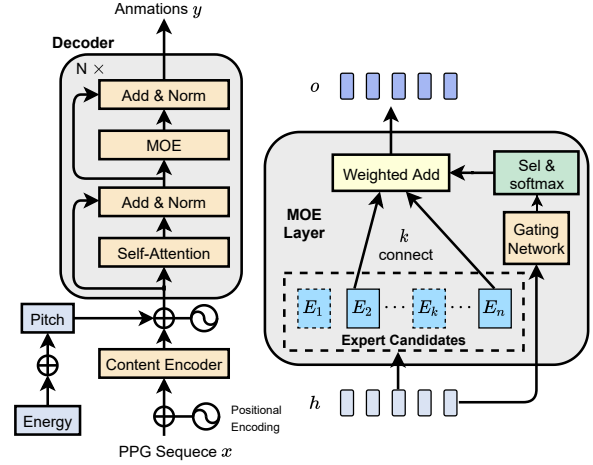

Fig. 2. Architecture of Transformer-S2A. Prosody features are concatenated while positional encoding is added to PPGs.

S2A model failing in capturing animation variations, additional prosody features, pitch and energy, are introduced to provide clear guidance for the model. It is consistent with the common sense that the louder and higher-pitched voice always makes the wider mouth open.

Spectral features [2, 13] are not directly used in this work for two reasons. 1) The spectral features are speaker-dependent. 2) Empirically, redundant information (e.g. background noise) in spectral features that has no relationship with pronunciation disturbs the training and causes instability in generation. By means of PPGs augmented with the prosody features, the proposed approach obtains noise, speaker and language auto-adapted ability in a more data-efficient way.

2.2. MOE-based Transformer

Transformer [14], a typical sequence-to-sequence network, has achieved great success in machine translation [15] and speech processing [1]. It mainly consists of two components: the self-attention mechanism and the feed-forward network. Compared with RNN-based models, Transformer generates outputs without conditioning current input on the previous output, and thus enables parallel training. In the meantime, the self-attention captures long-range and multi-level dependencies across the whole sequence by presenting the global context to each input frame. Nevertheless, the many-to-many mapping problem [8] as well as the cross-modal nature of S2A demands for specific architectures with more powerful context extraction ability, different from those employed for language understanding with single-modal.

The architecture of the proposed model is shown in Figure 2, comprising a content encoder and an expert decoder. Let $x = (x_1, \cdots, x_{t1})$ be the sequence of input PPGs, and $y = (y_1, \cdots, y_{t2})$ be the sequence of target facial animations. Due to the different sampling rate of PPGs and anima-

tions, we first interpolate x to become the same length $t2$ with y [4]. The content encoder that is stacked with the blocks in the Transformer encoder [14], takes PPGs as inputs and produces the hidden representation. The goal of the encoder is to extract robust text-related information and model the co-articulation effects from PPGs.

The expert decoder is constructed with several identical Transformer blocks that based on the MOE layer [16, 17]. As illustrated in Figure 2, we concatenate the hidden representation and prosody features pitch and energy to form the input sequence to the decoder. After being processed by the self-attention layer, the sequence h is further passed into the MOE layer, which consists of multiple experts $E_1, \cdots, E_n$ and a gating network G. The experts are of the same architecture but separated parameters, each comprising a convolutional layer with stride 1 and a fully connected layer. The gating network gives scores on how much effort each expert may contribute over the input. Only the experts with the top-k values are kept, and the $softmax$ function is then applied over the experts.

$$Prob(h) = Softmax(Score(h)) \tag{1}$$

$$Score(h)_i = \begin{cases} G_i(h), & \text{for top } k \text{ elements in } G(h) \\ -\infty, & \text{otherwise.} \end{cases} \tag{2}$$

The output of MOE layer can be formulated as

$$o = \sum_{i=0}^{n} G_i(h) Prob_i(h) \tag{3}$$

The number of expert candidates is set as 48 which roughly equals to the number of distinct phonemes in Mandarin or English, and k is set as 16 according to the defined basic visemes in [18]. Although MOE is originally designed for enlarging the model capacity without proportionally increasing computation [16], we find it works for S2A task under such setting. This can be attributed to the inductive bias for phoneme-to-viseme introduced by the MOE structure, where all the experts collaborate to search the specific mapping from the hypothesis space in a more efficient way.

3. EXPERIMENTS

3.1. Experimental Setup

Rendering engine. In this work, the digital avatar character is developed based on UE4, that is, a game creation engine commonly used in the gaming industry. Benefiting from the UE creation nature that the animation is disentangled with the character, the animation predicted by the pre-trained S2A model is able to drive any digital character that is designed following the developing rule of UE4. Hence, these features contribute to a reliable character-independent pipeline while

almost no prior knowledge about game creation is required. This creation tool also can be replaced with other 3D computer graphics application, e.g. Unity and Maya. Synthesized demo videos are available online[1].

Data collection. We utilize the mobile software Live Link Face[2] and the 3D character kite-boy[3] that is freely available for non-commercial use to collect training data. The blendshape coefficients recorded by the software contain timestamps of dictionary that describe the facial expression of the detected face, and we only select 32 dimensional coefficients that related to pronunciation as the target animation for training. Compared with other works [2, 19] that involve commercial system or extra depth estimation algorithm, the only thing we need is an iPhone that supports face recognition, which greatly reduces the complexity and cost in high-quality data collection. Only **one-hour** paired speech and animation data is collected from a single person for model training. Ten-minute outside data is reserved as the test set.

PPG extraction. we pre-train an automatic speech recognition (ASR) model with easily-available speech corpora of thousands of Chinese and English speakers, and extract the bottleneck of the model with dimension 64 as PPG. Speech data is sampled at 16 kHz and 80-dimensional filter banks computed with a sliding hamming window of 40ms width and 20ms overlap are used as the inputs to the ASR model.

Data pre-process for S2A. The prosody features pitch and energy are extracted with the same setting with filter banks in the PPG model. Voice activity detection is performed to remove silence frames at the start and end portion of the utterance. In particular, data normalization on the facial animation is vital to synthesize smoother lip movements.

3.2. Evaluations

3.2.1. Objective Evaluation

To evaluate the performance of the proposed approach, we conduct quantitative evaluation by calculating root mean square error (RMSE) between the predicted animations and ground-truth. The RMSEs on the entire animations and two crucial animations (jawOpen and mouthClose) are separately described. We compare our approach with 1) NVP [6], the state-of-the-art (SOTA) code-available approach in photo-realistic talking face generation task (using the implementation of Audio2ExpressionNet[4], which predicts facial expression in frame level) 2) BLSTM, which has been adopted by several S2A systems [2, 5, 10].

As described in Table 1, the RMSE of the entire animation decreases with the sliding window widening, which indicates that more contextual information does lead to more accurate facial movements. Though BLSTM-based model can

---

[1]https://thuhcsi.github.io/icassp2022-Transformer-S2A
[2]https://apps.apple.com/us/app/live-link-face/id1495370836
[3]https://www.unrealengine.com/marketplace/en-US/product/kite-demo
[4]https://github.com/keetsky/NeuralVoicePuppetry

Table 1. RMSE of different methods. The inputs are the same for all methods. The values are divided by 100 for simplicity. P: pitch. E: energy. winLen: the length of the window.

Method	Entire	Crucial
NVP winLen=3	4.048 ± 0.640	7.951 ± 1.242
NVP winLen=9	3.899 ± 0.397	8.121 ± 1.445
BLSTM	3.981 ± 0.510	8.185 ± 1.257
Proposed	$\mathbf{3.886 \pm 0.458}$	$\mathbf{7.378 \pm 1.231}$
Proposed w/o MOE	3.930 ± 0.387	7.709 ± 1.258
Proposed MFCC	4.240 ± 0.712	8.013 ± 1.565
Proposed w/o P&E	3.891 ± 0.424	7.490 ± 1.230

Fig. 3. Rendered images with animations predicted from NVP (left) and the proposed (right) after phoneme /i/.

capture long-term correlation in sequence level, the synthesis of each animation frame fails to attend the relevant speech frames while NVP imposes manual constraint about attentive region with the sliding window. The proposed model obtains lowest RMSE on both entire and crucial animations.

The ablation study is conducted to verify the effectiveness of the modules. Without MOE layer, RMSE on the crucial animations rises due to the poor model generalization. When using MFCC instead of PPGs as input, significant performance drop can be observed. Removing pitch and energy would also results in performance drop, which demonstrates the effectiveness of prosody features. This phenomenon tends to be more evident if the S2A model is trained on a more expressive dataset.

A detailed comparison is shown in Figure 3. For rendered image frames from NVP (winLen=9) and the proposed approach after speaking Mandarin phoneme /i/, the upper lip in NVP trembles towards the middle position, while the mouth transits to the next phoneme with more expressive and smooth movements in the proposed approach. Similar patterns can be widely observed in the evaluation.

3.2.2. Inference Efficiency

As shown in Table 2, we compare the average inference real time factor (RTF) of the proposed model with that of NVP [6] and BLSTM [2, 5, 10]. Each utterance for testing is around 12 seconds. The proposed model speeds up the animation synthesis by $17\times$, compared with BLSTM model. In regard to the frame-level model, our model achieves acceptable speed decline but obtains animations of higher quality.

Table 2. The comparison of inference RTF. The evaluation is conducted on a single NVIDIA P100 GPU. RTF denotes the average time to generate one-second animations.

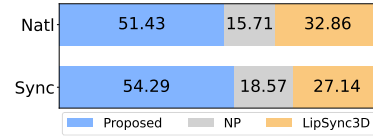| Method	RTF	Speedup
BLSTM	$0.09486 \pm 0.02077\text{s}$	$1.00 \times$
NVP winLen=9	$\mathbf{0.00049 \pm 0.00031s}$	$\mathbf{193.59 \times}$
Proposed	$0.00534 \pm 0.00219\text{s}$	$17.76 \times$

Fig. 4. ABX Results. NP: no preference. Sync: lip-speech synchronization. Natl: naturalness.

3.2.3. Subjective Evaluation

To further evaluate the perceptual quality of the proposed and the SOTA method when transferring across speakers and languages, ABX test is carried. Thirty-five judges with proficient English ability are invited to choose the better video in terms of lip-speech synchronization and naturalness. Judges are informed to focus on speech-to-animation performance, and to ignore the difference between characters. LipSync3D [13] (not open-source) is selected as the SOTA method to be compared, and its official demo video [5] is reused in the ABX test directly. Audio from the demo video is extracted as input to the proposed approach for driving a 3D avatar.

Evaluation result of ABX is presented in Figure 4. The proposed system has outperformed the baseline in both lip-speech synchronization and naturalness. Specially, the proposed S2A model is trained on the Mandarin dataset only and none English data is used, while LipSync3D is directly trained on English datasets.

4. CONCLUSION

In this paper, we propose a robust and efficient S2A system to generate high-quality facial animation from given speech. PPG and prosody features are introduced as inputs to ensure the cross-speaker and cross-language ability while maintain variants in speech. MOE-based Transformer is introduced to better model contextual information while provide significant optimization on computation efficiency. Experiments demonstrate the effectiveness of the proposed in both quality and efficiency evaluations.

[5]https://www.youtube.com/watch?v=L1StbX9OznY

5. REFERENCES

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

[2] Guanzhong Tian, Yi Yuan, and Yong Liu, "Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks," in *IEEE international conference on Multimedia Expo Workshops (ICMEW)*, 2019, pp. 366–371.

[3] Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu, "Photorealistic audio-driven video portraits," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, pp. 3457–3466, 2020.

[4] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black, "Capture, learning, and synthesis of 3d speaking styles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10093–10103.

[5] Huirong Huang, Zhiyong Wu, Shiyin Kang, Dongyang Dai, Jia Jia, Tianxiao Fu, Deyi Tuo, Guangzhi Lei, Peng Liu, Dan Su, Dong Yu, and Helen Meng, "Speaker independent and multilingual/mixlingual speech-driven talking head generation using phonetic posteriorgrams," *arXiv preprint arXiv:2006.11610*, 2020.

[6] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *Computer Vision – ECCV*, 2020, pp. 716–731.

[7] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, and Andrew Y.Ng, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[8] Wesley Mattheyses, Lukas Latacz, and Werner Verhelst, "Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis," *Speech Communication*, vol. 55, pp. 857–876, 2013.

[9] Constantinos Charalambous, Zerrin Yumak, and A. Frank van der Stappen, "Audio-driven emotional speech animation for interactive virtual characters," *Computer Animation and Virtual Worlds*, vol. 30, pp. e1892, 2019.

[10] Danny Websdale, Sarah Taylor, and Ben Milner, "Speaker-independent speech animation using perceptual loss functions and synthetic data," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[11] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2016, vol. II, pp. 1–6.

[12] M. Gadermayr, M. Tschuchnig, L. Gupta, N. Krämer, D. Truhn, D. Merhof, and B. Gess, "An asymmetric cycle-consistency loss for dealing with many-to-one mappings in image translation: A study on thigh mr scans," in *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 1182–1186.

[13] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler, "Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, vol. II, pp. 2755–2764.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[16] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[17] Jiaao He, Jiezhong Qiu, Aohan Zeng, Zhilin Yang, Jidong Zhai, and Jie Tang, "Fastmoe: A fast mixture-of-expert training system," *arXiv preprint arXiv:2103.13262*, 2021.

[18] Zhiyong Wu, Shen Zhang, Lianhong Cai, and Helen M Meng, "Real-time synthesis of chinese visual speech and facial expressions using mpeg-4 fap features in a three-dimensional avatar," in *Interspeech*, 2006, vol. 4, pp. 1802–1805.

[19] Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, et al., "Durian: Duration informed attention network for multimodal synthesis," *arXiv preprint arXiv:1909.01700*, 2019.