Interspeech 2024
1-5 September 2024, Kos, Greece

FastLips: an End-to-End Audiovisual Text-to-Speech System with Lip Features Prediction for Virtual Avatars

Martin Lenglet, Olivier Perrotin, Gérard Bailly

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, France
{martin.lenglet,olivier.perrotin,gerard.bailly}@grenoble-inp.fr

Abstract

In this paper, we introduce FastLips, an end-to-end neural model designed to generate speech and co-verbal facial movements from text, animating a virtual avatar. Based on the FastSpeech2 Text-to-Speech model, FastLips integrates an audiovisual Transformer-based encoder with distinct audio and visual neural decoders. This model combines audiovisual representations computed by the shared encoder with asynchronous generation of audio and visual features. Furthermore, we enhance the model with explicit predictors of lip aperture and spreading, adapted from prosodic FastSpeech2's variance adaptor. The proposed model generates mel-spectrograms and facial features (head, eyes, jaw and lip movements) to drive the virtual avatar's action units. In our evaluation, we compare FastLips with a baseline audiovisual-Tacotron2, demonstrating the advantages of the FastSpeech2 architecture for lip generation. This benefit becomes particularly prominent when implementing explicit lip prediction.

Index Terms: End-to-end TTS, audiovisual synthesis, facial animation, lip sync

1. Introduction

Virtual avatars are gaining popularity in various applications, such as video games, interactive chatbots, and immersive experiences in virtual or augmented reality. Embodied avatars enable multimodal interactions with computer systems, contributing to more engaging experiences that closely mirror natural human-to-human interactions [1, 2].

Expectations of users from these interactions are high, as their successful management can result in an increased sense of presence and competence of the virtual agent [1]. However, since speech and gesture originate from a common communication intent [3], even subtle discrepancies may be poorly rated by humans [4], complicating the audiovisual generation process. Among facial movements, the correct synchronization of jaw and lip movements (known as lip sync) is crucial in the successful fusion of modalities: inconsistent audio and lip movements have been showed to negatively impact intelligibility [5, 6].

To generate consistent audiovisual features, the dominant paradigm – illustrated by latest entries to the GENEA Challenge [7] – is a two steps approach: 1) a Text-To-Speech (TTS) model is used to generate the audio from the text input, and 2) a gesture model uses the synthesized audio (potentially augmented with the text) to generate animation features for the virtual avatar. However, this process induces sub-optimal computations [8] and negates the benefits of building unified audiovisual representations directly from text, in contrast with the conception of co-planned speech and gesture [3].

In this paper, we propose a unified end-to-end neural model to generate speech and facial gesture from text, which we refer to as an Audiovisual Text-To-Speech (AVTTS) model. The proposed AVTTS model, called FastLips, is based on FastSpeech2 [9], which is established as one of the main state-of-the-art neural TTS architectures. Through this contribution, we explore the potential of the FastSpeech2 architecture to generate audiovisual synthetic speech as an unified process. We evaluate FastLips in comparison with a baseline unified audiovisual-Tacotron2 model [10]. Through an ablation study, we highlight the main contributing components to the evaluated benefits of the proposed FastLips model. Related unified AVTTS models are discussed in Section 2. The audiovisual dataset as well as the avatar used in this study is detailed in Section 3. The proposed FastLips architecture is further described in Section 4. The results of the ablation study are stated in Section 5, and further discussed in Section 6.

2. Related Works

The latest propositions of unifying speech and gesture into the same generative models [8] have allowed the emergence of AVTTS models with remarkable performance [10, 11, 12]. Notably, DurIAN [11] and AVTacotron2 [10] have extended the TTS-architecture from Tacotron [13] and Tacotron2 [14], respectively, in order to generate visual features directly from audiovisual embeddings computed from text. These unified models have been showed to compete with the two-stages generative process [8] and even surpassed it in terms of naturalness of synthesized gestures [10, 12].

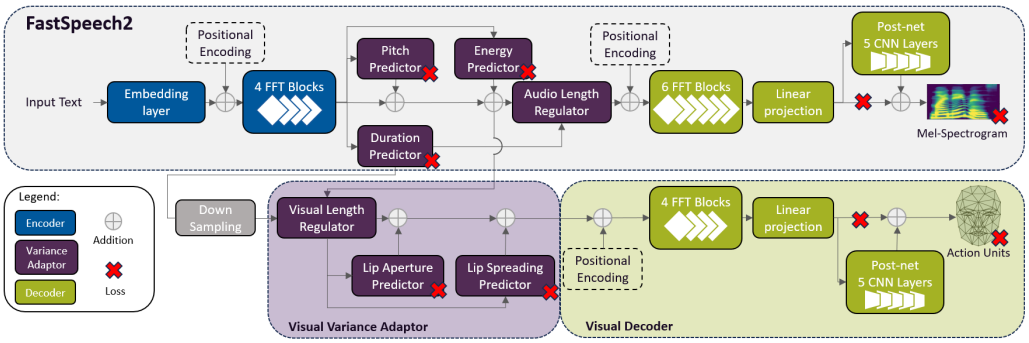
In both models, visual features are predicted from the very end of the autoregressive decoder. This setup ensures the computation of audiovisual embeddings in the whole model. Although we also believe that the computation of audiovisual embeddings should benefit to the consistency of generated synchronous features, the visual modality may exhibit asynchronous behaviors such as pre-phonatory gestures or more general anticipatory articulatory activations [15]. Therefore, we consider that an earlier distinction between the audio and visual decoders should benefit to both modalities. This made us consider the implementation of distinct audio and visual decoders from a shared audiovisual encoder, closer to the Tacotron2-Integrated Speech and Gesture (Tacotron2-ISG) [8]. However, we avoided the challenging training of the autoregressive Tacotron2 architecture by using the parallel Transformer-based FastSpeech2 [9] implementation instead.

3. Audiovisual Dataset

For this study, we recorded an audiovisual French dataset. This dataset was uttered by a French professional theater actress. Sentences are taken from the SIWIS database [16], which is composed of isolated extracts from French Novels and French parliament debates, spoken without any particular style instruc-

10.21437/Interspeech.2024-462

(a) *View of the virtual avatar with lip aperture (A) and lip spreading (S). The cropped section is used for the evaluation of lip sync.*



(b) *FastLips architecture.*

Figure 1: *Proposed model and preview of the virtual avatar.*

tions. This corpus contains 5.04 hours of audiovisual recordings, segmented into 6538 utterances (2.77 s±1.22 s). The audio is recorded at a 22 050 Hz sampling rate and 32-bit depth.

The face of the speakers was recorded with a Logitech StreamCam. The speaker's facial animation parameters are then tracked by an external service provider[1]. These facial animations – including head, eyes, jaw and lip movements – are used to emulate the virtual avatar by morphing the tracked facial features into deformations of the 3D model of the avatar. 152 elementary action units (AU) are computed to animate the avatar, sampled at 60 Hz. The avatar is illustrated in Figure 1a.

Since some of these control features co-vary in time, we used Principal Component Analysis (PCA) to compute a set of 37 facial features (FF) from the original 152 AU. These 37 FF are distributed among the main facial segments: 6 degrees-of-freedom for the head, 6 for the eyes, 3 for the eyelids, 4 for the eyebrows, 5 for the jaw, 10 for the lips and 3 for the nose. This dataset was recorded at GIPSA-lab, as part of the Theradia project [17]. Due to its intended use for commercial purposes, we are unable to share the complete dataset for public access or distribution. But two hours of this corpus were shared for the Spoke Task of the Blizzard Challenge 2023 [18].

4. Proposed Audiovisual TTS Model

This section describes the proposed FastLips architecture. Our implementation, the avatar player and pre-trained checkpoints are available online[2]. Hyperparameters used in this study are specified in the configuration files shared with the implementation.

4.1. FastLips Architecture

The proposed FastLips model is illustrated in Figure 1b. FastLips is a end-to-end neural model trained to predict mel-spectrograms and facial features from text and/or phones. Following the FastSpeech2 framework [9], FastLips adopts the encoder/decoder architecture, with both components being stacks of Feed-Forward Transformer (FFT) layers [19]. As opposed to LSTM units [20], FFT layers allow for the contextualization of the input sequence irrespective of the proximity of symbols in the sequence. Thus, FFT layers facilitate the learning of long-range dependencies [21]. Moreover, FFT layers allow for the parallel computing of the whole input sequence, reducing the training and inference time compared to recurrent models (inference speed reduced by a factor of 5, see Table 1).

The original single decoder of FastSpeech2 is duplicated to distinguish between one audio decoder and one visual decoder. In both cases, the output of the last FFT layer is projected into the dimension of the corresponding modality: 80 mel-spectrogram energies for the audio decoder and 37 FF for the visual decoder. The original Mean Absolute Error (MAE) is kept for the spectral prediction loss. A Tacotron2-like postnet is added after the mel-spectrogram prediction. This postnet models finer-grained temporal patterns through a stack of 5 convolutional layers. The spectral residual computed by the postnet is added to the prediction of the decoder. The postnet is trained with MAE spectral reconstruction loss after the addition of the residual. Similarly, the visual loss is computed before and after the visual postnet, also with MAE, and is added to the total loss during training. The benefit of the visual postnet is explored in the evaluation presented in Section 5.

Both decoders consume the contextualized sequence of input embeddings computed by the shared encoder. Note that the backpropagation is not stopped at the input of the visual decoder. Therefore the text encoder is constrained to produce audiovisual embeddings, as advocated by Wang et al.[8].

4.2. Visual Variance Adaptor and Lip Prediction

At the interface between the text encoder and the audio decoder, FastSpeech2 implements a variance adaptor which goal is twofolds: 1) two explicit predictors for pitch and energy are trained with a Mean Squared Error (MSE) loss function. In our implementation, pitch and energy prediction are trained at the phone-level. Values are first computed by frame [22] and averaged by phone following the phone-alignment. Pitch and energy values are normalized. 2) An explicit duration predictor is trained with MSE. The duration prediction is used at inference to produce the text-to-audio alignment through the length regulator [19]. Duration values are predicted as $log(1+\#frames)$. Ren et al. [9] reported better perceived voice quality thanks to these explicit prosodic predictions.

Similar to this audio variance adaptor, FastLips implements a visual variance adaptor, illustrated in Figure 1b. The visual variance adaptor implements the explicit prediction of two visual features: the lip aperture (A) and spreading (S) defined as the external lip height and width respectively (in millimeters), as illustrated in Figure 1a. Lips account for only 10 of the 37 FF of the avatar; the prediction of A and S is therefore an additional constraint to avoid conflicts between the visual and the audio modalities. A and S predictors are implemented after the length regulator in order to take into account the anticipatory lip movements [15]. The effect of the lip predictors on the visual loss and perceived audiovisual quality are evaluated in Section 5.

A and S predictors are trained with MSE losses, which are added to the total loss of the model. Similar to pitch and energy,

[1] DynamicXYZ© performed the tracking with the software Grabber.
[2] https://github.com/MartinLenglet/FastLips

each predictor computes a scalar value for A and S respectively, which is converted into an embedding that is added to the corresponding frame embeddings of the phone. Note that the aperture and spreading embeddings (resp. pitch and energy embeddings) are only added to the embeddings sequence of the visual decoder (resp. audio decoder). The losses are not weighted.

We did not implement a duration predictor in the visual variance adaptor. Training two duration predictors may cause global asynchronicity at inference between the audio and visual modalities. Instead, only the audio duration predictor is trained. The audio duration predictor is trained to predict the number of mel-spectrogram frames to produce from each symbol of the input sequence. On the one hand, the mel-spectrogram is predicted with a temporal sampling-frequency of $\sim$86 Hz[3]. On the other hand, FF are predicted with a sampling-frequency of 60 Hz. Thus, the number of frames predicted by the audio duration predictor is down-scaled by a factor $60 \div 86 \approx 0.7$.

4.3. Avatar Generation from Audiovisual Features

Mel-spectrograms and facial features are combined to drive the animation of the virtual avatar. The vocoder used is Waveglow [23]. The original architecture remains unchanged[4]. The pre-trained model shared with this implementation is fine-tuned on the French corpus shared for the Blizzard Challenge 2023 organizers [18] for 50 epochs. Facial features are reconstructed from the reduced set of 37 FF to the original 152 AU using the inverse PCA transformation.

5. Experiments

In this section, we evaluate the proposed FastLips model in comparison with an AVTTS baseline: AVTacotron2 [10]. The contribution of each specific layer is first evaluated through objective metrics to asses the minimization of the visual distortion between models. Best models according to objective metrics are selected for perceptual evaluations.

5.1. Baseline AVTacotron2 Implementation

To the best of our knowledge, Hussen et al. [10] have not shared their implementation of AVTacotron2. Therefore, we implemented our version based on their description. Our implementation is available online[5]. Since the presented experiment is focused on the audiovisual synthesis of neutral speech, we did not implement the emotion encoder. Our implementation consists in the original Tacotron2 architecture enhanced with a linear projection from the hidden states of the second LSTM layer of the autoregressive decoder to the 37 FF. Using a shared audiovisual decoder forces the same sampling rate between the two modalities. During training, FF are thus linearly extrapolated to match the audio sampling rate. At inference, predicted FF are interpolated at 60 Hz to compare all models at equivalent visual sampling rate.

Similar to [8, 10], visual features are not transmitted to the autoregressive process through the prenet, and no postnet is implemented for the visual features. Our AVTacotron2 uses the same hyperparameters as the original Tacotron2 [14].

5.2. Training Procedure

Five variants of the FastLips model are trained for this experiment: the complete architecture described in Section 4 (**Complete**), without lip aperture predictor (**-A**), without lip spreading predictor (**-S**), without lip predictors (**-A&S**) and without lip predictors and visual postnet (**-PN -A&S**). Technical specificities of the proposed FastLips variants, the baseline (**AVTacotron2**) and the Waveglow vocoder are given in Table 1.

All AVTTS models are trained on the same subset of the dataset presented in Section 3. 5% of this dataset (327 utterances) are randomly selected as the test set. Utterances are presented twice by epoch to the model: once with text inputs and once with phone inputs. This procedure enables the use of both text and phones at inference as showed by [24]. All models are trained from scratch for 40 000 iterations with a batch size of 32, which is equivalent to 100 epochs.

5.3. Objective Evaluation

We conducted an ablation study to evaluate the individual contribution of each component of the proposed FastLips model. The 327 utterances of the test set were synthesized with phone inputs to ensure correct pronunciation. Since synthesized stimuli may vary in duration compared to the original recordings, all objective metrics are computed on predicted stimuli aligned with the Ground-Truth (GT) via Dynamic Time Warping (DTW) [25]. All statistical test performed are pairwise Wilcoxon test. None of the evaluated models showed any statistical difference for the audio modality, so this evaluation focuses on the visual features. Figure 2 summarizes this evaluation.

5.3.1. Visual Distortion

The visual distortion between predicted and GT features was evaluated by root mean squared error (RMSE) on aligned FF sequences. Distributions of RMSE by utterance are showed in Figure 2a. These results validate that FastLips variants with at least one lip predictor (**-A**, **-S** and **Complete**) produce FF that are closer to GT than other models. However, this metric does not highlight benefits of the FastSpeech2 architecture in comparison with Tacotron2 for FF prediction.

Although visual distortion gives equal consideration to all degrees of freedom of the virtual avatar, not every feature carries the same weight in determining the animation quality. Specifically, the accurate modeling of lip sync is anticipated to have a more decisive impact on the perceived quality of the synthetic model. Consequently, we also conducted an assessment of the precision of lip sync.

5.3.2. Lip Sync

The lip sync is evaluated as the accuracy of models in predicting the aperture and spreading of the lips. Given that synthetic sequences are temporally aligned with GT, note that this metric evaluates lip sync based on the correct synchronization of jaw and lip movements' amplitude with the audio stimulus. Aperture and spreading errors are showed in Figures 2b-2c.

Both figures highlight the advantages of explicitly predicting lip aperture and spreading. Models equipped with the aperture predictor (**Complete** and **-S**) exhibit fewer aperture errors compared to all other models. Interestingly, **AVTacotron2**

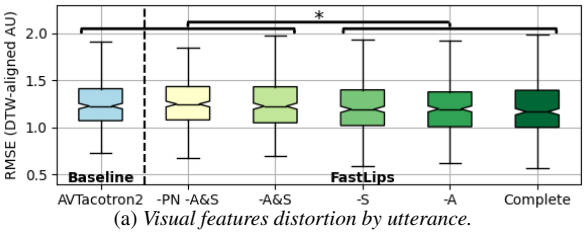
Table 1: *Technical specificities and performances of the models. Inference speed is reported as the Real-Time Factor (RTF). Performances are computed on a single GPU Quadro RTX 8000.*

Model	# Parameters	Inference Speed (RTF)
AVTacotron2	28 241 318	9.95×10^{-2}
FastLips -PN -A&S	47 683 942	1.89×10^{-2}
FastLips -A&S	51 811 797	2.11×10^{-2}
FastLips -S	52 272 597	2.17×10^{-2}
FastLips -A	52 272 597	2.21×10^{-2}
FastLips Complete	52 733 397	2.31×10^{-2}
Waveglow	87 879 272	5.31×10^{-2}

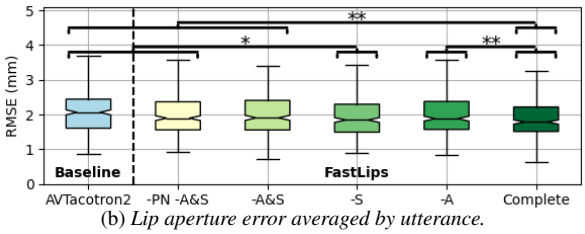
[3]Audio is recorded at 22 050 Hz and mel-spectrograms are computed with a hop length of 256.

[4]https://github.com/NVIDIA/waveglow

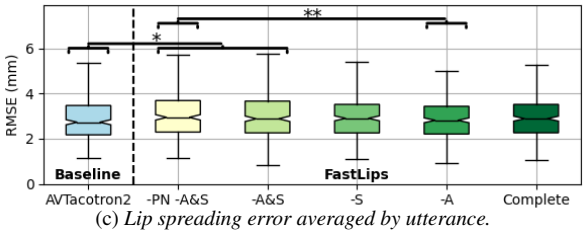
[5]https://github.com/MartinLenglet/AVTacotron2



(a) *Visual features distortion by utterance.*



(b) *Lip aperture error averaged by utterance.*



(c) *Lip spreading error averaged by utterance.*

Figure 2: *Objective evaluation of FF predicted by models. "*" vs. "**" indicates that the distribution of each model of a group differs statistically from all models of the other group via a pairwise Wilcoxon test ($p < .05$ and $p < .01$ respectively).*

predicts lip spreading relatively accurately. This can be attributed to the infrequent occurrence of lip protrusions with **AVTacotron2**, which tends to favor neutral positions, more commonly found in the corpus.

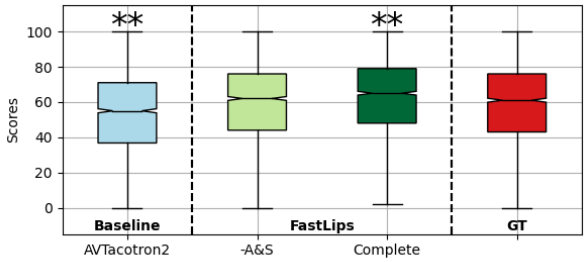
5.4. Perceptual Experiments

5.4.1. Perceptual Lip Sync

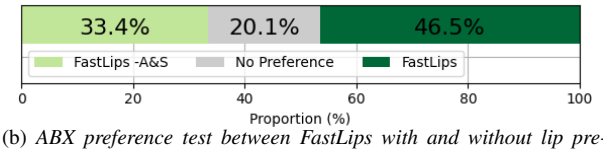
The objective evaluation highlighted the benefits of explicitly predicting lip for FastLips variants. Additionally, **AVTacotron2** exhibited better lip spreading prediction. Consequently, we conducted a perceptual test with three models to explore the impact of the observed differences on perceived lip animation quality: **AVTacotron2**, **-A&S**, and **Complete**.

We conducted an online MUSHRA-like experiment [26] with 50 participants, run with the HEMVIP framework [27]. Participants were asked to rate the lip animation quality of the avatar on a scale from 0 ("bad") to 100 ("excellent"). The reference (i.e. the GT animation features tracked from the original recordings) is hidden among the models, resulting in 4 conditions to evaluate. In all videos, only the GT audio is played to focus participants solely on the visual modality. Predicted FF are temporally aligned with GT. Head is fixed and only the lips of the avatar are displayed for this experiment, i.e. only the cropped section in Figure 1a is displayed. The 40 test stimuli for which the models predict the most different FF are selected for his experiment. Results of this experiment are given in Figure 3a.

On average, **AVTacotron2** was rated as "fair" (>40), whereas FastLips variants and GT were rated as "good" (>60). **Complete** was rated as producing significantly better lip animation quality than all other conditions, whereas **AVTacotron2** was rated significantly lower than other conditions. The outcomes of this evaluation is twofold: 1) Despite similar objective performances, **AVTacotron2** was judged as producing worse



(a) *MUSHRA results for lip animation quality. "**" indicates that the distribution statistically differs from all others according to a pairwise Wilcoxon test ($p < 0.01$).*



(b) *ABX preference test between FastLips with and without lip predictors. The effect of the model is validated by a chi-squared test ($p < 0.01$).*

Figure 3: *Perceptual Experiment Results.*

lip animation quality than **-A&S**. Although the difference was not significant, the higher errors of lip aperture on average may have impacted participants' ratings. 2) The explicit prediction of lip aperture and spreading improves lip animation quality. Note that this evaluation revealed tracking errors from GT in our recordings, leading to an overall lower animation quality than **Complete** on average. This score may be attributed to the stimuli selection method, which implicitly favored GT errors that artificially differ from FF predicted by other conditions.

5.4.2. Preference Test

In order to assess the potential of the explicit lip prediction for the audiovisual synthesis as a whole, we finally conducted an ABX preference test between the two best-performing models from the perceptual lip sync test: **-A&S** and **Complete**. 30 participants took part in this preference test. 80 utterances of the test set were selected for this test, also based on the maximum FF differences. For each utterance, the two models were presented as videos of the entire face of the avatar (with placement of models randomized). Results showed in Figure 3b indicate a statistical preference for FastLips with explicit lip predictors.

6. Conclusions and Discussion

In this work, we proposed FastLips, an end-to-end neural AVTTS model based on the FastSpeech2 architecture [9]. We showed that this model was able to generate better lip animation quality than the baseline AVTacotron2 [10]. This model emphasizes the advantages of an early distinction between the audio and visual modalities, enabling more effective support for asynchronous behaviors. This feature is crucial to take into account the anticipatory lip movements [15]. The preference test confirmed the objective results regarding the benefits of explicit lip aperture and spreading prediction in enhancing the overall quality of the audiovisual synthesis for virtual avatars. Samples are available on our online demo page[6].

We will consider applying the FastLips architecture to expressive audiovisual synthesis in the future. Additionally, we may explore the integration of other visual features, such as eye blinks and head nods, as explicit predictions in the proposed visual variance adaptor.

[6] `http://ssw2023.org/demo/FastLips/index.html`

7. Acknowledgements

This research has received funding from the BPI project THERADIA and MIAI@Grenoble-Alpes (ANR-19-P3IA-0003). This work was granted access to HPC/IDRIS under the allocation 2023-AD011011542R2 made by GENCI.

8. References

[1] D. Potdevin, "Vers des agents conversationnels animés sociaux: Quelle influence de l'intimité virtuelle sur l'expérience utilisateur et la relation-client?" Ph.D. dissertation, Université Paris-Saclay, 2020.

[2] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, and M. Neff, "A comprehensive review of data-driven co-speech gesture generation," in *Computer Graphics Forum*, vol. 42, no. 2. Wiley Online Library, 2023, pp. 569–596.

[3] D. McNeill, *Gesture and thought*. University of Chicago press, 2019.

[4] L. D. Rosenblum, "Primacy of multimodal speech perception," *The handbook of speech perception*, pp. 51–78, 2005.

[5] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[6] V. Van Wassenhove, K. W. Grant, and D. Poeppel, "Temporal window of integration in auditory-visual speech perception," *Neuropsychologia*, vol. 45, no. 3, pp. 598–607, 2007.

[7] T. Kucherenko, R. Nagy, Y. Yoon, J. Woo, T. Nikolov, M. Tsakov, and G. E. Henter, "The genea challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings," in *Proceedings of the 25th International Conference on Multimodal Interaction*, 2023, pp. 792–801.

[8] S. Wang, S. Alexanderson, J. Gustafson, J. Beskow, G. E. Henter, and É. Székely, "Integrated speech and gesture synthesis," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 177–185.

[9] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

[10] A. Hussen Abdelaziz, A. P. Kumar, C. Seivwright, G. Fanelli, J. Binder, Y. Stylianou, and S. Kajareker, "Audiovisual speech synthesis using Tacotron 2," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 503–511.

[11] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei, D. Su, and D. Yu, "DurIAN: Duration Informed Attention Network for Speech Synthesis," in *Proc. Interspeech 2020*, 2020, pp. 2027–2031.

[12] S. Mehta, S. Wang, S. Alexanderson, J. Beskow, E. Szekely, and G. E. Henter, "Diff-TTSG: Denoising probabilistic integrated speech and gesture synthesis," in *Proc. 12th ISCA Speech Synthesis Workshop (SSW2023)*, 2023, pp. 150–156.

[13] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of Interspeech*, Stockholm, Sweden, 8 2017, pp. 4006–4010. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1452

[14] 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 *ICASSP*. IEEE, 2018, pp. 4779–4783.

[15] J. X. Maier, M. Di Luca, and U. Noppeney, "Audiovisual asynchrony detection in human speech." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 1, p. 245, 2011.

[16] P.-E. Honnet, A. Lazaridis, P. N. Garner, and J. Yamagishi, "The SIWIS French speech synthesis database – Design and recording of a high quality french database for speech synthesis," Idiap, Tech. Rep., 2017.

[17] F. Tarpin-Bernard, J. Fruitet, J.-P. Vigne, P. Constant, H. Chainay, O. Koenig, F. Ringeval, B. Bouchot, G. Bailly, F. Portet, S. Alisamir, Y. Zhou, J. Serre, V. Delerue, H. Fournier, K. Berenger, I. Zsoldos, O. Perrotin, F. Elisei, M. Lenglet, C. Puaux, L. Pacheco, M. Fouillen, and D. Ghenassia, "Theradia: Digital therapies augmented by artificial intelligence," in *International Conference on Applied Human Factors and Ergonomics*. Springer, 2021, pp. 478–485.

[18] O. Perrotin, B. Stpehenson, S. Gerber, and G. Bailly, "The Blizzard Challenge 2023," in *Proc. 18th Blizzard Challenge Workshop*, 2023, pp. 1–27.

[19] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems*, 2019, pp. 3171–3180.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] 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.

[22] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Audio Engineering Society, 2009.

[23] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP*. IEEE, 2019, pp. 3617–3621.

[24] M. Lenglet, O. Perrotin, and G. Bailly, "The GIPSA-lab Text-To-Speech system for the Blizzard Challenge 2023," in *Proc. 18th Blizzard Challenge Workshop*. ISCA, 2023, pp. 34–39.

[25] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1. IEEE, 1993, pp. 125–128.

[26] ITU, "1534-1, method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunications Union, Geneva, Switzerland*, vol. 14, 2003.

[27] P. Jonell, Y. Yoon, P. Wolfert, T. Kucherenko, and G. E. Henter, "Hemvip: Human evaluation of multiple videos in parallel," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, ser. ICMI '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 707–711. [Online]. Available: https://doi.org/10.1145/3462244.3479957