
Proceedings Paper:
Que, S. and Ragni, A. orcid.org/0000-0003-0634-4456 (2025) VisualSpeech: Enhancing Prosody Modeling in TTS Using Video. In: Scharenborg, O., Oertel, C. and Truong, K., (eds.) Proceedings of Interspeech 2025. Interspeech 2025, 17-21 Aug 2025, Rotterdam, The Netherlands. International Speech Communication Association (ISCA), pp. 3778-3782. ISSN: 2958-1796. EISSN: 2958-1796.

https://doi.org/10.21437/Interspeech.2025-1494

Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

VisualSpeech: Enhancing Prosody Modeling in TTS Using Video

Shumin Que, Anton Ragni

Department of Computer Science, The University of Sheffield, United Kingdom
{sque1, a.ragni}@sheffield.ac.uk

Abstract

Text-to-Speech (TTS) synthesis faces the inherent challenge of producing multiple speech outputs with varying prosody given a single text input. While previous research has addressed this by predicting prosodic information from both text and speech, additional contextual information, such as video, remains under-utilized despite being available in many applications. This paper investigates the potential of integrating visual context to enhance prosody prediction. We propose a novel model, VisualSpeech, which incorporates visual and textual information for improving prosody generation in TTS. Empirical results indicate that incorporating visual features improves prosodic modeling, enhancing the expressiveness of the synthesized speech. Audio samples are available at https://ariameetgit.github.io/VISUALSPEECH-SAMPLES/.
Index Terms: Text-to-speech Synthesis, Video, Visual Features, Prosody

1. Introduction

With advancements in deep learning [1, 2, 3], text-to-speech (TTS) [4, 5] models have become capable of generating speech that closely mimics human speech. Despite the high-quality output, the prosody of the generated speech sometimes becomes inappropriate for the context. For example, most TTS approaches would fail to convey a different degree of happiness within the same sentence given a different context. This limitation arises because the same text input can produce multiple speech outputs with varying prosodic patterns. This is a well-known challenge in speech synthesis, known as the one-to-many mapping problem [6].

Previous research has attempted to mitigate the one-to-many problem by incorporating additional inputs, including both textual and speech-related features. Commonly used text features encompass linguistic elements (e.g. part-of-speech, word boundaries, and word position) as well as syntactic structures and semantic meanings. In addition, some models incorporate an extra module to extract latent features, including pitch, energy, and duration, from reference speech [7, 8, 9], capturing both global and local latent factors. Although these studies have enhanced prosodic modelling in TTS, there is still much work to be done.

In addition to text and speech, many common application scenarios, such as video game characters navigating dungeons or engaging in battles, feature visual information, which could play a crucial role in accurately predicting prosody. For instance, the speech patterns of these video game characters are unlikely to resemble those of read or acted speech due to the dynamic nature of visual contexts. There have so far been limited research on the usefulness of visual information in TTS. For instance, [10] successfully generated speech samples with accurate reverberation effects in specific environments by learning physical environment information from video data. However, no previous studies have explored the impact of visual information on prosody in TTS.

Thus, this paper explores the possibility of improving prosody prediction in TTS with visual information (video). It makes three key contributions. First, it demonstrates visual cues carry valuable prosodic information. Second, this visual information complements existing textual features. Finally, it reveals that the integration of both textual and visual inputs leads to substantial improvements in prosody prediction in TTS models.

2. The Proposed Method

2.1. Overview

This paper first verifies that visual information is predictive of prosody. It then demonstrates that the prosodic information derived from visual features complements that extracted from text. Lastly, the visual features are integrated into a modern TTS model, as shown in Figure 1. The architecture of the proposed model is based on FastSpeech2 [7] due to its explicit prosody modeling and widespread adoption in the field.

As shown in Figure 1(a), the VisualSpeech model mainly consists of four standard FastSpeech2 modules (text encoder, variance adaptor, Mel-spectrogram decoder, and pre-trained vocoder) and two new modules (visual encoder and visual-text fusion). Please refer to [7] for details on the former modules.

2.2. The Visual Encoder

The visual encoder shares a similar structure with the text encoder in FastSpeech2. Visual features, which refer to the characteristics extracted from video frames (such as facial expressions, body gestures, speaking environment, and other relevant visual cues), are represented as $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\}$. These extracted features $\mathbf{F}$ are then passed into the visual encoder to generate the visual hidden sequence $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$.

2.3. The Visual-Text Fusion

Since the sequence lengths of the text encoder output $\mathbf{T}$ and the visual encoder output $\mathbf{V}$ are different, a simple addition or concatenation approach is infeasible in this case. Simple methods like average pooling can be adopted to compress $\mathbf{V}$ into one global visual vector which can be added to the text encoder output $\mathbf{T}$. However, local information might be lost if global visual vector is utilized due to the possibility of information over-compression. Thus, a more dedicated module (*i.e.* the visual-text fusion module) is proposed to better summarize the visual information and to shed light on how different parts of a video

10.21437/Interspeech.2025-1494

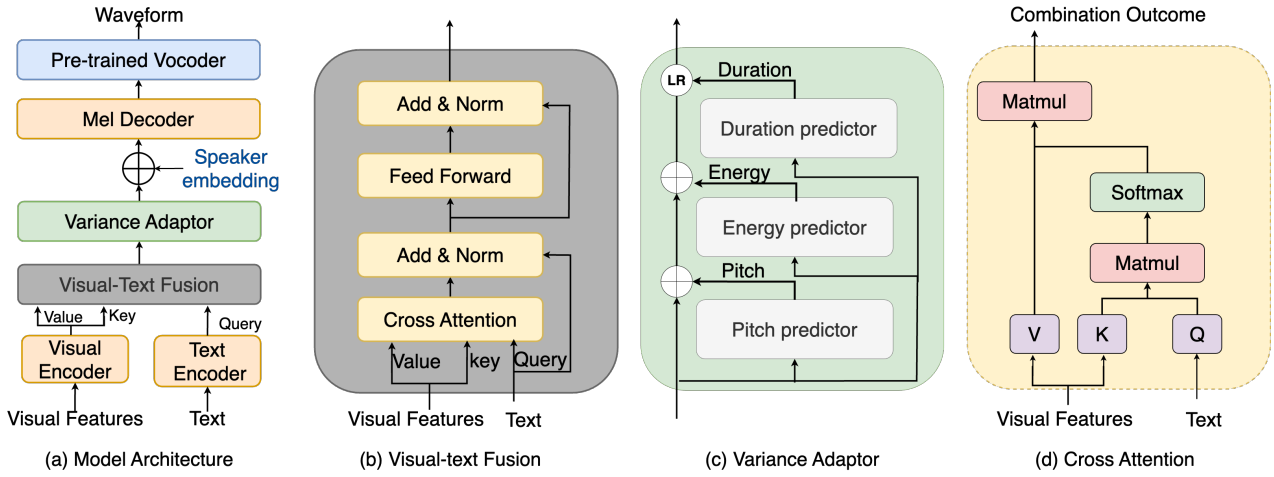

Figure 1: *The proposed multi-modal TTS model: VisualSpeech.*

are correlated to the text. The proposed visual-text fusion module is illustrated in Figure 1(b). It is designed to capture and leverage the complex interactions between textual and visual modalities. At its core, this module employs cross-attention that enables bidirectional information flow between the two modalities. This design choice is motivated by the observation that prosodic patterns often correlate with both linguistic content and visual context cues. The fusion architecture consists of two key components: (1) a cross-attention

$$\delta(\mathbf{V}, \mathbf{T}) = \text{Softmax}\left(\frac{\mathbf{T}\mathbf{V}^T}{\sqrt{d_{\mathbf{V}}}}\right)\mathbf{V} \tag{1}$$

which is illustrated in Figure 1(d), and (2) a feedforward network for feature refinement. In the cross-attention layer, text features are used as queries to attend to visual features, enabling the model to learn which visual cues are most relevant to specific textual elements and how these cues contribute to prosodic variation. Using the output of the text encoder $\mathbf{T}$ as the query can ensure that the output length of the visual-text fusion module is the same as the length of the phone embedding sequence, thus the output sequence of the module could be used to predict the pitch, energy, and duration at the phone level. $\mathbf{T}$ and $\mathbf{V}$ undergo a matrix-matrix product operation to obtain attention scores, which are divided by the square root of $d_{\mathbf{V}}$ (*i.e.* dimensionality of $\mathbf{V}$) and then passed through a Softmax function to get the normalized scores/weights. These normalized attention scores are used to obtain a weighted sum of $\mathbf{V}$.

3. Experiments

3.1. Experimental Setup

To assess the impact of visual information on speech generation performance, a dataset containing both diverse prosodic and visual data is essential. However, research exploring the role of visual features in speech synthesis from a prosody perspective is still scarce, and existing datasets are limited in scope. While well-known speech synthesis datasets, such as LJSpeech [11] and LibriTTS [12], provide rich textual data, they lack corresponding visual features. On the other hand, datasets that do include visual information, like TED and Ego4D [13], suffer from limitations in prosodic or visual diversity. To address this gap, this paper utilizes the Condensed Movies Dataset (CMD) [14],

Table 1: *The objective evaluations of the original speech (i.e. Raw), speech after vocal extraction (i.e. VE), and speech after vocal extraction and speech denoising and enhancement (i.e. VE + SDE). ↑ denotes the higher the better.*

	STOI↑	**PESQ ↑**	**SI-SNR ↑**
Raw	0.78	1.55	5.05
VE	0.84	1.85	9.06
VE+SDE	0.95	2.90	16.49

an open-sourced, large-scale video dataset in English. CMD consists of short video clips extracted from over 3,605 films, spanning various genres, countries, and decades, making it a valuable resource for studying the integration of visual features in speech synthesis. These videos capture high-quality scenarios such as conversations at funerals and weddings, providing a rich array of prosodic and visual variations.

The original duration of the video clips ranges from 10 to 300 seconds. Each clip was split into sentence segments by means of Whisper [15] to facilitate model training [16]. A subset of the resulting dataset was randomly selected for the experiments in this study, named CMD2, comprising approximately 33 hours of video and speech [16]. This subset includes 70,599 video clips ranging from 2 to 30 seconds along with corresponding speech and text transcriptions.

Compared to standard TTS datasets, CMD2 includes various types of background noise, music, and even some singing data. To prepare the CMD2 dataset for TTS training, this work applies vocal extraction techniques to separate music from speech, employs speech denoising and enhancement methods to reduce background noise, and implements filtering criteria to remove singing samples.

This work utilizes MVSEP [17] for vocal separation and Resemble-Enhance [18] for speech denoising and enhancement. Objective evaluations were conducted to assess the impact of vocal extraction, speech denoising, and enhancement on the speech quality. Following the methodology in [19], the evaluation metrics include Short-Time Objective Intelligibility (STOI) [20], Wideband Perceptual Evaluation of Speech Quality (PESQ) [21], and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [22]. As reference (clean) speech is unavail-

Table 2: *MSE Loss of FNN-ResNet50 and FNN-Omnivore Models on Test Datasets.*

Models	Pitch ↓	Energy ↓
Mean-predictor	23.72	28.97
FNN-ResNet50	0.93	1.20
FNN-Omnivore	0.84	1.11

able, reference-free scores for these metrics were estimated using the method in [23]. A random selection of 1,000 samples from the dataset was used to compute these metrics. The results, presented in Table 1, clearly demonstrate that, compared to the raw dataset, vocal extraction yields relative improvements of 7.7%, 19.4%, and 79.4%, while speech denoising and enhancement provide even more substantial improvements, highlighting the effectiveness of both techniques.

The following filtering criteria are applied to remove singing clips from the dataset. First, speech samples with pitch values exceeding a predefined threshold (i.e., 500 Hz) are discarded. Second, we compute the mean and standard deviation of pitch, energy, and duration for each phoneme. Any speech sample with pitch or duration values that fall outside 2.5 times the standard deviation of these statistics is considered an outlier and is removed. After applying these filters, the resulting dataset(44,665 training, 2,000 validation, and 200 test samples) is prepared for training the TTS model.

Following the common practice in [4], we resampled all speech into 22.05k sample rate, and converted it into 80-dimensional Mel-spectrograms. We used the Montreal Forced Aligner (MFA) to perform forced alignment and extract phoneme duration.

3.2. Preliminary Study I: Prosody Clues in Visual Features

To verify the hypothesis of visual features in prosody modeling, a preliminary experiment is conducted using a FeedForward Neural Network (FFNN) to predict prosodic features (pitch and energy), solely from visual features.

Given a sequence of visual features $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ and a sequence of prosodic features $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_t\}$, where $t \gg n$, we average every t/n prosody features to obtain a target prosody feature corresponding to each visual feature. As a result, the final prosody feature sequence $\mathbf{P_V} = \{\mathbf{p}_{\mathbf{v}_1}, \mathbf{p}_{\mathbf{v}_2}, \ldots, \mathbf{p}_{\mathbf{v}_n}\}$ aligns with the visual feature sequence. In this setup, the visual feature sequence $\mathbf{V}$ is used as the input to the model, and the prosody sequence $\mathbf{P_V}$ serves as the target.

In this experiment, two types of visual features were investigated: one extracted from the Omnivore model [24][1] and the other from the ResNet50 model [25].[2] The FFNN models consist of two hidden layers, each with 256 dimensions and a dropout rate of 0.5. Each model was trained for 400,000 steps with a batch size of 32. The loss function used is mean squared error (MSE). The Adam optimizer [26] was employed for parameter updates, with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a learning rate of 1×10^{-5}.

The results are presented in Table 2. For reference, we also include the results from a mean predictor, where the predicted value is simply the mean of the ground truth sequence, used to compute the MSE loss. The MSE loss for the model using visual features is significantly smaller than that of the mean predictor, clearly demonstrating that visual features are predictive of prosody. Moreover, the MSE loss obtained using visual features from the Omnivore model is slightly smaller than that from ResNet50, indicating that Omnivore extracts more relevant information for prosody prediction than ResNet50.

Table 3: *MSE Loss of three models (Text-based, Text + Omnivore Visual Features, and Text + ResNet50 Visual Features) on test datasets.*

Models	Pitch ↓	Energy ↓	Duration ↓
Text	0.43	0.50	0.35
Text+VF-Omnivore	0.39	0.59	0.34
Text+VF-ResNet50	0.40	0.58	0.35

3.3. Preliminary Study II: Visual Features as a Complement to Textual Information

Having confirmed visual information can predict prosody, this section investigates whether visual features complement textual features. To this end, we extend the pitch, energy, and duration (PED) predictor in FastSpeech2 [7] to incorporate both textual and visual information. We then compare performance of the model using both types of features with that of the model using only textual features. This comparison allows us to assess the added value of visual features in enhancing prosody prediction.

The text-only PED predictor, following the methodology outlined in [7], relies solely on textual input for prosody prediction. In contrast, the visual-textual PED predictor integrates both textual and visual features to predict prosody. The prosodic features—pitch, energy, and duration (P/E/D)—are extracted from the ground-truth (original) audio and serve as the target for both predictors. The visual encoder shares the same architecture as the text encoder. The cross-attention module in the visual-text fusion model has a hidden size of 256, with two attention heads and a dropout rate of 0.2. The training setup follows that of [7], with the exception of a batch size of 32 and a total of 400,000 training steps. The Adam optimizer [26] is used for parameter updates, with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a learning rate scheduler, as specified in [7].

The results presented in Table 3 show that the inclusion of visual features improves pitch prediction slightly, with the Text + VF-Omnivore model achieving the lowest MSE loss (0.39). In terms of duration prediction, the Text + VF-Omnivore model achieves a relative 3% improvement compared with the text-only model, and the Text + VF-ResNet50 model gains no improvement. However, for energy prediction, the addition of visual features has a negative impact on performance.

3.4. TTS Experiments

Previous results demonstrate that the combination of visual and textual information outperforms text-only prosody prediction, including pitch and duration. To verify whether these improvement in prosody feature prediction will lead to better speech generation, visual features are integrated into the FastSpeech2 [7] model,[3] incorporating a visual encoder and visual-text fusion, to synthesize speech. Two VisualSpeech models were trained, one with Omnivore and another with ResNet50 visual features. Each model is trained for 400,000 steps with a batch

[1] https://github.com/facebookresearch/omnivore

[2] https://github.com/v-iashin/video_features

[3] Experiments are implemented based on the open-sourced repository: https://github.com/ming024/FastSpeech2

Table 4: *MSE loss of three Models (FastSpeech2, VisualSpeech with Omnivore Features, and VisualSpeech with ResNet50 Features) on Test Datasets*

Models	**Pitch** ↓	**Energy** ↓	**Duration** ↓
FastSpeech2	0.27	0.39	0.41
VisualSpeech (Omnivore)	0.18	0.37	0.21
VisualSpeech (ResNet50)	0.18	0.34	0.24

Table 5: *The Mel-cepstral Distance (MCD) (in dB) and Log F0 RMSE Loss of Three Models*

Models	**MCD**↓	**Log F0**↓
FastSpeech2	7.64	0.46
VisualSpeech (Omnivore)	7.38	0.44
VisualSpeech (ResNet50)	7.34	0.45
FastSpeech2 (+ GT PED)	5.04	0.32
VisualSpeech (Omnivore) (+ GT PED)	4.83	0.30
VisualSpeech (ResNet50) (+ GT PED)	4.82	0.31

size of 32. The Adam optimizer is used for parameter updates [26], with hyper-parameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, and the same learning rate scheduler as [7].

As shown in Table 4, the proposed models significantly outperform the baseline FastSpeech2 model on all three metrics. Specifically, the proposed model achieves about 33%, 5%, and 49% relative improvement over the baseline model in pitch, energy, and duration prediction, respectively. We find that the improvement in the energy prediction is relatively less significant, which may be attributed to energy being less predictable and trivial. Likewise, FastPitch [27] finds removing energy prediction while maintaining pitch and duration prediction could generate high-quality speech.

Table 5 demonstrates that the proposed model significantly outperforms the baseline model in terms of Mel-Cepstral Distortion (MCD) and UTMOS [4] and offers slight improvement in Log F0. Similar trends are observed when using ground truth (GT) PED, suggesting that visual features also influence mel-spectrogram prediction beyond PED. These findings indicate that the current approach extracts some useful prosodic information from video, but there is still significant room for improvement. The remaining gap in MCD suggests that further refinement in the modeling of visual features or a more effective utilization of the available data could enhance their contribution to prosody modeling, extending beyond PED prediction alone.

3.4.1. Case Study: Prosody Visualization

To better understand the differences between models trained in the TTS experiments and evaluate their performance on test data, we used these three well-trained models to generate speech given the same text input and visualize them.

In Figure 2, we visualize the predicted pitch contours from three models and compare them with the ground truth. As shown, the pitch contours generated by the two models incorporating additive visual features are closer to those produced by the text-only model. A similar trend is observed in energy

[4]UTMOS `https://github.com/sarulab-speech/UTMOS22` results for these models (i.e. 2.91, 3.06 and 3.13, respectively) suggest that these improvements have led to an enhanced perceptual quality.

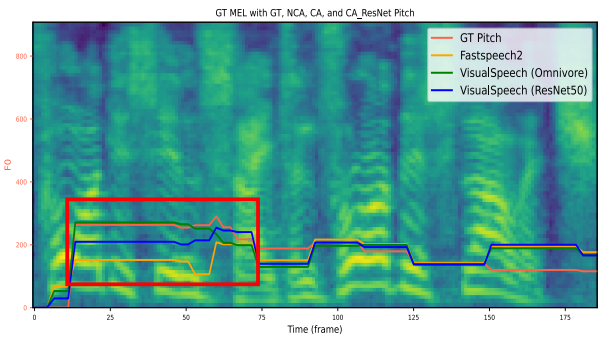

Figure 2: *The Pitch Contour*

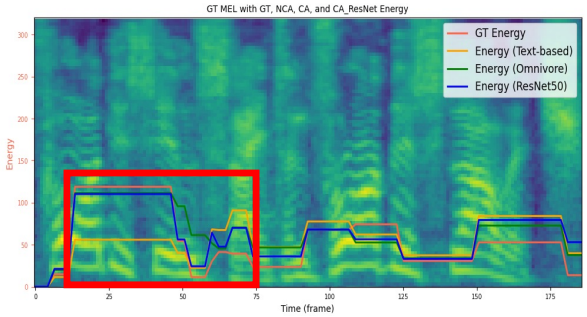

Figure 3: *The Energy Contour*

prediction, as illustrated in Figure 3.

These findings align with other studies that have demonstrated the use of visual features to improve speech across various dimensions [28, 10, 29, 30], further confirming the effectiveness of visual features in speech synthesis. Additionally, the improvement in prosodic prediction through visual features is consistent with the discussion in [31], which suggested that contextual information enhances prosody in generated speech. As previously mentioned, visual features provide additional context, such as the speaking environment, which plays a critical role in more accurately predicting prosody.

4. Conclusion

In this work, we introduce VisualSpeech, the first visual text-to-speech synthesis model that incorporates visual features from corresponding videos to complement text features, significantly enhancing the prosody of the generated speech. Using two distinct video feature extractors, we demonstrate that these visual features encapsulate prosodic information. By integrating these features with text, our results show that visual information complements textual information, leading to improved prosody prediction. This study lays a foundation for future research on leveraging visual information to improve prosody performance in TTS. Additionally, the CMD2 dataset created in this work provides a valuable resource for future studies.

While this study has shown promising results, there are some limitations. Despite employing speech denoising and enhancement, the quality of the CMD dataset still impacts the speech generation quality. Therefore, a high-quality dataset is essential for future research.

5. References

[1] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3844–3848.

[2] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4470–4474.

[3] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He *et al.*, "NaturalSpeech: End-to-end text-to-speech synthesis with human-level quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[4] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.

[5] R. Huang, M. W. Lam, J. Wang, D. Su, D. Yu, Y. Ren, and Z. Zhao, "FastDiff: A fast conditional diffusion model for high-quality speech synthesis," *arXiv preprint arXiv:2204.09934*, 2022.

[6] M. Babiański, K. Pokora, R. Shah, R. Sienkiewicz, D. Korzekwa, and V. Klimkov, "On granularity of prosodic representations in expressive text-to-speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 892–899.

[7] 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," *arXiv preprint arXiv:2006.04558*, 2020.

[8] T. Kenter, V. Wan, C.-A. Chan, R. Clark, and J. Vit, "Chive: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3331–3340.

[9] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6189–6193.

[10] H. Liu, R. Huang, X. Lin, W. Xu, M. Zheng, H. Chen, J. He, and Z. Zhao, "VIT-TTS: visual text-to-speech with scalable diffusion transformer," *arXiv preprint arXiv:2305.12708*, 2023.

[11] K. Ito and L. Johnson, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[12] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[13] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 995–19 012.

[14] M. Bain, A. Nagrani, A. Brown, and A. Zisserman, "Condensed movies: Story based retrieval with contextual embeddings," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[16] J. Sanders, "CMD+: A D.I.Y. Audiovisual Dataset for Multi-Speaker TTS," 2023, unpublished undergraduate thesis, University of Sheffield.

[17] G. Fabbro, S. Uhlich, C.-H. Lai, W. Choi, M. Martínez-Ramírez, W. Liao, I. Gadelha, G. Ramos, E. Hsu, H. Rodrigues *et al.*, "The sound demixing challenge 2023 music demixing track," *arXiv preprint arXiv:2308.06979*, 2023.

[18] R. AI, "Resemblyzer," https://github.com/resemble-ai/Resemblyzer, 2024, accessed: 2024-08-22.

[19] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan *et al.*, "Audiobox: Unified audio generation with natural language prompts," *arXiv preprint arXiv:2312.15821*, 2023.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4214–4217.

[21] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[22] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[23] A. Kumar, K. Tan, Z. Ni, P. Manocha, X. Zhang, E. Henderson, and B. Xu, "Torchaudio-Squim: Reference-less speech quality and intelligibility measures in torchaudio," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[24] R. Girdhar, M. Singh, N. Ravi, L. Van Der Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 102–16 112.

[25] 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.

[26] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592.

[28] Z. Xie, S. Yu, Q. He, and M. Li, "Sonicvisionlm: Playing sound with vision language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 866–26 875.

[29] R. Mira, K. Vougioukas, P. Ma, S. Petridis, B. W. Schuller, and M. Pantic, "End-to-end video-to-speech synthesis using generative adversarial networks," *IEEE Transactions on Cybernetics*, vol. 53, no. 6, pp. 3454–3466, 2022.

[30] J. Choi, M. Kim, and Y. M. Ro, "Intelligible lip-to-speech synthesis with speech units," *arXiv preprint arXiv:2305.19603*, 2023.

[31] Z. Hodari, A. Moinet, S. Karlapati, J. Lorenzo-Trueba, T. Merritt, A. Joly, A. Abbas, P. Karanasou, and T. Drugman, "Camp: a two-stage approach to modelling prosody in context," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6578–6582.