

ArtSpeech: Adaptive Text-to-Speech Synthesis with Articulatory Representations

Anonymous Author(s)

Submission Id: 2547

ABSTRACT

In this supplementary, we provide: (1) Comparative Mean Opinion Scores for Overall Similarity (CMOS-O) to facilitate a detailed analysis of the relative differences between models; (2) linear fusion of style vectors between two different speakers and visualization of the synthesized results; (3) detailed information on ablation experiments to demonstrate the effectiveness of *ArtSpeech* components; (4) detailed information about the subjective evaluation process.

A SUPPLEMENTARY EXPERIMENTS

A.1 CMOS-O

Table 1: The comparison Mean Opinion Score of Overall Similarity (CMOS-O) with P-values from the Wilcoxon test relative to other models.

Model	CMOS-O (p-value)
Ground Truth	+0.01 (0.769)
YourTTS + HiFi-GAN	-0.12 (0.002)
VALL-E X	-0.10 (0.009)
StyleTTS + HiFi-GAN	-0.08 (0.033)
StyleTTS 2	-0.05 (0.082)
ArtSpeech + HiFi-GAN	0.00

We utilized the Comparison Mean Opinion Score of Overall Similarity (CMOS-O) for a more nuanced analysis of the relative differences between models. we applied the Wilcoxon test to the CMOS-O to detect significant differences in the synthesis results.

The results show that the similarity of *ArtSpeech* synthesis results to the reference speech exceeds that of all baseline models. The p-value with all other baseline methods except StyleTTS2 is less than 0.05, indicating that the similarity of *ArtSpeech* synthesized results with the reference speech is significantly higher than these methods.¹

¹Note: In the *LibriTTS-test-clean* dataset, stylistic variations within speeches by the same speaker caused inconsistencies between the target and GT audio. We deleted the 10 speeches where GT scored low in MOS-O. Without this exclusion, *ArtSpeech* exceeded GT in both MOS-O and CMOS-O. The GT's similarity score in CMOS-O was -0.03, lower than *ArtSpeech*'s.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM MM, 2024, Melbourne, Australia

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

A.2 Style Control

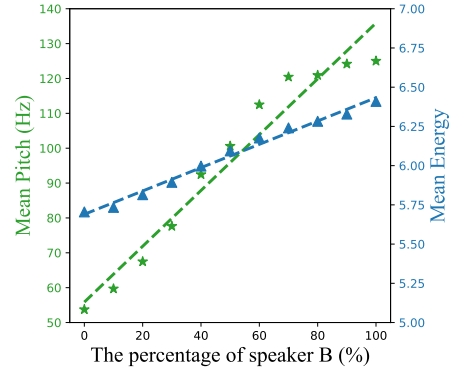


Figure 1: Utilize the style vectors of speaker A and speaker B from Section 6, applying different proportional weightings to their summation. Visualize the mean energy(blue line) and the mean pitch(green line) of the generated mel spectrograms as the proportion of speaker B shifts from 0% to 100%.

In Figure 1, we visualize the mean energy and the mean pitch of the generated mel spectrograms as the proportion of speaker B's style vector varies from 0% to 100%. It is observed that with an increasing proportion of speaker B's style vector, there is a linear rise in both the mean energy and mean pitch of the synthesized speech. The smooth transition during the interpolation of style vectors in the latent space amply demonstrates that the latent space is sufficiently disentangled. Even without imposing additional constraints during the training process, the style vectors of *ArtSpeech* demonstrated remarkable linear separability in the latent space.

A.3 Ablation Study

We conducted ablation experiments to validate the efficacy of *ArtSpeech*. These experiments focused on several critical aspects: the selection of the articulatory features, the overall design of the model, and the implementation of training techniques. The experiments were carried out on the single-speaker LJSpeech dataset, and utilized the Comparison Mean Opinion Score (CMOS) rating system for an objective evaluation of the synthesized speech quality. In this section, we detailed the results and the corresponding discussions.

To verify the effectiveness of articulatory variations modeling in Section 3.3, we directly removed the vocal tract variables(TVs), resulting in a 0.20 decrease in CMOS. thereby underscoring the significance of these vocal features. Additionally, we simplified the multidimensional style vector mapping network. The various articulatory features and mel spectrograms were directly concatenated,

and the overall style of the target audio was extracted by a single mapping network. This led to a 0.11 decrease in CMOS.

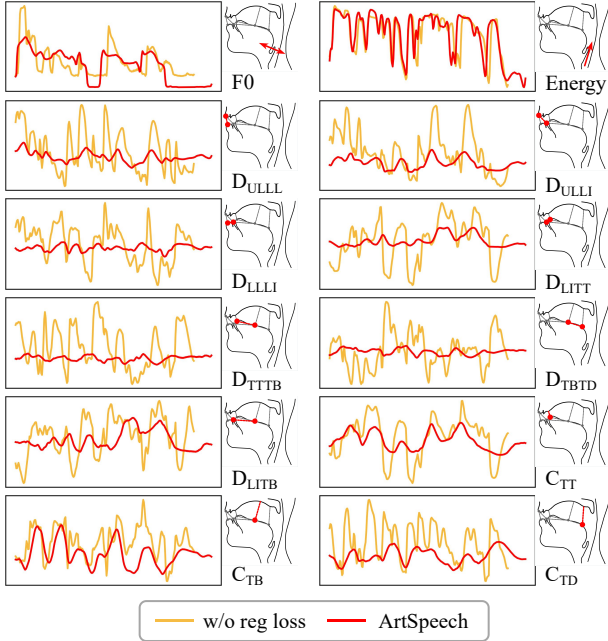


Figure 2: Visualizations were made for articulatory variations predicted by *ArtSpeech* (red line) and w/o L_{reg} loss (yellow line). The two subfigures above show the trajectories of pitch and energy, while the 10 subfigures below display the trajectories of 10-dimensional TVs for the same speech. The synthesized text is: "came up in all respects to modern requirements."

To verify the effectiveness of the model design method in section 3.4, we respectively removed the additional articulatory encoder and duration encoder, using the output of the encoder in the mel predictor as phoneme features to estimate articulatory variations or duration. This resulted in a 0.32 and 0.19 decrease in CMOS, respectively. This outcome demonstrates that independent encoder designs are instrumental in achieving more nuanced feature modeling, significantly enhancing the quality of speech synthesis.

In order to validate the effectiveness of the multi-step training method mentioned in section 3.5, we first removed the L_{reg} loss from the first training step, which meant that the training of the F_0 extractor and TVs extractor proceeded without any explicit constraints. This alteration precipitated a substantial decrease of 0.60 in the CMOS score, thereby highlighting the constructive role of explicit articulatory features in enhancing the speech synthesis process. Figure 2 presents the visualization results of articulatory variations predicted from the text by both *ArtSpeech* and w/o L_{reg} loss models. It is observed that compared to the predictions of *ArtSpeech*, the F_0 and energy estimated by w/o L_{reg} loss from text exhibit considerable irregular noise, making it challenging to represent true pronunciation conditions. Since the extraction of energy is directly calculated from the mel spectrogram, its accuracy is

less affected. This suggests that extracted articulatory features may gradually deteriorate into latent variables during subsequent training if constraints are not applied to the pre-trained articulatory extractors, leading to unpredictable results in model synthesis. Additionally, we eliminated the warm-up step of the second training step and fine-tuned the entire model directly. This led to a 0.14 decrease in CMOS.

B SUBJECTIVE EVALUATION DETAILS

In this section, we provide detailed information about the evaluation process. For a comprehensive and unbiased evaluation, native English-speaking raters from the United States were recruited through Amazon Mechanical Turk. Each test sentence was rated by at least 20 raters, with the sequence of model presentations randomized to prevent any order bias. To ensure high-quality evaluation results, all questionnaires were distributed with two filters activated, including HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 95 and the Number of HITs Approved greater than or equal to 50.

We utilized the *LibriTTS-test-clean* dataset to evaluate the synthetic performance of the zero-shot style transfer of Out-of-Domain (OOD) custom voice by our model. The official pre-trained models of StyleTTS², YourTTS³, and StyleTTS 2⁴ were used to generate the test speech. We also use an open-source implementation of VALL-E X⁵ zero-shot TTS model. To ensure a robust assessment, we removed the four speakers with the fewest speech samples from *LibriTTS-test-clean*. From the remaining 35 speakers, we randomly selected two speech samples per speaker as reference speech and chose two texts at random for voice synthesis. We used pre-trained HiFiGAN⁶ from ESPnet to synthesize the final speech for *ArtSpeech*. The evaluation of the synthesized speech's similarity to the prompt speech was conducted using the Mean Opinion Score of Overall Similarity (MOS-O) and Comparison Mean Opinion Score of Similarity (CMOS-O).

MOS-O was used to measure the similarity between test speech and prompt speech, with a scoring scale ranging from 1 to 5 in increments of one. For each prompt speech, we provided five audios to be tested, including ground truth, and synthesized audio from *ArtSpeech* or the three baseline methods. To enable raters to better compare the gap between prompt speech and the test speech, we provided pairs of speech to the raters in a random order. To avoid auditory fatigue, each rater was required to evaluate 35 pairs of speech. The following shows the instructions presented to the rater for MOS-O evaluation.

²<https://github.com/y14579/StyleTTS>

³<https://github.com/Edresson/YourTTS>

⁴<https://github.com/y14579/StyleTTS2>

⁵<https://github.com/Plachtaa/VALL-E-X>

⁶[parallel_wavagan/libritts_hifigan.v1](https://github.com/libritts_hifigan.v1)

The purpose of this test is to measure the similarity between two speeches. You will be provided with 35 audio pairs, each containing a speech prompt and an audio clip.

Please consider their overall similarity, as if they were recorded in the same place by the same speaker using the same style.

During the test, please wear a headset and adjust the sound volume to a level that is comfortable for you. Please maintain the sound volume throughout the test and refer to the following scale to give a rating:

- Bad - Completely dissimilar speech - 1
- Poor - Mostly dissimilar speech - 2
- Fair - Equally similar and dissimilar speech - 3
- Good - Mostly similar speech - 4
- Excellent - Completely similar speech - 5

CMOS-O was used to assess whether the audio produced by *ArtSpeech* is more similar in style to prompt audio compared to other audio to be tested. The CMOS-O scoring range is -2 to +2, with intervals of one point. We randomly divided 280 pairs of speech to be tested into 8 batches of 35 groups each. One sentence in each pair is the result of *ArtSpeech* synthesis, and the other is the result of the baseline method or ground truth. The following shows the instructions presented to the raters for CMOS-O evaluation.

The purpose of this test is to compare the overall similarity between the two clips and the target prompt. You will be provided with 35 tasks, each containing a speech prompt and two clips. Please compare which clip is more similar to the target audio. Please consider their overall similarity, as if they were recorded by the same speaker using the same style.

During the test, please wear a headset and adjust the sound volume to a level that is comfortable for you. Please maintain the sound volume throughout the test and refer to the following scale to give a rating:

- clip 1 closer: -2
- clip 1 slightly closer: -1
- clip 1 and clip 2 are about the same : 0
- clip 2 slightly closer: 1
- clip 2 closer: 2

To further evaluate the speech quality of *ArtSpeech*, we randomly selected 50 texts from the test set of LJSpeech. The evaluation incorporated two subjective metrics: the Mean Opinion Score of Similarity (MOS-S) and the Mean Opinion Score of Quality (MOS-Q). Official pre-training models of StyleTTS, VITS⁷, and FastSpeech 2⁸ were used to generate the test audio. We used the pre-trained HiFiGAN⁹ from ESPnet to synthesize the final speech for *ArtSpeech*.

The purpose of MOS-S is to compare the voice tone and prosody similarity between synthesized speeches and the ground truth. The following shows the instructions presented to the rater for MOS-S evaluation.

The purpose of this test is to measure the voice tone and prosody similarity between two speeches. You will be provided with 25 audio pairs, each containing a speech prompt and an audio clip.

Please consider whether the voice tone and prosody of the clip, including the speed, rhythm, and accent, are similar to that in the speech prompt.

During the test, please wear a headset and adjust the sound volume to a level that is comfortable for you. Please maintain the sound volume throughout the test and refer to the following scale to give a rating:

- Bad - Completely dissimilar speech - 1
- Poor - Mostly dissimilar speech - 2
- Fair - Equally similar and dissimilar speech - 3
- Good - Mostly similar speech - 4
- Excellent - Completely similar speech - 5

MOS-Q focused on evaluating the overall audio quality, scrutinizing aspects such as clarity, continuity, accuracy, and the presence of distortions or noise. The following shows the instructions presented to the rater for MOS-Q evaluation.

The purpose of this test is to evaluate the speeches' quality. You will be provided 5 clips.

To rate each clip, please consider their overall quality score, you can consider the following aspects: overall sound quality and background noise; the clarity of pronunciation, any missing words, and mumbling. Please note that you should not consider the speaking style, including timbre, emotion, and prosody during the rating.

During the test, please wear a headset and adjust the sound volume to a level that is comfortable for you. Please maintain the sound volume throughout the test and refer to the following scale to give a rating

- Bad - 1
- Poor - 2
- Fair - 3
- Good - 4
- Excellent - 5

⁷<https://github.com/jaywalnut310/vits>

⁸<https://github.com/ming024/FastSpeech2>

⁹[parallel_wavagan/ljspeech_hifigan.v1](https://github.com/ljspeech_hifigan)