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

Thai Speech Spoofing Detection Dataset with Variations in Speaking Styles

Ticho Urai, Pachara Boonsarngsuk, Ekapol Chuangsuwanich

Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Thailand
ticho.urai@gmail.com, pacharawinboon@gmail.com, ekapol.c@chula.ac.th

Abstract

We develop the Chula Spoofed Speech (CSS) dataset, a spoofing dataset for Thai, which contains 1,332,120 utterances of both bona fide and synthetic speech. Synthetic speech samples were generated using five distinct high-quality text-to-speech (TTS) systems, all based on the same utterances as the bona fide data. The data covers various age ranges and speaking styles. Strong baselines such as AASIST and RawNet2 are trained under different conditions to uncover aspects that affect the performance of the models. Besides unseen attacks, unseen speaking styles also have a big impact on performance, indicating a need for diversity in speaking styles in anti-spoofing datasets. Furthermore, we investigate the models in telephony scenarios against additional TTS systems. The results reveal that the models still face certain challenges in this context.
Index Terms: anti-spoofing, anti-spoofing dataset, text-to-speech, speech spoofing detection

1. Introduction

Automatic speaker verification (ASV) is a system for identity authentication that relies on an individual's speech. ASV has gained widespread adoption and is used in various systems, such as voice-controlled smart assistants (e.g., Alexa and Google Assistant) and phone-based banking systems [1]. However, with rapid advances in generative artificial intelligence, recent models can produce highly realistic, natural-sounding speech that closely mimics genuine human voices, making detection increasingly difficult for both machines and humans, especially in telephony situations with lower-quality sound. Additionally, the increasing availability of large language models (LLMs) raises concerns regarding possible automated large-scale text-to-speech (TTS) spoofing attacks. This creates a need for countermeasures in order to catch TTS-based spoofing attacks.

According to [2], spoofing attacks can be broadly categorized into four main types: impersonation, replay, voice conversion, and synthetic speech. The first two types rely on either reproducing or imitating the original voice, which is human-generated, whereas the latter two produce artificial speech generated by algorithms that is nearly indistinguishable to human listeners. Among these four attacks, synthetic speech and voice conversion are commonly used to exploit vulnerabilities in voice-based logical access (LA) systems.

To address these vulnerabilities, researchers have worked to develop and evaluate anti-spoofing systems using the ASVspoof [3, 4, 5, 6] benchmark datasets. The most recent dataset, ASVspoof5 [7], comprises both bona fide (genuine) and spoofed (synthesized) speech samples, where the spoofed speech is generated using synthetic speech and voice conversion techniques derived from state-of-the-art models. However, Tamayo Flórez et al. indicated that variations in pronunciation and accents can significantly degrade the performance of detection models when applied across different languages [8]. This finding has encouraged researchers to develop language-specific anti-spoofing datasets. For instance, the VSASV dataset [9] was created for Vietnamese and the CFAD dataset [10] for Chinese. For Thai, the ThaiSpoof dataset [11] was recently proposed. ThaiSpoof comprises both bona fide speech, sourced from the Common Voice dataset, and spoofed speech samples generated using text-to-speech (TTS), F0 modification, and pitch shifting. The TTS models used in this dataset include the VAJA 9.0 system[1] along with a multilingual TTS model [12], totaling five different TTS voices. However, the bona fide speakers and the spoofed voices come from different sets of speakers. Consequently, it remains unclear whether anti-spoofing models trained on this dataset properly detect spoofed speech based on artifacts inherent to synthetic speech or merely differentiate them by the distinct vocal characteristics of the speakers. This limitation constrains the applicability of the ThaiSpoof dataset for evaluating anti-spoofing models.

We present the Chula Spoofed Speech (CSS) dataset, a Thai-language dataset specifically designed for the development and evaluation of anti-spoofing systems. Our dataset takes into account sex, age, and speaking style of the speakers during data collection. For spoofed data generation, we utilize various text-to-speech models to produce utterances that align with their bona fide counterparts. Additionally, we conduct further studies to investigate the impact of speaking styles on model performance and evaluate the model's effectiveness in a telephony scenario. Our results demonstrate that speaking style greatly affects the performance of spoof detection models.

2. Dataset Construction

Our dataset consists of audio samples from 20 speakers, covering a diverse range of genders, age groups, and speaking styles, with both the bona fide and spoofed data originating from the same speakers and the same utterances. The overall dataset construction pipeline is illustrated in Figure 1.

2.1. Data collection

The bona fide portion of our dataset comprises recordings from 20 professional voice actors, consisting of 10 male and 10 female speakers. These individuals were selected to cover a wide range of ages and speaking styles, thus ensuring diversity within the dataset. Specifically, the speakers were categorized into three age groups: 8 adolescents, 9 working adults, and 3 elderly

[1] https://aiforthai.in.th/services/vaja9.html

10.21437/Interspeech.2025-1895

individuals, to capture the various speech characteristics associated with different stages of life. Each speaker was assigned exactly one speaking style out of three distinct categories: formal speech (6 speakers), casual speech (10 speakers), and excited speech (4 speakers). This design aims to create a dataset that is diverse, comprehensive, and reflective of real-world scenarios.

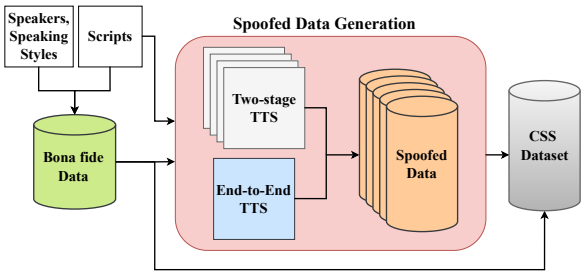


Figure 1: *Overview of the CSS Dataset creation.*

The 20 speakers received the same set of 11,101 scripts, which were gathered from various sources. The selected scripts were adapted from [13]. By having each of the 20 speakers record all 11,101 scripts, the resulting dataset consists of 222,020 utterances, totaling approximately 270 hours of recorded speech. All audio samples were recorded at a 44.1 kHz sampling rate in single-channel format, provided in 16-bit linear PCM (RIFF) format. This dataset was manually verified, focusing on both audio quality and script accuracy, to ensure high-quality and reliable data.

2.2. Spoofed data generation

To generate spoofed data for our dataset, we focus on the Text-to-Speech (TTS) approach, where raw waveforms are generated directly from textual input. In training the text-to-speech model, we use data from the bona fide portion to train the model. The audios were downsampled to 22.05 kHz for TTS training. Once the training process was completed, we used the model to generate spoofed data using the same text scripts as those in the bona fide set. To cover a wide range of neural models, we investigated both two-stage and end-to-end TTS systems. This approach diversified the synthetic speech output.

2.2.1. Two-stage TTS

For the two-stage TTS system, the process consists of two main steps. The first step involves generating intermediate speech representations, such as mel-spectrograms from textual input [14]. In the second step, these intermediate speech representations are transformed into raw waveforms by a vocoder. In this work, we employ two different spectrogram generator models, namely **FastPitch** [15] and **Tacotron 2** [14], along with two different vocoder models, **HiFi-GAN** [16] and **UnivNet** [17].

In training the two-stage TTS system, we adopt a single-speaker TTS configuration, where each model is designed to synthesize speech for only one speaker. This is to ensure that the generated speech is of high quality and completely matches the speaker's identity. The spectrogram generators were trained from random initialization. For the vocoder, we initialized from pretrained models[2] trained on the English LJSpeech dataset [18]. The models were fine-tuned using the mel-spectrograms, ground truth alignments, and durations in teacher-forcing mode, following [14].

This process resulted in spoofed data generated from four different combinations: Tacotron 2 + HiFi-GAN, Tacotron 2 + UnivNet, FastPitch + HiFi-GAN, and FastPitch + UnivNet.

2.2.2. End-to-End TTS

An end-to-end TTS system is a Text-to-Speech architecture that integrates the two modules of a two-stage TTS system into a single unified model. This approach enables the direct transformation of textual input into raw waveforms without the need for intermediate speech representations. In this work, we utilize the **VITS** model [19], a widely used end-to-end TTS system, to generate spoofed data for our end-to-end TTS component. Additionally, we followed YourTTS's [20] training methodology to train a VITS model for multi-speaker TTS, enabling a single model to synthesize speech for multiple speaker identities. For speaker embedding extraction, we utilize the publicly available H/ASP model[3] [21] to serve as the speaker encoder.

In constructing the multi-speaker TTS system with the VITS model, we use an H/ASP model to extract speaker embeddings for each file, and then average the embeddings from the same speaker to obtain a representative embedding for each speaker. These embeddings, alongside the bona fide dataset, are then used to train the VITS model, where the speaker embedding serves as a conditioning factor specifying each speaker's voice characteristics. After the training is complete, we utilize the resulting VITS model to generate spoofed data, thereby producing spoofed data sourced from the VITS model.

Table 1: *Summary of the CSS dataset, detailing the number of speakers and utterances for each data type.*

Type	Algorithm	#Speakers	#Utts
Bona fide	-	20	222,020
Spoofed	Tacotron 2 + HiFi-GAN	20	222,020
Spoofed	Tacotron 2 + UnivNet	20	222,020
Spoofed	FastPitch + HiFi-GAN	20	222,020
Spoofed	FastPitch + UnivNet	20	222,020
Spoofed	VITS	20	222,020

2.3. CSS: Chula Spoofed Speech Dataset

The CSS dataset is a Thai spoofing dataset that comprises both bona fide speech and spoofed speech, recorded and generated from 20 speakers, with the bona fide and spoofed speech paired, originating from the same speaker and utterance. This extensive dataset includes a total of 1,332,120 utterances, amounting to approximately 1,620 hours of speech data[4]. The spoofed data is generated using five distinct text-to-speech (TTS) approaches, ensuring diversity in synthesis methods[5]. Detailed information regarding the algorithms and the composition of the dataset can be found in Table 1, which highlights the distribution of bona fide and spoofed speech across different TTS approaches.

We compared our CSS dataset with other spoofing datasets, such as the Vietnamese VSASV dataset, the Chinese CFAD dataset, the HABLA dataset [8] with five Spanish accents, and

[2] `https://catalog.ngc.nvidia.com/models`

[3] `https://github.com/Edresson/YourTTS`

[4] More information at `https://github.com/SLSCU/CSS`

[5] To validate the quality of our TTS, we conducted a small survey asking 20 listeners. They could not reliably distinguish between bona fide and spoofed speech.

Table 2: *Comparison of various spoofing datasets. VB denotes Vocoder-Based, a process that converts an intermediate speech representation (Mel spectrogram) into a raw audio waveform using a vocoder. Augmented refers to pitch shifting and F0 modification. AP denotes Adversarial Perturbation. PS denotes Partial Spoof. #TTS algorithms count only real TTS-based attack algorithms. #TTS voices include only speakers from TTS and exclude those from VC voices. HABLA contains one TTS-only algorithm and two TTS+VC algorithms.*

Dataset	**Language**	**#Attack algorithms**	**Spoofed Type**							**#TTS algorithms**	**#TTS voices**
			VB	TTS	VC	Replay	Augmented	AP	PS		
ThaiSpoof	Thai	4		✓			✓			2	5
VSASV	Vietnamese	3			✓	✓		✓		0	0
CFAD	Chinese	12	✓	✓					✓	2	24
HABLA	5 Spanish accents	6		✓	✓					3*	10
CSS	Thai	5		✓						5	20

the existing Thai-language ThaiSpoof dataset, across various aspects. Detailed information is provided in Table 2. Since generating high-quality TTS with multiple speaker voices in low-resource languages is challenging, these languages tend to avoid relying solely on TTS-based spoofing and instead employ alternative techniques. However, we consider TTS-based attacks highly scalable and dangerous, as they can be combined with LLMs to create deceptive, large-scale scam call bots. As a result, when compared to all existing datasets, ours contains the highest number of TTS-based attack algorithms. Compared to ThaiSpoof and VSASV, our dataset provides a greater variety of attack algorithms and a larger number of TTS speaker voices, enabling a more comprehensive evaluation of model robustness against diverse spoofing threats. While ours contains more TTS speakers than HABLA, it has slightly fewer than CFAD. Overall, our dataset features a diverse range of TTS-based spoofing methods, allowing for a thorough assessment of model performance in handling TTS-based attacks.

3. Experimental Setup

In this section, we elaborate on the different experimental setups to showcase how our dataset can facilitate research in spoofing countermeasures.

3.1. Training and Testing set

Since our dataset includes a diverse range of speakers, attack types, and speaking styles, we designed our experiment to evaluate both the performance and generalizability of the anti-spoofing models. To achieve this, we implemented three distinct train-test splitting setups:

1. **Split by speaker** We selected 18 out of the 20 speakers in our dataset and divided their data into a training set, validation set, and seen speaker test set. To assess the model's generalizability on unseen speakers, we used the remaining data from two speakers, one male and one female, to form the unseen speaker test set.
2. **Split by attack type** From the five types of TTS attacks in our dataset, we used data from all four of the two-stage TTS-based attack as the training set, validation set, and seen attack test set. Data from the VITS model is used as the unseen attack test set.
3. **Split by speaking style** To the best of our knowledge, no prior research has explored the performance of anti-spoofing models on unseen speaking styles. We observed that most speaking styles in other datasets tend to be formal. Therefore, we selected formal speaking style data to develop our anti-spoofing model. Similar to the previous two setups, we split the formal speaking style data into a training set, a validation set, and a seen-style test set. The remaining casual and excited speech data were used for the unseen style test set.

In every setup, we randomly selected data from our entire dataset, resulting in roughly 210,000 utterances for the training set, around 50,000 for validation, and about 109,000 for the seen test set. For each unseen test set, we randomly sample approximately 55,000 utterances. This sampling ensured that each of the three settings had a similar amount of training data. The proportion of bona fide and spoofed data in all sets followed [7]. All audios were downsampled to 8,000 Hz, with the raw waveform input duration fixed at approximately 4 seconds (32,000 samples), following [22, 23].

3.2. Anti spoofing models

As baselines, we selected two well-known end-to-end models: **AASIST** [22] and **RawNet2** [23]. These models take raw 4-second audio as inputs for the spoofed classification model. We trained each model on the prepared training set for each experimental setup, using a binary classification format with a weighted cross-entropy loss function.

3.3. Evaluation metrics

There are two standard metrics for evaluating spoof detection models: equal error rate (EER) and minimum detection cost function (min-DCF), as described in [7].

4. Results and Discussions

Table 3 presents the performance of the anti-spoofing models on our three setups. Overall, both AASIST and RawNet2 detected spoofed speech more effectively in the seen test sets than in the unseen test sets across all setups, which is expected. Since one of our key objectives is to evaluate the model's generalizability when trained on our dataset, we will focus our analysis on the AASIST model, as it shows a lower relative increase in both EER and min-DCF between the unseen and seen test sets, indicating better generalization. For AASIST, the EER in the unseen test sets was approximately 4, 11, and 57 times higher than in the seen test sets for the speaker, attack, and speaking style setups, respectively. Notably, the speaking style setup exhibited

Table 3: *Evaluation results of AASIST and RawNet2 models under different experimental setups, presenting EER and min-DCF values.*

Setup	Testing set	EER (%) ↓		min-DCF ↓	
		AASIST	RawNet2	AASIST	RawNet2
Split by speaker	Seen 18 speakers	0.27	0.28	0.0076	0.0072
Training set: 18 speakers	Unseen 2 speakers	1.04	1.13	0.0287	0.0323
Split by attack type	Seen 4 attacks	0.33	0.24	0.0091	0.0063
Training set: 4 attacks	Unseen 1 attack (VITS)	3.62	4.45	0.1011	0.0992
Split by speaking style	Seen formal style	0.12	0.14	0.0031	0.0036
Training set: formal style	Unseen style (casual, excited)	6.86	8.83	0.1870	0.2260

the highest relative increase in EER, rising 57 times from the seen style result, suggesting that speaking style greatly affects the spoof detection performance. To verify this, we conducted further investigation.

4.1. Impact of Speaking Style

To examine the impact of speaking style on model performance, we conducted an additional study using the AASIST model. We selected speech from four out of six formal-style speakers to form the training set, while the remaining two were used as the in-domain test set, serving as the baseline for this study. To form the out-of-domain test sets, we selected two speakers' data for the casual test set and two for the excited test set. In other words, each test set contained unseen speaker data, but the formal-style test set shared the same speaking style as the training set, whereas the out-of-domain test set contained data from an unseen speaking style.

Table 4: *Performance of the AASIST model across speaking styles. To account for variability in speaker selection, the speakers in the training and testing sets were shuffled three times. We report the mean ± standard deviation across these runs.*

Testing set	EER (%) ↓	min-DCF ↓
Formal	3.77 ± 0.42	0.1305 ± 0.0501
Casual	6.85 ± 1.44	0.1743 ± 0.0419
Excited	8.17 ± 1.47	0.2045 ± 0.0457

As shown in Table 4, AASIST achieved an EER of 3.77% on the in-domain test set (formal style). However, performance significantly deteriorated on the out-of-domain style sets, particularly on the excited style, where the EER increased by over 200%. We also conducted an experiment training the model with casual-style data, and the results followed the same trend. This result shows that the anti-spoofing model is sensitive to variation in speaking style and suggests that speaking style should be a key consideration in future research.

4.2. Real-World Telephony Use Cases Evaluation

To evaluate the performance of the anti-spoofing models developed using our dataset in a real-world use case, we simulated spoofing attacks transmitted through telephony channels. Audio files were processed through simulated telephony channels with effects such as packet loss, background noise, and jitter. The split by speaker configuration was used, with 18 speakers for training and 2 speakers for the in-domain test set. Six different external test sets, referred to as "A"-"F", gathered from six external sources, were used for further evaluation. The same set of text was used to generate the audio. All audio files were processed through simulated telephony conditions. The evaluation was conducted using the AASIST model.

The results in Table 5 show that detecting spoofing attacks in telephony channels is more challenging than under normal conditions. The EER for the in-domain test set reached 6.90%, which is nearly 7 times higher than that of normal evaluation. For the external test sets, three out of six test sets (D, E, and F) demonstrated that our models effectively detected spoofing attacks. Notably, the results from the D and E test sets achieved EERs lower than that of the in-domain test set. However, for the remaining 3 test sets, the models showed degraded performance compared to the results obtained on the in-domain test set. These results indicate that our dataset is of sufficient quality to support the development of anti-spoofing models for the Thai language in real-world telephony situations. However, there remain several aspects where detection performance is still suboptimal, indicating challenges that require further investigation and improvement in future work.

Table 5: *Performance of the AASIST model under simulated telephony conditions.*

Testing set	EER (%) ↓	min-DCF ↓
Unseen 2 speakers	6.90	0.1692
A test set	19.12	0.4800
B test set	15.73	0.4483
C test set	25.00	0.7217
D test set	6.62	0.1417
E test set	0.17	0.0033
F test set	9.27	0.2617

5. Conclusion

We proposed the CSS dataset, a new Thai spoofing dataset comprising bona fide data from 20 speakers with diverse speaking styles, as well as five types of spoofing attacks generated using different Text-to-Speech (TTS) approaches while maintaining the same utterances as the bona fide speech. Our proposed baseline focuses on evaluating the generalizability of anti-spoofing models across various aspects, including speakers, attacks, and speaking styles. Our results indicate that speaking styles must be taken into account, as they significantly impact the performance of anti-spoofing models.

6. Acknowledgement

This research was jointly supported by the PMU-C grant (C05F660049) and Amity Accentix Co., Ltd.

7. References

[1] T. Arif, A. Javed, M. Alhameed, F. Jeribi, and A. Tahir, "Voice spoofing countermeasure for logical access attacks detection," *IEEE Access*, vol. 9, pp. 162 857–162 868, 2021.

[2] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: from the perspective of ASVspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, p. e2, 2020.

[3] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Interspeech 2015*, 2015, pp. 2037–2041.

[4] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Interspeech 2017*. International Speech Communication Association, 2017, pp. 2–6.

[5] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[6] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.

[7] X. Wang, H. Delgado, H. Tak, J. weon Jung, H. jin Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. H. Kinnunen, N. Evans, K. A. Lee, and J. Yamagishi, "ASVspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale," in *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, 2024, pp. 1–8.

[8] P. A. Tamayo Flórez, R. Manrique, and B. Pereira Nunes, "HABLA: A Dataset of Latin American Spanish Accents for Voice Anti-spoofing," in *Interspeech 2023*, 2023, pp. 1963–1967.

[9] V. Hoang, V. T. Pham, H. N. Xuan, N. Pham, P. Dat, and T. T. T. Nguyen, "VSASV: A Vietnamese dataset for spoofing-aware speaker verification," in *Proc. Interspeech*, 2024, pp. 4288–4292.

[10] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, and R. Fu, "CFAD: A Chinese dataset for fake audio detection," *Speech Communication*, vol. 164, p. 103122, 2024.

[11] P. Leepagorn, K. Z. Mon, K. Galajit, J. S. Hada, N. N. Thilakarathne, and J. Karnjana, "ThaiSpoof: An extension to current methods and database catering advanced spoof detection," in *2024 19th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*. IEEE, 2024, pp. 1–6.

[12] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[13] T. Kongthaworn, B. Naowarat, and E. Chuangsuwanich, "Spectral and Latent Speech Representation Distortion for TTS Evaluation." in *Interspeech*, 2021, pp. 2741–2745.

[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 *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[15] 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.

[16] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[17] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation," in *Interspeech 2021*, 2021, pp. 2207–2211.

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

[19] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[20] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[21] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, "Clova baseline system for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2009.14153*, 2020.

[22] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 6367–6371.

[23] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6369–6373.