Audio Engineering Society

Conference Paper 31

Presented at the AES International Conference on Machine Learning and Artificial Intelligence for Audio
2025 September 8–10, London, UK

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Broadcast-quality synthetic narration: a workflow for fine-grained text-to-speech intonation and emotion control

Luiz F. Kruszielski[1], Pedro H. L. Leite[1,2], Myllene P. Fernandes[1], André Pereira[1], and Luiz W. P. Biscainho[2]

[1]*Globo Group*
[2]*Federal University of Rio de Janeiro*

Correspondence should be addressed to Luiz F. Kruszielski (luiz.fk@g.globo)

ABSTRACT

The ability to generate high-quality synthetic speech with fine-grained control over intonation and emotion is crucial for broadcast applications such as narration and voice-over production. This paper presents a novel text-to-speech (TTS) workflow designed to generate expressive and user-controllable speech synthesis. By leveraging a one-hot encoded emotional embedding approach through application-aware, manually labeled emotion classes and a fine-tuned neural model, our method allows for precise control over speech tone and intention. We detail the data preparation pipeline, including manual annotation and automatic emotion classification, and describe the implementation of our speech synthesis model using VITS2. Additionally, we introduce an intuitive prosody control mechanism that enables word- and phoneme-level adjustments for pitch and duration. Subjective evaluation results indicate that our synthetic narration achieves parity with professionally recorded voice-overs, with some preference for synthetic speech due to its enhanced customization. This research contributes to the advancement of AI-driven voice production, enabling scalable, high-fidelity, and fine-grained controlled speech synthesis for media and entertainment industries.

1 Introduction

Speech is one of the most primal and indispensable forms of direct transfer of complex information between humans. Natural speech does not only carry information through the semantics of the words being expressed; tone and intonation are just as crucial to conveying information. The tone intention of the speech can change the semantic interpretation of the words; for instance, the meaning of a word can be reversed if a sarcastic tone is used. Furthermore, the tone and emotions in natural speech can vary widely in a single sentence. For this reason, producing emotionally diverse speech samples has recently been a hot topic for general speech generation (e.g. [1, 2, 3, 4, 5, 6, 7]).

A common way to create natural emotion and intonation is to use only the semantics of the text to direct the response in tone of voice. However, this approach does not offer flexibility over the same text input. To avoid this issue, we focus on text-to-speech (TTS) with parallel and user-customized emotional and tone input in this work, allowing for the representation of different tones and intonations for the same text prompt. With parallel emotion information, we may control the output speech prosody to input user-based semantic intention, ranging

from lively and enthusiastic speech to neutral, sad or sober.

The established approach to intentionally customize speech style in TTS is to generate a high-dimensional latent space generally represented by embeddings [8] that encode the desired features into vectors, which will serve as input to the neural layers together with or independently of the text [9, 10, 11, 12, 2, 3, 4, 13]. For example, [2] explores pre-trained embedding spaces. In contrast, [3] generates the embeddings through text content, [4] proposes a specific algorithm to extract an embedding space from a previously labeled dataset, and [14] uses an intuitive prosodic encoder to drive speech features utterance-wise. Although these approaches have proven to be easily generalizable, they do not account for application-specific intonations and may fail when the inference domain may interfere with emotional overlapping, and/or a more fine-grained adjustment is required.

From a production/industrial perspective, a particular speech tone and a high voice quality standard are required to create broadcast narrations and voice-overs. Furthermore, voices must connect with the audience, expressing the necessary emotion and intonation to convey the desired message. Any perception by the audience of something "strange" or "unnatural" in a voice can break trust and disconnect from the immersive diegesis of the audiovisual content.

Due to the advances in complex models with well-validated training architectures in that environment, we argue that it may be better to let neural networks produce an embedding space directly through application-aware and perceptually labeled data. Also, employing "one-hot" generated embeddings can leverage the quality of the generated space by reducing ambiguities, since the generating vectors are orthogonally representing semantic features. This approach is beneficial when implementing TTS systems in practice, as it simplifies the inference pipelines while maintaining controllability and operability, even more so when applied together with fine-grained prosody control for eventual adjustments.

Conversely, languages like Brazilian Portuguese (the target language of our models, as seen further) often lack sufficiently large and high-quality datasets suitable for production-grade applications and high-quality demands for speech [15]. This creates a pressing need for optimized data pipelines and algorithms that are explicitly tailored to the linguistic characteristics of such underrepresented languages.

We thus propose a complete workflow for preparing data, generating classes and embeddings, training domain-specific and language-specific models, and creating robust and easy-to-use application interfaces to enable a TTS model scale that can deliver content for millions of users daily.

2 Emotional Embeddings and one-hot encoding vectors

Pre-trained or generalistic embedding networks can be adapted for unseen emotions, and generate meaningful latent spaces from realistic distributions. However, this approach faces some problems when dealing with application-aware emotions. Although some speech-emotional mapping has already been developed (e.g. [16]), it is not straightforward to use pre-defined emotional dimensions when employing TTS systems in production-ready applications, because generally this requires a level of emotional fine-tuning. Furthermore, encoding speech into those dimensions is also a challenge, usually requiring complex and data-demanding pre-training stages or the adaptation of generalist embeddings, which may bring other undesired components into the resulting latent dimensions. In this work, we propose to mitigate these issues by using one-hot encoding vectors to encode emotions, which may be a better choice when emotion classes can be well defined beforehand and coupled with manually labeled data with application context being taken into account.

When using this one-hot encoding approach, the networks may learn the intrinsic emotional dimensions without needing an early latent space design, which is simpler and less contaminative. The downside may be that this learned space is not entirely focused on emotion, but rather on a generalist speech style as seen in [9]; however, when this style is driven from manually labeled data, this contamination can be controlled by adequately generating label classes for emotions and guaranteeing intraclass semantic coherence.

We kept the one-hot embedding approach as our design choice for these reasons. Also, as the use case involves emotion classes from application employments, orthogonally separating the generating vectors should ensure that the intention that drives the embedding generation is the same as when defining the classes, maintaining application-defined requirements until inference stages.

3 Data conditioning and preparation

We used approximately 5 years of archived promo content voice recordings for the dataset corpus, mainly consisting of 30s commercial inserts, recorded by the same narrator. To address TTS applications, general data is needed for prepared speech with corresponding transcripts, and the emotional, language, and accent context is required at inference time. Thus, these recordings were cut into smaller segments of at least four and at most 15 seconds, following silences larger than 0.5 seconds detected in the original audio. Silence and noise-only audio signals have been discarded, resulting in 25800 short sentences. The Python Pydub library[1] was used for this first processing step. The next task was selecting the intentions and emotions of interest, considering their correspondence with the available data and intended application. The subsequent annotation procedure is based on the classes thus defined. Since the original use of the available audio content is to promote various programs, including sports, several film genres, soap operas, and news, it allows for a wide range of different emotions and intonations. All sentences were automatically transcribed using the Whisper [17] library[2], and later manually reviewed for eventual errors.

3.1 Emotion and intonation Classes

The emotion classes were defined based on content and potential application for a synthetic narrator voice. Several distinct states of excitement were established, ranging from a "very excited" tone (e.g., in promotions such as soccer matches) to a "neutral" tone (for instance, during technical network maintenance announcements), including "serious" (e.g., for action movie promotions) and "sad/down" tones. In addition to these emotion-related categories, other non-emotional categories were also defined. The "signature" category captures the narrator's unique way of announcing the broadcast program's name. In contrast, the "date/signature" category is used when the narrator specifies the month or day of the week that the program will air. Because interrogative sentences appeared less frequently in the primary dataset —– resulting in challenges for properly rendering a questioning tone —– a dedicated "question" category was created. This category is the only one that contains duplicated examples to help balance the dataset. Furthermore, a "slate" category was introduced for the introductory words in each recording section, where the narrator announces the date and the targeted program using a "non-narrator" voice. Lastly, a "discarded" category was created for recordings with poor quality or excessive noise; this category was not included in the training dataset. The dataset comprises 11 categories, 10 of which are used for voice synthesis. In Table 1, we show the final class choices and a brief description with text examples representing the samples within them.

[1] https://github.com/jiaaro/pydub

[2] https://github.com/openai/whisper

3.2 Language specific tokenizer

For Brazilian Portuguese (in short, Pt-Br) speech, some text-speech ambiguities (such as contextual and/or syntax-dependent vowel openness and nasalization) should be resolved beforehand, reducing the need for rule-related learning by the neural networks. Therefore, we introduced a stress-dependent phoneme tokenizer in the text processing layers, transcribing the written text into IPA [18] representations (unique text-phoneme mappings) and adapting stress marked phonemes into distinguishable tokens.

While we targeted a language-specific problem with this tokenizer, this kind of adaptation will usually be needed for languages other than English or Mandarin, which are the languages these models have generally been designed for.

3.3 Class predictor

After the classes had been defined and we had manually labeled around a fifth of the dataset (5400 samples totaling around 12 h), we trained an automatic class predictor to generate labels accordingly, for the remaining dataset. The classifier was structured as a generalist pretrained feature embedding generator followed by a small feedforward network adapted to minimize the cross-entropy for the manually labeled data. The generalistic encoder used was HuBeRT [19], which is unsupervisedly trained with many hours of speech and fine-tuned to work in many different domain-specific tasks. After HuBeRT generates the embedding, a simple feedforward network with one hidden layer is applied to the data. The output layer uses a softmax activation function to produce a probability distribution over the emotion classes. Then, the model is trained using cross-entropy loss to minimize the divergence between predicted and ground-truth labels.

Table 1: Defined classes and brief descriptive textual examples.

Class	Class Description
0 - Slate	"Recording: NHA 3/18"
1 - Very excited	Soccer match: "Flamengo versus Botafogo. Who will become the leader?"
2 - Enthusiastic	Variety programs: "At dancing with the famous, a beauty showoff..."
3 - Happy	Film: "3 Times more adventure, Spider-Man 3"
4 - Narrative	Series: "The town hero cannot do anything to save her"
5 - Serious/Neutral	Telenovela: "Now, Sophie will have the help of her daughter..."
6 - Sad/Down	Film: "Sometimes the only thing to change her life is..."
7 - Signature	Production name: "Secret lies"
8 - Question	Questions: "Can we improve our mind through food?"
9 - Date/Signature	Production name + date/day of the week: "Next Monday, Sunday, tomorrow: the Martin"
10 - Discarded	Coughs, noisy recordings, etc.

To measure the performance of the result before extrapolating the actual labels, we split the labeled dataset into training, validation, and test subsets. We performed rounds of fine-tuning on the first two until satisfactory accuracy was achieved. After checking the adjusted model on the test subset, it was applied to the unlabeled data.

To illustrate that generated labels were satisfyingly accurate for our pipeline, Figure 1 shows the confusion matrix for the generated labels in our test subset. We can see a significant correlation between ground-truth labels and the predicted data.

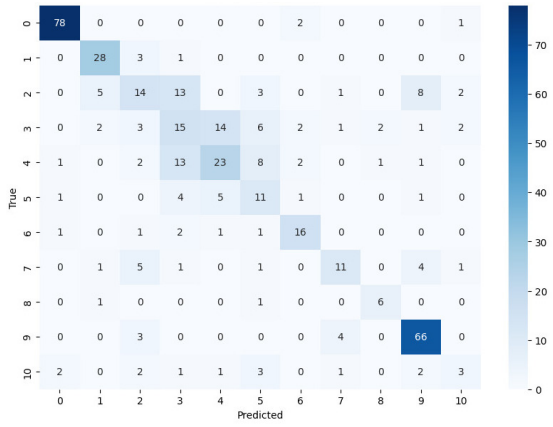


Fig. 1: Confusion matrix for the automatic class predictor against manually labeled data.

However, we can observe that the confusion matrix reveals a few situations in which emotions are not clearly differentiated (e.g. classes 3 and 4 in Figure 1); in fact, they reflect the very difficulty of subjectively classifying the nature of emotion in such cases, which will generally produce class overlap. Although this overlap may suggest a possible class reduction, misclassifications are not harmful in this case, as they can mean that intentions can be interchanged according to the context. This is another advantage of generating application-aware classes, as the overlapping classes may still fit the end-user intention.

Class 10 (which represents "bad quality" samples) is the only one with significant prediction spread, which should be an effect of this different class nature. This is another reason why we discarded this class afterwards.

This way, we could further enhance the synthesis quality by drastically increasing the dataset size without semantic loss, given that the predictor is trained to follow the human criteria to generate labels.

With this automatic class predictor, we completed the necessary data processing. The complete block diagram for all the data pre-processing pipelines is illustrated in Figure 2.

4 Model implementation and training

With the data formatted as a `.csv` file, we used the VITS2 [11] open-source model repository. VITS2 is an evolution of VITS [10], a well-established TTS model with official training implementation procedures out-of-the-box, which made it a community validated codebase. With VITS2, the embedding input architecture was enhanced to modify the phoneme features deeply,

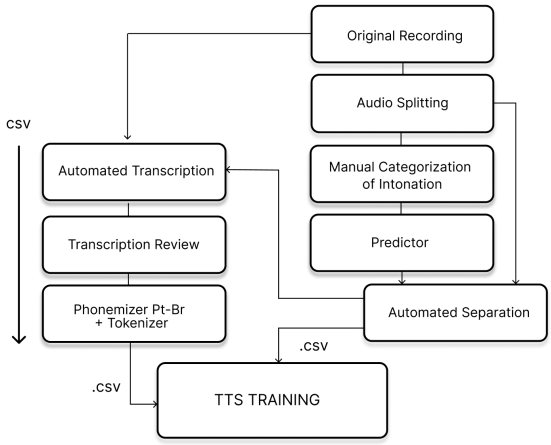


Fig. 2: Block diagram of data-processing pipelines.

and we take advantage of that to get increased emotion control precision and quality.

The VITS2 model was trained with the original hyperparameters (except for batch size) and mostly the original architecture available on the official repository[3], and the training took around 300 thousand steps to converge, which took a week using four NVIDIA RTX 8000 GPUs. The follow-up early stop procedure was to perceptually determine when prosody and quality stopped improving for some time.

With a free embedding network and labeled data, the optimization could find a latent space that encoded emotional features from data content. By following this strategy, we could successfully control intonations by replicating the data present for each label without losing audio quality, as can be verified in the examples available on our website[4]. There, we present samples of the exact text spoken with distinct and controlled emotions/prosody.

5 Fine-grained control

By our experience, even with intonation class control, the editorial team sometimes needs to fine-tune the prosody for a syllable or word. Because of this, we also designed a fine-grained prosody control where the user can choose the duration and pitch with phoneme or word granularity. With this, we can avoid the common text-changing iterations generally needed for high quality speech when the TTS system does not capture the intended prosody.

We enabled fine-grained control by dissecting the VITS2 model. Inside the text feature extraction modules, VITS2 predicts duration and pitch for each token (in our case, phoneme representation). We extracted these phoneme-level features by inferring just the early layers in the network for the given text, which gives us duration and pitch information for each of the corresponding IPA characters, producing two arrays of floats: one for character pitch, and one for character duration, with the length of the phrase in IPA characters plus spacing.

These phoneme-level pitch and duration arrays were then made available for user customization through a graphical interface, in two options: phoneme- or word-level controls. When using phoneme-level controls, the changes are applied directly to the corresponding phonemes, as shown in Figure 3. For word-level knobs, on the other hand, we have to first aggregate the features by taking the mean pitch and duration sum for each word, and then disaggregate them after customization by scaling the individual features using the same factor by which the word-level value was scaled in the interface.

Figure 4 shows an example of the word-level controls in the user interface. The prosody customization is done, in practice, as follows: the user may drag the boxes around a word to increase or decrease its duration and change the slider level inside the boxes to customize pitch for a specific word. The same controls could be more fine-grained, controlling phoneme per phoneme with the same interface.

After the necessary changes, the inference is resumed from the user-tweaked features, telling the models to adapt the speech for slower/faster pronunciation or higher/lower pitch. In order allow controlling the general tendency for a word or space between words, we also grouped same-word phonemes and spaces into chunks, so the user can change the tendency for each chunk as a whole without having to access each character, which is generally preferred for speed.

[3] https://github.com/daniilrobnikov/vits2

[4] https://gpa-smt-ufrj.github.io/aes-aimla2025

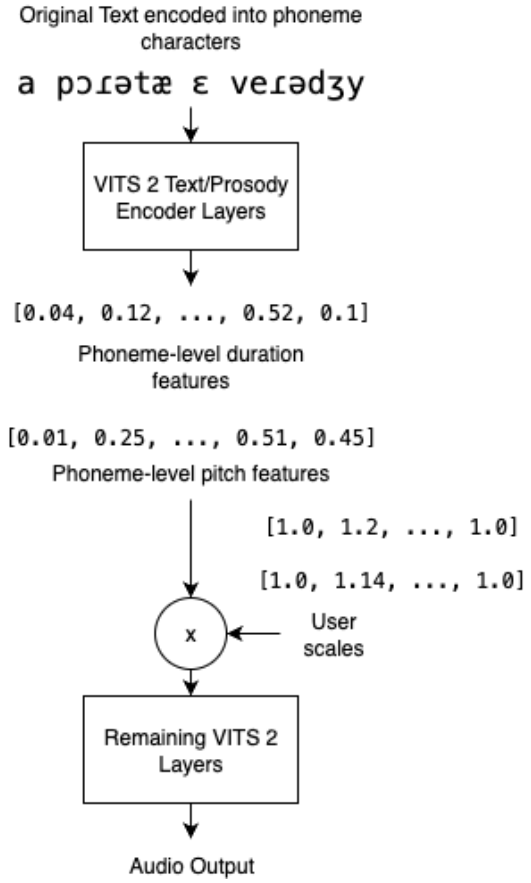


Fig. 3: Block diagram for the phoneme-level feature customization system. The VITS2 architecture is dissected to extract character prosodic features, which are scaled individually as desired by the user.

6 Subjective evaluation

A blind subjective test was conducted to validate the quality of the synthetic voice. In this test, three pairs of video clips were tested in a comparison setup, containing promos of a sports match, a drama series, and a documentary. The same video, music, and sound effects background was maintained for all comparison pairs, and only the narration was changed. The test samples were chosen to match the applications of the promo content, using real promos that aired recently before the time they were selected, and corresponding synthetic replications of them. To ensure consistency, they were also normalized to match the ITU 1770 loudness

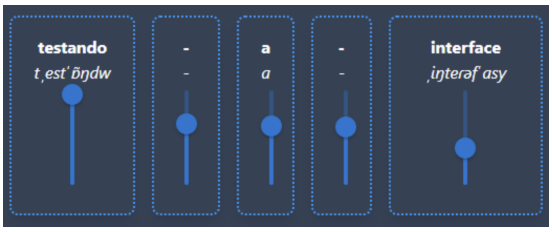


Fig. 4: Fine-grained word-level prosody controls. Each word has a box and a slider. Changing the size of the box by dragging it around customizes the duration of the word, while moving the slider changes the desired pitch intention for that word.

recommendation [20] at -23 LUFS using the *Waves WLM Plus Loudness Meter*[5] plugin for the *Avid Pro-Tools*[6] DAW, and were subject to the same equalization patterns.

For each pair, one clip featured the original recorded narration, while the other presented the synthetic voice generated using our proposed pipeline, with fine-grained adjustments applied to specific words or phonemes when the operator understood them as unnatural. The synthetic samples were also generated to mimic the emotions and intonations that are historically used in such narrations for each context, but without guidance from the real samples. We used a five-point discrete scale where users could evaluate the generated audio against original content, with the given options as in Table 2, where synthetic and original samples were randomly chosen as Sample A or Sample B for each content pair.

The idea was to assess whether or not there was a difference between synthetic-generated narrations and original recordings on the end-user-ready content, and, if there was such a difference, how significant would it be. To address that, we opted for a categorical answer list instead of a continuous/numerical scale, aiming to simplify the evaluation process and minimize potential misunderstandings regarding numerical ratings and their relationship to perceived audio quality.

The subject panel consisted of 29 specialists, resulting in 87 pair evaluations. All subjects were chosen for having notable previous audio and video post-production,

[5] https://www.waves.com/plugins/wlm-loudness-meter

[6] https://www.avid.com/pro-tools

Table 2: Possible options to be chosen by subjects in the audition tests, with assigned linear indexes.

Question Answer options
Sample A is MUCH BETTER than Sample B.
Sample A is SLIGHTLY BETTER than Sample B.
Both samples are EQUALLY ACCEPTABLE.
Sample B is SLIGHTLY BETTER than Sample A.
Sample B is MUCH BETTER than Sample A.

editing, and/or mixing professional experience. To avoid biases, we did not give the subjects prior information about the nature of each piece of content. Also, there was no indication that artificial intelligence was used, and the subjects were asked to optionally leave their observations about the content using a textual field, right after the test.

The equipment used for the tests was a 42-inch TV and a Sony 7506 headphone with equal sound volume for every subject, played from a notebook computer. The original recordings in this test were not used for any data or training procedure.

6.1 Results

Assuming a linear preference scale for the discrete results where 0 represents the response "*Original content is MUCH BETTER than synthetic content*" and 4 represents "*Synthetic content is MUCH BETTER that original content*", we draw a box plot for each content as shown in Figure 5.

We can infer from the results that there is no significant difference between the synthetic and the original recordings, and we even observe a slight tendency towards synthetic preference. This tendency can be explained by the enhanced controllability given to the voice directors, who could adjust prosody according to the finalized content. In the human case, narrations are generally the first asset to be created, and cannot be easily customized afterwards due to logistical difficulties.

As for the optional textual feedback, none of the participants reported any degradation in the speech samples. In fact, some explicitly stated they couldn't detect any

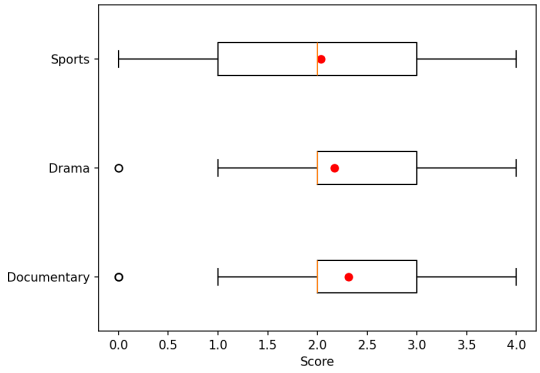


Fig. 5: Box plot for the audition tests. The whiskers show the response range without outliers (white circles), the box represents where 50% of the responses were, orange lines are the medians, and red dots represent the means.

difference in quality between them. We then highlight that these comments should also be good indicators that the synthetic content is ready for production, as even attentive listeners were unable to perceive any notable differences in quality.

The pieces used in the subjective tests and other generated speech samples are available at our website[7]. We also make the individual datapoints from the test available, so it is possible to verify the detailed response distribution.

7 Conclusion and future works

Considering these results, we advanced this workflow to production, since we concluded that delivering the content to a broader audience without any significant perception of loss in quality was feasible.

The proposed method of producing synthetic speech has introduced a new paradigm for scalable and controllable speech production, which is validated by a fast-paced production environment in a relevant broadcast company. Training and pre-processing steps were also optimized and adequate for fast-paced broadcast environments, not to mention the increased controllability and robustness for generating speech content driven

[7] https://gpa-smt-ufrj.github.io/aes-aimla2025/samples

by the operator's intention, with selectable emotions and fine-grained prosody control.

Also, this tool allowed the scriptwriters to test their text choices before the recording stages started and eventually make early changes that could avoid retakes later in the pipeline.

The same methodology was also tested and applied in production for another broadcast environment: voice-overs for pre-recorded clips in a variety show, with new classes and another notable voice in Brazil. With the pipeline validated for promo content, this extension was straightforward and very similar (the only necessary adaptation being to change the data entries and the definition of classes), showing that the proposed workflow can be easily replicated as needed.

As can be seen, this field has vast potential for future research that directly impacts many audio production areas. In the future, we aim to have the same process in a broad-spectrum form, with multiple speakers, accents, and emotions. We hope this can be achieved by decoupling and isolating latent spaces through curated datasets and carefully designed embedding systems, which should employ unseen dimensions for different datasets. We hope this workflow documentation can further impact the industry and academy as a starting point for good practices and an example of a validated and scalable end-to-end synthetic speech system design.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) Finance Code 001.

References

[1] Tang, H., Zhang, X., Cheng, N., Xiao, J., and Wang, J., "ED-TTS: Multi-Scale Emotion Modeling Using Cross-Domain Emotion Diarization for Emotional Speech Synthesis," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12146–12150, 2024.

[2] Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I. L., and Wu, Y., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *32nd International Conference on Neural Information Processing Systems (NeurIPS 2018)*, pp. 4485–4495, 2018.

[3] Kwon, O., Jang, I., Ahn, C., and Kang, H.-G., "Emotional Speech Synthesis Based on Style Embedded Tacotron2 Framework," in *2019 International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pp. 1–4, 2019.

[4] Um, S., Oh, S., Byun, K., Jang, I., Ahn, C. H., and Kang, H.-G., "Emotional Speech Synthesis with Rich and Granularized Control," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7254–7258, 2020.

[5] Kharitonov, E., Vincent, D., Borsos, Z., Marinier, R., Girgin, S., Pietquin, O., Sharifi, M., Tagliasacchi, M., and Zeghidour, N., "Speak, read and prompt: High-fidelity text-to-speech with minimal supervision," *Transactions of the Association for Computational Linguistics*, 11, pp. 1703–1718, 2023.

[6] Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., and Wei, F., "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," `https://arxiv.org/abs/2301.02111`, 2023, online; accessed 1 Oct 2024.

[7] Zhao, W. and Yang, Z., "An Emotion Speech Synthesis Method Based on VITS," *Applied Sciences*, 13(4), 2023.

[8] Bengio, Y., Ducharme, R., and Vincent, P., "A neural probabilistic language model," *Journal of Machine Learning Research*, 3, pp. 932–938, 2000.

[9] Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., and Saurous, R. A., "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *35th International Conference on Machine Learning (ICML)*, pp. 5180–5189, 2018.

[10] Kim, J., Kong, J., and Son, J., "VITS: Conditional variational autoencoder with adversarial Learning for end-to-end text-to-speech," in *2021 International Conference on Machine Learning (ICML)*, pp. 5530–5540, 2021.

[11] Kong, J., Park, J., Kim, B., Kim, J., Kong, D., and Kim, S., "VITS2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design," in *24th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, volume 24, pp. 4374–4378, 2023.

[12] Kim, J., Kim, S., Kong, J., and Yoon, S., "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," in *34th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 8067–8077, 2020.

[13] Doddipatla, R. S., Braunschweiler, N., and Maia, R., "Speaker adaptation in DNN-based speech synthesis using d-vectors," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3404–3408, 2017.

[14] Raitio, T., Rasipuram, R., and Castellani, D., "Controllable Neural Text-to-Speech Synthesis Using Intuitive Prosodic Features," in *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, volume 21, pp. 4432–4436, 2020.

[15] Leite, P. H. L., Hoyle, E., Álvaro Antelo, Kruszielski, L. F., and Biscainho, L. W. P., "Neutral TTS female voice corpus in Brazilian Portuguese," in *41st Brazilian Symposium on Telecommunications and Signal Processing (SBrT)*, pp. 1–5, 2023.

[16] Zhou, K., Sisman, B., Liu, R., and Li, H., "Emotional voice conversion: Theory, databases and ESD," *Speech Communication*, 137(1), pp. 1–18, 2022.

[17] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I., "Robust speech recognition via large-scale weak supervision," `https://arxiv.org/abs/2212.04356`, 2022, online, accessed 15 May 2023.

[18] "International Phonetic Alphabet," `https://internationalphoneticassociation.org/content/ipa-chart`, 2023, online; accessed 06 Mar 2023.

[19] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A., "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, pp. 3451–3460, 2021.

[20] International Telecommunication Union, Radiocommunication Sector (ITU-R), "ITU-R BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level," Technical Report BS.1770-4, International Telecommunication Union, Geneva, Switzerland, 2015, accessed: 2025-07-22.