INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Controlling Multi-Class Human Vocalization Generation via a Simple Segment-based Labeling Scheme

Hieu-Thi Luong[1], *Junichi Yamagishi*[1]

[1] National Institute of Informatics, Japan

luonghieuthi@nii.ac.jp, jyamagis@nii.ac.jp

Abstract

As prompt-based generative models have received much attention, many studies have proposed a similar model for sound generation. While prompt-based generative models have an intuitive interface for non-professional users to experiment with, they lack the ability to control the generated sounds via a more direct means.

In this work, we investigated the use of a simple segment-based labeling scheme for human vocalization generation, which is a specific subset of sound generation. By conditioning the generative models on the label sequence which marks the vocalization class of the segment, the generated sound can be controlled in a more detailed manner while maintaining a simple and intuitive input interface.

Our experiments showed that simply switching the label scheme from global to segment-based does not degrade the quality of the generated samples in any way and provides a new method of controlling the generation process.

Index Terms: human vocalization, sound generation, speech generation

1. Introduction

Sound generation systems are useful for many practical applications such as generating sound effects for video games [1, 2], movies [3], interactive media, or conveying emotions in human-machine interactions [4, 5]. Hence, there is growing interest in developing a new method for generating sounds through a simple and intuitive input interface. Prompt-based sound generation is one such method, and it has gained much attention [6, 7, 8] due to the increasing interest in prompt-based image generation [9, 10]. While generating by describing the desired output via text input is an intuitive interface in theory, the reality is that it is unreliable and lacks precision. Many subsequent studies on text-to-image generation have focused on developing a new mechanism that enables more direct control over the generated output [11].

Compared with image generation, there is less research focusing on sound. The most straightforward method for sound generation is to train a model conditioned on a one-hot vector label of the desired class [12, 13]. Barahona-Rios *et al.* [4] trained a WaveGAN model [14] conditioned on emotion intent to generate knocking sounds infused with a particular emotion. Kong *et al.* proposed using the SampleRNN [12] to train an auto-regressive model that generates sound from a class label. Similarly, Liu *et al.* [13] used a three-stage approach to generate an intermediate quantized latent feature vector to capture long-term dependencies. Yang *et al.* [8] proposed a prompt-based sound generation system using a diffusion model to generate sounds from quantized tokens. While a prompt-based system provides an engaging and intuitive input interface for regular users, it cannot control the generated samples' structure.

In this study, we propose a labeling scheme for controlling human vocalization generation. Our results demonstrate that simply switching from global labels to segment-based labels make it possible to manipulate the generated samples in a more detail-oriented manner while at the same time maintaining a simple input interface for human intervention. The proposed labeling is model-agnostic and can be integrated into a more complex system. For the experiment in this paper, we tested the method using the auto-regressive sound generation system proposed by Liu *et al.* [13]. The results showed that the quality of the generated samples was not affected and the proposed approach opened up a new control functionality. The remainder of the paper is organized as follows: Section 3 describes our method and system, Section 4 provides detailed information about the experiments and evaluation results, Section 5 explains different ways of manipulating generated samples using the new interface, and Section 6 concludes our findings.

2. Related Work

Our work closely relates to sound generation [15, 16, 14], procedural audio [2], and human vocalization generation [17, 18] such as laughter [19, 20] and affect bursts [21].

Video-to-sound generation is a task of automatically generating a sound sample to accompany a silent video segment [22, 3]. The purpose of such models varies from learning a relationship between sound and material perception [22] to supporting video production by reducing the cost of creating high-quality sound effects [3]. Since these models are tied to video inputs, there is no method for human intervention to obtain the desired output. Subsequently, Cui *et al.* [23] proposed a timbre-controllable video-to-sound model by introducing an acoustic encoder to model timbre information.

Prompt-based sound generation, or text-guided audio synthesis, is a model that generates sound on the basis of a natural-language description [7, 8, 24]. Due to the success of prompt-based generative models for images [25, 10] and video [26], there have been many subsequent studies on similar models for audio. Unlike video-to-sound, which cannot be controlled, prompt-based models are intuitive and suitable for human users. However, due to the impreciseness of natural language, it is tricky to manipulate the output samples through the text interface. Hence, research is being conducted toward a more precise interface [11].

Class-based generation is the most simplistic approach among the three mentioned in this section. We condition the generative model on a class label and enable it to generate new samples [12, 4]. Due to its simplicity, there is no mechanism to

10.21437/Interspeech.2023-1175

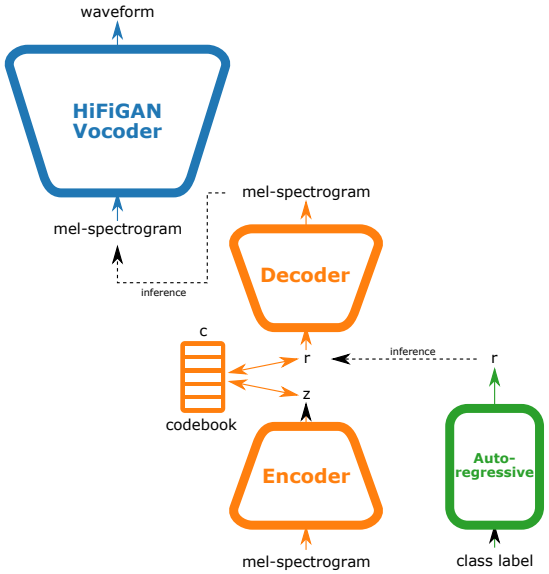

Figure 1: *Multi-stage sound generation system conditioned on vocal class labels. It includes three main components: the HiFiGAN Vocoder that transforms acoustic features to waveform (left), VQ-VAE representation learning model (middle), and auto-regressive model that generates the quantized latent feature vector (right).*

control the output, and users must depend entirely on the model. Our proposed method focuses on improving the flexibility of this class-based approach.

3. Multi-Class Human Vocalization Generation System

3.1. Conditional Sound Generation System

The conditional sound generation system proposed by Liu *et al.* [13] serves as the base for our multi-class human vocalization generation system. Human vocalizations are essentially a specific category of sounds, so we used the source code released by the authors[1] as is. We summarize the basis of the system in this section; readers can refer to the original paper for more details. The sound generation system comprises three main models. Figure 1 illustrates the system's overall structure.

The *HiFiGAN Vocoder* [27] transforms spectrograms into waveforms. We did not use the pretrained model released by Liu *et al.* and instead trained our model on our data using the source code released by Kong *et al.*[2] as our experiments focused on human vocalization rather than general sounds.

The *VQ-VAE representation learning model* was trained on the speech and sound data in a self-supervised fashion. It learned a non-linear mapping from a spectrogram to latent representation z that was further quantized using a learned codebook c into a simpler representation r. The model assumed a fixed-size spectrogram input extracted from a four-second waveform sampled at 22.05 kHz with a window length of 1,024 points and a window shift of 256 points.

The *conditional sound generation model* is an auto-regressive model based on PixelSNAIL [28], which combines causal-convolutions layers with a self-attention mechanism. The model is conditioned on a one-hot vector class label and trained to generate a codebook k index sequence one after another. We turn the generated sequence into the feature map r using the codebook and then into a spectrogram using the VQ-VAE decoder.

The setups of these models were kept the same as in their original papers. Our focus is testing a novel labeling scheme that can create a new interface for controlling generation, which will be explained in the following section.

[1] https://github.com/liuxubo717/sound_generation

[2] https://github.com/jik876/hifi-gan

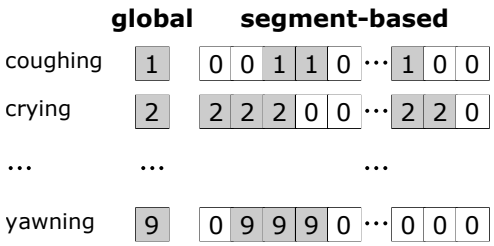

Figure 2: *Global and segment-based class labeling schemes.*

3.2. A Simple Scheme of Segment-based Labeling

Given the model proposed by Liu *et al.* [13] described in the previous section, we generate a sound sample of a specific class by feeding the class label to the auto-regressive model without any other mechanism to manipulate the outcome. We proposed using a segment-based label scheme instead of a single global label for the entire sample to address this limitation. Specifically, given a four-second sample, we prepared an 86-dimensional label using a sequence of overlapped windows. The value of each element of the label vector will either be 0, representing a silence frame, or k, the index of the specific vocalization class, depending on the samples within the corresponding window. Figure 2 presents the difference between the global and the segment-based labeling schemes. Our proposal is a more straightforward and highly abstract version of the waveform silhouette described in [20].

For example, given a training sample of a laughing sound (with laughing being assigned the index of 3), we used a sequence of windows with a window length of 4,096 points and hop length of 1,024 points to extract and define a segment-based label sequence. The element will receive the label value 0 if the root mean square (RMS) value within its window is below a specified threshold, which was -24 dB in our experiments. If the RMS value is above the threshold, then the element will receive its class value, which is 3 in this case. The 0 index was reserved for silence frames, and it was used across samples of all classes.

To use this segment-based label scheme to generate a new four-second sample, we feed the generation model an integer sequence with 86 values. This interface is simple enough for human intervention but more flexible than a global class label.

4. Experiments

4.1. Data

We used VCTK [29] and a subset of the commercial Deeply Nonverbal Vocalization dataset[3] for training. The VCTK speech corpus was used for pretraining as speech data were expected to

[3] https://www.babba.ai/nonverbal

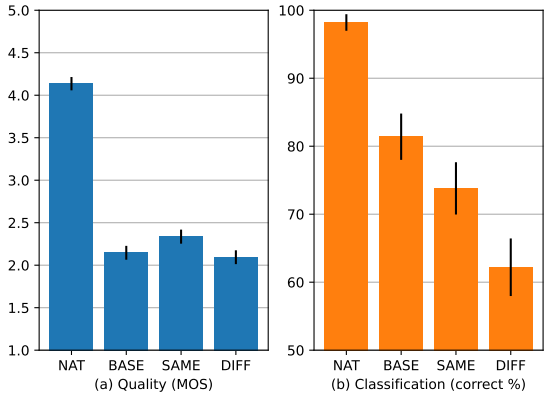

Figure 3: *Subjective evaluation with human perception.*

be beneficial for vocalization sound generation. As the Deeply corpus includes vocalization perofrmed by non-professional individuals and the samples were recorded using the participants' mobile phones, the overall quality is relatively low, but we deemed it to be sufficient for our experiments. To enhance the quality, we used the portion marked as 'clean' by the provider and discarded samples marked as 'noisy.' Moreover, we only used nine out of 16 classes included in the dataset, which are coughing (2,732), crying (966), laughing (1,155), moaning (1,745), panting (745), screaming (843), sighing (3,206), throat-clearing (4,414), and yawning (3,090), as it is questionable to classify the rest as vocalization. The amount of data (the number in parentheses) varied between classes as shown. In total, there were 18,896 samples in the training set, 45 for validation, and 100 for testing. All samples were 16 bits and sampled at 22.05 kHz. For the acoustic features, the mel-spectrogram was extracted as described in the HiFiGAN paper [27].

4.2. Model and training configuration

In our experiments, we tested a baseline system that uses global labels, as described in [13], and the proposed system that uses segment-based labels. As we focus on the labeling schemes rather than the model itself, the configuration between the two systems was kept the same wherever possible. More specifically, the same HiFiGAN Vocoder and VQ-VAE model were used for both systems. For the HiFiGAN Vocoder, we first trained it with the VCTK dataset for 120,000 steps at a learning rate of 0.001 and batch size of 12, then continued to fine-tune with the human vocalization data. Similarly, we trained the VQ-VAE model with speech data for 7,000,000 steps at a learning rate of 0.0001 and batch size of 24, then switched to the vocalization data. The difference between the two systems is whether their auto-regressive models are conditioned on global or segment-based labels. We trained these auto-regressive models for 700,000 steps on human vocalization data at a learning rate of 0.0001 and batch size of 24. A single NVIDIA A100 GPU was used to train all models.

4.3. Subjective evaluation setup

Unlike previous work on sound generation [12, 13] which used indirect metrics for evaluations, we believe people should evaluate generation systems as they are the end users. Our evaluation using human listeners is also one of the main contributions of our research. We asked ten participants two types of questions regarding the presented sound samples. First, we asked participants to judge the general quality of a sample out of a typical five-point scale mean opinion score (MOS). Second, we asked participants to classify a sample into one of five classes: coughing, crying, laughing, panting, and yawning (only these five classes were evaluated). The classification task was similar to that presented by Kong *et al.* [12], but instead of using a trained automatic system, we directly asked the listeners. In the end, each participant completed ten sessions. Each session included 20 quality and 20 classification questions. We prepared these sessions so that they had samples from all evaluated systems and all evaluated vocal classes.

Table 1: *Subjective quality evaluations and classification results by vocal classes.*

(a) BASE

Class	MOS	Classification (%)				
		1	2	3	4	5
1. Coughing	2.44	**80.0**	12.0	7.0	0.0	1.0
2. Crying	2.01	2.0	**82.0**	12.0	2.0	2.0
3. Laughing	2.12	1.0	11.0	**86.0**	0.0	2.0
4. Panting	1.88	6.0	9.0	0.0	**71.0**	14.0
5. Yawning	2.28	5.0	1.0	0.0	6.0	**88.0**

(b) SAME

Class	MOS	Classification (%)				
		1	2	3	4	5
1. Coughing	2.73	**92.0**	3.0	3.0	1.0	1.0
2. Crying	2.21	0.0	**64.0**	16.0	3.0	17.0
3. Laughing	2.13	2.0	15.0	**81.0**	2.0	0.0
4. Panting	2.20	3.0	18.0	3.0	**52.0**	24.0
5. Yawning	2.41	6.0	2.0	1.0	11.0	**80.0**

(c) DIFF

Class	MOS	Classification (%)				
		1	2	3	4	5
1. Coughing	2.31	**75.0**	3.0	21.0	0.0	1.0
2. Crying	2.17	5.0	**61.0**	17.0	3.0	14.0
3. Laughing	2.14	3.0	29.0	**63.0**	2.0	3.0
4. Panting	1.74	7.0	20.0	4.0	**31.0**	38.0
5. Yawning	2.11	2.0	9.0	0.0	8.0	**81.0**

The evaluated systems included the natural samples (NAT) held out for evaluation purposes, the baseline system (BASE), which used global labels, and two of our proposed systems, SAME and DIFF, which used segment-based labels. The difference between them is the segment-based labels used for generation; the SAME system used labels extracted from natural samples of the same vocal class to generate sounds, while the DIFF system used labels of natural samples from different classes. For example, the SAME system uses segment-based labels extracted from natural laughing samples to generate laughing samples. In contrast, the DIFF system used labels extracted from coughing, crying, panting, yawning, and screaming (screaming was exclusively used for label extraction). Even though the segment-based label is a highly abstract representation, the label pattern may still affect the nature of generated samples, so we decided to test it in two different ways.

4.4. Evaluation results

Figure 3 shows the evaluation results of the listening test survey. In terms of the perceived quality of the generated samples, all three synthesis systems produced samples that were of significantly lower quality than the natural samples, as expected. Our proposed method, specifically the SAME system, yielded

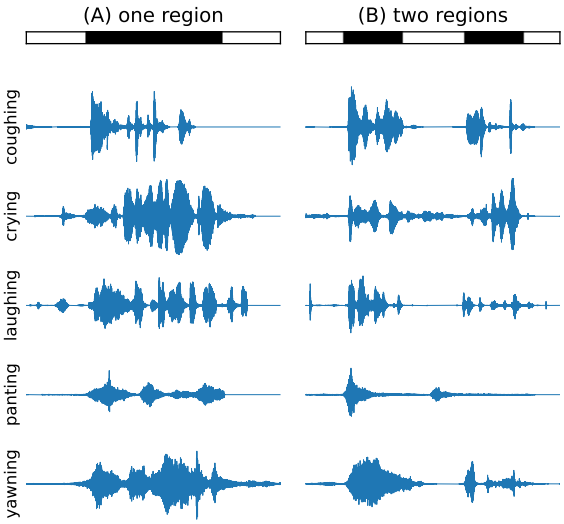

Figure 4: *Using same label shapes to generate different types of vocalization sounds.*

higher quality than the baseline, and the difference was statistically significant, as indicated by the 95% confidence interval. The DIFF system, however, yielded slightly lower quality than BASE, but the difference is not significant. Given the results presented in Fig. 3(a), we concluded that switching from global labeling to segment-based labeling does not affect the quality of generated samples. In many cases, it improves the perceived quality. For the classification questions, listeners were asked to identify the sound type. The accuracy of each system is presented in Fig. 3(b). Interestingly, the BASE system received more correct answers than the proposed systems, with SAME being significantly higher than DIFF. We can interpret this result to indicate that the BASE system can generate samples with clearer characteristics because the output samples do not have to conform to label shapes.

Table 1 contains detailed results of each vocal class of the BASE, SAME, and DIFF systems. We do not include NAT as it has high quality and accuracy scores. Among the vocal classes, coughing received the highest score for all systems, while panting received the lowest. However, this may be due to data imbalance as there were the fewest panting samples (745) and most coughing samples (2,732) out of all training classes. A counterargument is that yawning has 3,090 samples, but the perceived quality is not the highest. Thus, we conclude that the amount of data is an important aspect but not the only one.

We focused on human vocalization and set up the experiments with five vocal classes because we wanted to evaluate the sound generation system in a more nuanced manner rather than simply generating a random high-quality sample. As shown by the classification results of each vocal class, coughing was primarily mistaken for crying or laughing. Moreover, laughing/crying and panting/yawning were often confused by the listeners. Between BASE and SAME, our proposed model yielded a higher correct rate for coughing but a lower rate for crying and panting.

5. Controlling Vocalization Generation

In this section, we introduce several methods of utilizing the proposed segment-based label scheme for controlling synthetic sounds. The generated samples can be found at the associated website[4]. We used two label shapes A and B, as shown in Fig. 4, to demonstrate the effectiveness of the proposed method for controlling generation. The associated samples are also marked with either A or B shape whenever applicable.

5.1. Same shape with different labels

As the segment-based label is a sequence with 86 elements that received either 0 or the assigned value of vocalization, the values can be directly edited to manipulate the generated samples. Our scheme used a highly abstract representation for controlling audio generation, so the segment-based labels do not directly dictate the shapes of the output waveforms but act as a guide for the generation. Figure 4 shows several examples of samples of different vocalization classes when using the same label shape as the input. Depending on the vocal class, the model interpreted the same shape differently to produce the desired sounds. Generally speaking, the label inputs dictate the shape of the output waveform.

5.2. Multi-class within a single sample

Another way to utilize the segment-based label scheme is to generate multiple vocal classes within a single sample. To test this, we generated several samples on the basis of two scenarios. In the first scenario, we used the label shape A, assigned two vocal classes instead of one to the region, and attempted to generate sound samples. The results were inconsistent but still worth examining. For example, by assigning both coughing and crying, we obtained a sample reflecting both classes, but results for other combinations varied. In the second scenario, we used the label shape B and assigned one vocal class to the first region and another to the second. In this case, we generally received the expected samples. However, for several combinations, the second vocal class failed to generate any meaningful sound, suggesting that the position of the label affects the outputs due to the fixed-size nature of the model. Samples can be found at the associated website.

6. Conclusion

We have integrated a segment-based label scheme into a pre-existing sound generation model and demonstrated that controllability can be increased while still maintaining the quality of generated samples. Although the generated samples had a relatively low quality score as judged by human listeners, they received very high classification accuracy scores (81.4% for BASE and 73.8% for SAME), even though the focused vocal classes were quite similar. Controllability is the next step for the generation model as it provides an interface for human intervention. Toward this end, designing an intuitive and flexible input interface for sound generation is crucial. In future work, we will test our labeling scheme on more elaborate and powerful models such as diffusion [8] and expand our experiments to other sound classes besides human vocalization.

Acknowledgements This study is partially supported by JST CREST (JPMJCR18A6) and MEXT KAKENHI Grants (21K19808).

[4] https://soundcloud.com/xkytq59c/sets

7. References

[1] K. Collins, *Game sound: an introduction to the history, theory, and practice of video game music and sound design*. Mit Press, 2008.

[2] A. Farnell, "An introduction to procedural audio and its application in computer games," in *Audio Mostly Conference*, 2007, pp. 1–33.

[3] S. Ghose and J. J. Prevost, "Autofoley: Artificial synthesis of synchronized sound tracks for silent videos with deep learning," *IEEE Transactions on Multimedia*, vol. 23, pp. 1895–1907, 2020.

[4] A. Barahona-Rıos and S. Pauletto, "Synthesising knocking sound effects using conditional wavegan," in *17th Sound and Music Computing Conference, Online*, 2020.

[5] V. Maraev, *Who is laughing now? Laughter-infused dialogue systems*, 2022.

[6] V. Iashin and E. Rahtu, "Taming visually guided sound generation," *arXiv preprint arXiv:2110.08791*, 2021.

[7] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," *arXiv preprint arXiv:2209.15352*, 2022.

[8] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *arXiv preprint arXiv:2207.09983*, 2022.

[9] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang *et al.*, "Cogview: Mastering text-to-image generation via transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19 822–19 835, 2021.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[11] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv preprint arXiv:2302.05543*, 2023.

[12] Q. Kong, Y. Xu, T. Iqbal, Y. Cao, W. Wang, and M. D. Plumbley, "Acoustic scene generation with conditional samplernn," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 925–929.

[13] X. Liu, T. Iqbal, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Conditional sound generation using neural discrete time-frequency representation learning," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2021, pp. 1–6.

[14] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *International Conference on Learning Representations*.

[15] A. Anikin, "Soundgen: an open-source tool for synthesizing nonverbal vocalizations," *Behavior research methods*, vol. 51, pp. 778–792, 2019.

[16] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "Audiolm: a language modeling approach to audio generation," *arXiv preprint arXiv:2209.03143*, 2022.

[17] H. Mori, T. Nagata, and Y. Arimoto, "Conversational and social laughter synthesis with wavenet." in *Interspeech*, 2019, pp. 520–523.

[18] A. Anikin, "A moan of pleasure should be breathy: the effect of voice quality on the meaning of human nonverbal vocalizations," *Phonetica*, vol. 77, no. 5, pp. 327–349, 2020.

[19] J. Urbain, H. Cakmak, and T. Dutoit, "Evaluation of hmm-based laughter synthesis," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7835–7839.

[20] H.-T. Luong and J. Yamagishi, "Laughnet: synthesizing laughter utterances from waveform silhouettes and a single laughter example," *arXiv preprint arXiv:2110.04946*, 2021.

[21] K. Matsumoto, S. Hara, and M. Abe, "Controlling the strength of emotions in speech-like emotional sound generated by wavenet." in *Proc. INTERSPEECH*, 2020, pp. 3421–3425.

[22] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2405–2413.

[23] C. Cui, Y. Ren, J. Liu, R. Huang, and Z. Zhao, "Varietysound: Timbre-controllable video to sound generation via unsupervised information disentanglement," *arXiv preprint arXiv:2211.10666*, 2022.

[24] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," *arXiv preprint arXiv:2301.12661*, 2023.

[25] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[26] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," *arXiv preprint arXiv:2205.15868*, 2022.

[27] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020, pp. 1–12.

[28] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, "Pixelsnail: An improved autoregressive generative model," in *International Conference on Machine Learning*. PMLR, 2018, pp. 864–872.

[29] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017, http://dx.doi.org/10.7488/ds/1994.