ABSTRACT

Video-to-sound generation aims to generate realistic and natural sound given a video input. However, previous video-to-sound generation methods can only generate a random or average timbre without any controls of the generated sound timbre, leading to the problem that people cannot obtain the desired timbre under these methods sometimes. In this paper, we propose the task of generating sound with a specific timbre given a silent video input and a reference audio sample. To solve this task, we first use three encoders to disentangle each target sound audio into temporal, acoustic, and background information respectively, then we use a decoder to reconstruct the audio given these disentangled representations. To make the generated result achieve better quality and temporal alignment, we also adopt a mel discriminator and a temporal discriminator for the adversarial training. Our experimental results on the VAS dataset demonstrate that our method can generate high-quality audio samples with good synchronization with events in video and high timbre similarity with the reference audio. Our demos have been published on https://conferencedemos.github.io/icassp23/.

Index Terms— Audio Generation, Cross-Modality Generation, Generative Adversarial Networks, Information Disentanglement

1. INTRODUCTION

More and more people are devoting themselves to video creation, with the boom of microfilm and the rise of the short video market recently. However, many people do not produce studio-level video works due to the high cost of post-production, which has become a critical issue affecting the development of the personal video production industry. As machine learning continues to evolve in the video and audio fields, more and more AI-based software is being developed to help people create after-effects generation, background music production, video dubbing, and other post-production. Video sound effects are an exceptionally critical part of video post-production, and a fantastic sound effect can bring a better video viewing experience to the audience. The traditional methods of manual sound effects production are usually ex-

Fig. 1: Timbre Controllable Video to Sound Generation
As shown in Fig. 2, our method is a process of information disentanglement and re-fusion. We first disentangle the final audio information into three components: temporal information, timbre information, and background information, modeling them with three different encoders respectively, and then use a mel decoder to recombine these disentangled information for the reconstruction of the audio. The disruption operations (RR and EM) and the bottlenecks will break the irrelevant information and force the encoders to pass only the information that other encoders cannot supply, hence achieving the disentanglement; and with sufficient information inputs, the mel decoder could finish the reconstruction of the target mel-spectrogram, hence achieving the re-fusion. We also adopt adversarial training, which helps the model to fit the distribution of the target mel-spectrogram better and thus obtain higher quality and better temporal alignment generation results.

2.2. Training Losses

In the proposed method, we use a generative adversarial structure for training[15], including one generator $G(\cdot)$ with a time-domain alignment discriminator $D_t(\cdot)$ and a mel discriminator $D_m(\cdot)$. We denote the input video, reference audio and the generator’s generate result of our model as $v \sim p_{data}(v), m \sim q_{data}(m)$ and $G(v, m) = \hat{m} \sim q_{generated}(\hat{m})$, respectively.

The generator loss $L_G$ for training is defined by the following equation:

$$L_G(G, D_m, D_t) = \lambda_m \|\hat{m} - m\|_1 + \lambda_a L_{adv} \quad (1)$$

where $\lambda_m$ and $\lambda_a$ are hyper parameters for training and the definition of the adversarial loss $L_{adv}$ is:

$$L_{adv} = E_{v,\hat{m}}[\|1-D_t(\hat{m}|v)\|_2 + \|1-D_m(\hat{m})\|_2] \quad (2)$$

The loss to minimize when training the time-domain alignment discriminator is defined by the following equation:

$$L_{D_t}(G, D_t) = L_{real} + L_{fake} \quad (3)$$

and the definition of the positive sample loss $L_{real}$ and the negative sample loss $L_{fake}$ is:

$$L_{real} = E_{v,m}[\|1 - D_t(m|v)\|_2] \quad (4)$$

$$L_{fake} = E_{v,\hat{m},m}[\|1 - D_t(\hat{m}|v)\|_2 + \|1 - D_m(S[m]|v)\|_2] \quad (5)$$

where the $S[\cdot]$ donates the shift operation to build the negative samples with time misaligned.

It is worth noting that we use two negative samples in training the time-domain alignment discriminator: a generated samples of the generator and a real samples after time-shifting. The advantage of this approach is that we use these negative samples to increase the discriminator’s ability to discriminate generated spectra that are consistent in content but not aligned in time, thus allowing the generator to model the temporal location of voicing more accurately.

The loss of the mel discriminator is defined by the following equation:

$$L_{D_m}(G, D_m) = E_{\hat{m}}[\|1-D_m(m)\|_2 + \|D_m(\hat{m})\|_2] \quad (6)$$

Fig. 2: Information Disentanglement. $E_T$, $E_A$ and $E_B$ denotes the Temporal Encoder, the Acoustics Encoder and the Background Encoder. $RR$ denotes the random resampling transform in the Acoustics Encoder and $EM$ denotes Energy Masking.
2.3. Model Architecture

The model structure is shown in the Fig. 3(a). The generator consists of four parts, including three encoders for information disentanglement and one decoder for reconstruction. We also introduce two discriminators to improve the temporal alignment and the quality of the generated mel-spectrogram.

**Temporal Encoder** modeling the temporal location of the acoustic event from the video feature sequence. As shown in Fig. 3(b), the temporal encoder accepts video feature sequence as input and encodes hidden sequences with the same length as the video features with assist of a multi-layer 1D convolutional stack and a bidirectional LSTM layer.

**Acoustic Encoder** is proposed for encoding the timbre information which is required to reconstruct the mel-spectrogram, which consists of a multi-layer Self-Gated acoustic unit (SGAU) and a single layer of the bidirectional LSTM as shown in Fig. 3(c). Each SGAU accepts two inputs and gives two outputs with convolution and gated activation, the detailed description of the SGAU are also added to our demo page. The Random Resampling operation is added after the gated operation in SGRU, which refers to the operations of segmenting, expand-shrink transforming and random swapping of the tensor in the time sequence. We use the hidden vector of the last layer of the bidirectional LSTM as the timbre features and expand it to the same length as the temporal features in the temporal dimension.

**Background Encoder** is used to encode the acoustic independent background information required to reconstruct the mel-spectrogram. As shown in Fig. 3(d), the background encoder consists of an energy masking operation, a bidirectional LSTM layer and a linear layer. The input of the background encoder will be set to zero (i.e., silent audio) in the inference stage to get a clearer result.

**Mel Decoder** is used for reconstruct mel-spectrogram. As shown in Fig. 3(e), the outputs of the three encoders are concatenated along the time dimension and then fed to a mel-decoder to generate the final mel-spectrogram in parallel. The input hidden vector is first fed to an upsampling layer and then passed through several layers of feed-forward Transformer Block[16] to get reconstructed mel-spectrogram.

**Time-Domain Alignment Discriminator (TDAD)** is a conditional discriminator. As shown in Fig. 3(f), the discriminator receives two inputs: the mel-spectrogram as the discriminate target and the video feature sequence as the condition. In order to keep consistent with the mel-spectrogram in the time dimension, the video feature sequences are upsampled through an upsampling layer. The mel-spectrogram features are concatenated with the upsampled video features after a 1D convolution layer and fed into a multi-layer 1D convolutional structure for dimensionality reduction.

**Multi-Window Mel Discriminator** is used to discriminate the quality of the generated mel-spectrogram. The Multi-Window Mel discriminator shares the same structure in SynTaSpeech [17] which consists of 3 sub-discriminators with three 2D convolutional layers in each sub-discriminator.

3. EXPERIMENTS AND EVALUATION

This section briefly describes how our experiments were conducted and gives the results. Due to space limitations, we have published a more detailed experimental design and analysis of the results and some video demos as an appendix on the demo page¹.

3.1. Experimental Setup

We evaluate our method and model on the VAS dataset [4], which is proposed for audio-visual related machine learning tasks. VAS has a total of 12541 videos in eight categories, whose sound is highly synchronized with the visual content.

---

¹https://conferencedemos.github.io/icassp23/
We train our VarietySound on a single NVIDIA GeForce RTX 3090 GPU for 500 epochs in each category respectively, with a batch size of 48 samples. We use the AdamW optimizer [19] with \( \beta_1 = 0.5, \beta_2 = 0.999 \) and \( \epsilon = 10^{-8} \) and follow the same learning rate schedule in Regnet. During the inference, the output mel-spectrograms are transformed into waveform audio by a pre-trained HiFiGAN Vocoder [20].

We train our VarietySound on a single NVIDIA GeForce RTX 3090 GPU for 500 epochs in each category respectively, with a batch size of 48 samples. We use the AdamW optimizer [19] with \( \beta_1 = 0.5, \beta_2 = 0.999 \) and \( \epsilon = 10^{-8} \) and follow the same learning rate schedule in Regnet. During the inference, the output mel-spectrograms are transformed into waveform audio by a pre-trained HiFiGAN Vocoder [20].

### 3.2. Experiment Results and Evaluation

#### 3.2.1. Evaluation Metrics

Since there are no existing evaluation metrics for proposed tasks, we first defined a series of evaluation metrics. For the subjective evaluation, we use Mean Opinion Score (MOS) to evaluate the generated result in three areas: the quality of the generated audio, the temporal alignment with the video, and the similarity in timbre to the reference audio. For the objective evaluation, we used the cosine similarity of the timbre features extracted from third-party tools\(^2\) as an objective evaluation of the similarity in timbre.

#### 3.2.2. Evaluation Results

Through the third-party evaluation on the Amazon Mechanical Turk (AMT), we obtained the evaluation results of our model. Since we are the first to propose this task, the baseline model we choose to compare with is the cascade model of Regnet and SpeechSplit.

As shown in Table 1, the proposed model achieves a higher objective scores in terms of both audio realism and temporal alignment by comparing with the baseline model.

For the similarity of the timbre, the proposed model achieves higher scores both in the subjective and objective evaluation, which means the result of the proposed model has a timbre closer to the ground truth than the baseline model.

In summary, by obtaining the evaluation results above, we have successfully demonstrated the effectiveness of our method and model.

#### 3.2.3. Ablation Study

To demonstrate the effectiveness of all components in our model, we conducted several ablation experiments. We generate results with one encoder removed for ablation of encoders, and generate results using a re-trained model with one discriminator removed for ablation of discriminators. MCD [21] scores is use as the criterion, and a lower score stand for a better quality. As shown in Table 2, the generated samples of our model acquire the minimum MCD score on average, which has successfully demonstrated the usefulness of the three encoders for information disentanglement and the two discriminators for the adversarial training.

### 4. CONCLUSION

In this paper, for the problem of generating audio with a specific timbre for silent video, we define a new task called Timbre Controllable Video to Sound Generation (TCVSG). In addition, to accomplish this task, we propose a method based on information disentanglement, which first disentangles the target audio information into temporal, acoustic, and other background information and then re-fusion them to reconstruct the mel-spectrogram. Furthermore, according to our method, we propose a GAN-based model architecture called VarietySound, which can accept a video and a reference audio as inputs and generate high-quality, video time-aligned audio with the reference timbres. Finally, we have successfully demonstrated the effectiveness of our model and model through a series of experiments and evaluations of several aspects.

---

\(^2\)https://www.ffmpeg.org/

\(^3\)https://github.com/resemble-ai/Resemblyzer
5. REFERENCES

[1] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman, “Visually indicated sounds,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2405–2413.

[2] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg, “Visual to sound: Generating natural sound for videos in the wild,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3550–3558.

[3] Kan Chen, Chuanxi Zhang, Chen Fang, Zhaowen Wang, Trung Bui, and Ram Nevatia, “Visually indicated sound generation by perceptually optimized classification,” in Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0–0.

[4] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan, “Generating visually aligned sound from videos,” IEEE Transactions on Image Processing, vol. 29, pp. 8292–8302, 2020.

[5] Sanchita Ghose and John J Prevoist, “Foleyan: Visually guided generative adversarial network-based synchronous sound generation in silent videos,” arXiv preprint arXiv:2107.09262, 2021.

[6] Vladimir Iashin and Esa Rahtu, “Taming visually guided sound generation,” in British Machine Vision Conference (BMVC), 2021.

[7] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox, “Unsupervised speech decomposition via triple information bottleneck,” in International Conference on Machine Learning. PMLR, 2020, pp. 7836–7846.

[8] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, “Autovc: Zero-shot voice style transfer with only autoencoder loss,” in International Conference on Machine Learning. PMLR, 2019, pp. 5210–5219.

[9] Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh, “Disentangling disentanglement in variational autoencoders,” in International Conference on Machine Learning. PMLR, 2019, pp. 4402–4412.

[10] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles, “Learning to decompose and disentangle representations for video prediction,” Advances in neural information processing systems, vol. 31, 2018.

[11] Guillaume Carbajal, Julius Richter, and Timo Gerkmann, “Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement,” in 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2021, pp. 126–130.

[12] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, “Soundnet: Learning sound representations from unlabeled video,” Advances in neural information processing systems, vol. 29, pp. 892–900, 2016.

[13] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba, “Ambient sound provides supervision for visual learning,” in European conference on computer vision. Springer, 2016, pp. 801–816.

[14] Naoya Takahashi, Michael Gygli, and Luc Van Gool, “Aenet: Learning deep audio features for video analysis,” IEEE Transactions on Multimedia, vol. 20, no. 3, pp. 513–524, 2017.

[15] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, “Least squares generative adversarial networks,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2794–2802.

[16] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” in International Conference on Learning Representations, 2020.

[17] Zhenhui Ye, Zhou Zhao, Yi Ren, and Fei Wu, “Syntaspeech: Syntax-aware generative adversarial text-to-speech,” arXiv preprint arXiv:2204.11792, 2022.

[18] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in International Conference on Machine Learning. PMLR, 2015, pp. 448–456.

[19] Ilya Loshchilov and Frank Hutter, “Decoupled weight decay regularization,” in International Conference on Learning Representations, 2018.

[20] Jungil Kong, Jaehyeon Kim, and Jae young Bae, “Hifigan: Generative adversarial networks for efficient and high fidelity speech synthesis,” Advances in Neural Information Processing Systems, vol. 33, 2020.

[21] Robert Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing. IEEE, 1993, vol. 1, pp. 125–128.