THE METHODOLOGY FOR STEREO IMAGE LEARNING REPRESENTATION

Seongrae Kim[⋆] *Changwoon Choi*[†] *Sangwoo Han*[‡]

[⋆] Mechanical & Aerospace Engineering, Seoul National University
[†] Electrical & Computer Engineering, Seoul National University
[‡] Earth & Environmental Sciences, Seoul National University

ABSTRACT

Automating audio generation and synthesis is a key building block of advanced computer listening applications such as auto composition. Whilst various audio generative neural networks exist that target producing audio with both high-fidelity and global structure, far less attention has been paid to generate stereo audio, which is essential to satisfy the listeners. In this paper, we propose a simple but effective training scheme for stereo audio generation task by reparameterizing left-right channels to Mid-Side channels. To address this problem, we first introduce a novel public dataset[1] which features 480 short high-fidelity stereo audios. We also propose new representations, namely *'Side Distance'* and *'Short-time Side Distance'* that effectively capture the stereo image of stereo audio. Our results clearly show that proposed method is superior to the conventional method in generating stereo audio samples on both quantitative and qualitative evaluations.

Index Terms— Generative model, Multi-channel audio, Learning representation

1. INTRODUCTION

From classical acoustics research to the modern music industry, spatiality is one of the most important acoustical features. Classical acoustic studies deal with spatiality with parameters such as interaural cross-correlation coefficient (IACC), lateral fraction (LF), and apparent source width (ASW) [1], but in the two-channel digital audio environment most familiar to the public, spatiality is mainly discussed as stereo image. The stereo image is the perceived spatial locations of the sound sources, both laterally and in-depth, within a two-channel audio signal. Stereo image control is one of the most important tasks for satisfying modern audiences in the music industry.

With the recent development of deep learning in the field of audio, several outstanding audio generative model with the neural network, which synthesizes instrumental sounds and voices, is proposed. For example, there is a well-known WaveNet [2] that generates audio as an autoregressive manner in the time domain with a network, and GANSynth [3], which generates audio by image-learning based on audio converted to spectrogram with Generative Adversarial Network. Discussions on performance improvements for earlier audio generative models continued, but there are only studies to improve the fidelity of generated audio, such as NSF [4], no serious discussion has been conducted on multi-channel audio generation. Conventional neural audio synthesis models, including the aforementioned models, generate plausible quality audio, but only mono audio can be generated, or support multi-channel audio generation [5], but they do not consider inter-channel coherence, resulting in low-quality spatiality. Since at least a two-channel environment must be provided to satisfy the modern audience, the need for discussion of the multi-channel audio generation with plausible spatiality is emphasized.

Overall, we propose a methodology that allows a two-channel audio generative network to learn stereo images of a target audio set. We approach the problem from a channel split perspective to enable the network to learn stereo images. We apply our conjecture to an architecture based on GANSynth [3] to verify our proposed methodology. We show promising results on the stereo audio generation problem with our new dataset.

In summary, the main contributions of this work are: **1)** We propose a simple, yet effective, training scheme for stereo audio generation task. **2)** The introduction of a new public dataset composed of two-channel stereo audios that have high-fidelity and rich stereo images. **3)** We introduce a proper representation of the stereo image, namely 'Side distance' and 'STSD', and define the distance metric between them, enabling quantitative evaluation of the stereo image generation model.

2. RELATED WORK

2.1. Audio generative model

The earliest generated models for audio tend to focus on speech synthesis. In this case, models require handling variable length, and WaveNet [2] used the autoregressive method for variable-length inputs and outputs. A flow-based WaveGlow [6] has emerged that compensates for slow speed of

[1]Dataset, code and sampled audios are available on `https://github.com/changwoonchoi/ml2020`

autoregressive models. In comparison to speech, audio generation for music is relatively in the development stage. [7] proposed generation of a single musical note to use WaveNet, but it is still slow, and global latent conditioning was impossible. GANSynth [3] with Generative Adversarial Network solves these shortcomings and provides high-fidelity and locally-coherent audio represented with log magnitudes and instantaneous frequencies with sufficient frequency resolution in the spectral domain. As a result of the research to date, it is possible to learn a single note of a variety of instruments to convincingly describe the sound of real instruments and to create a variety of synthesizers by interpolation between two instruments. Despite such breakthrough development, the reason that the generated musical notes cannot be used for commercial music is that the generated result is a single channel. Despite this breakthrough, the generated musical notes cannot replace commonly used virtual instruments due to the limitation of a single channel.

2.2. Multi-channel audio

In the machine listening community, multi-channel audios are mainly used for speech separation [8], speech enhancement [9], and speech recognition [10] in the light of their ability to utilize information about speech source location. On the other hand, [11] proposed an upmixing conversion of the mono signal to pseudo-stereo in order to enhance the audio effect. However, to the best of our knowledge, there are still no attempts to generate high-fidelity multi-channel audio with a neural network. It is necessary to learn the difference while maintaining the coherency of both channels to form a spatiality of sound. A technique referred to as "mid-side coding" exploits the common part of a stereophonic input signal by encoding the sum and difference signals of the two input signals rather than the input signals themselves. ([12] described the concept of mid-side coding of using the interchannel redundancies.) Therefore, for a high-performance multi-channel audio generative model, we would like to train the state-of-the-art audio generative model([3], so-called GANSynth) through the mid-side coding to verify whether this attempt can effectively learn stereo image.

3. PROPOSED METHOD

3.1. Training scheme

As we mentioned above, channel coherency is an important property for the plausible spatiality formation in two-channel audio. However, when the network generates two-channel audio, if left and right channels are created without a guide of channel coherency, the spatiality of the generated audio will not be appropriate. Since creating additional networks associated with channel coherence caused overhead, it is recommended to find a different method to avoid burdening the network. Therefore, it is valid to make the following conjecture: Network will learn stereo image better when using the mid-side (M-S) channel than using the left-right (L-R) channel. When y is stereo audio and y_M and y_S are mid and side channel of stereo audio y, y_M and y_S as following:

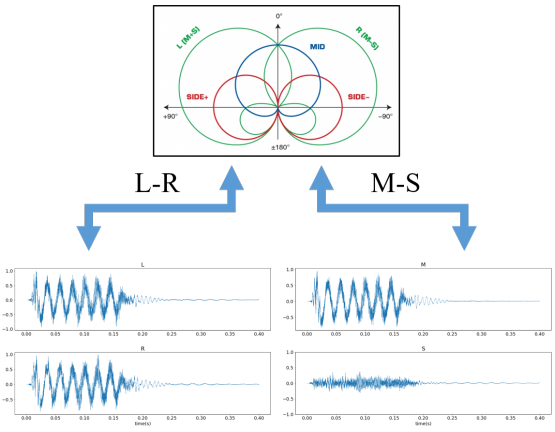


Fig. 1: Example of channel representation for two-channel audio. The left one is the L-R channel representation, and the right one is the M-S channel representation.

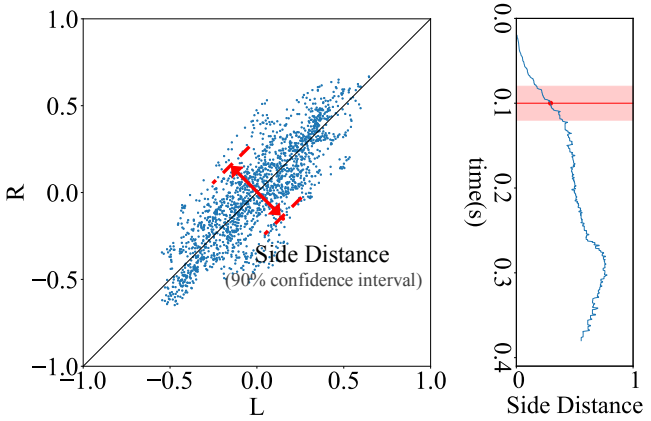
$$y_M = \frac{y_L + y_R}{2}, \quad y_S = \frac{y_L - y_R}{2}$$

where y_L and y_R are left and right channel of stereo audio y.

3.2. Custom dataset

To validate the aforementioned methodology for the stereo image learning for the network, we needed a new dataset. Since we focus on a stereo audio generation model, we needed an audio set with the drastic and various stereo images for our conjecture, but none of the conventional datasets were appropriate. Inspired by the NSynth dataset, which is mainly used by previous studies on neural audio synthesis [3, 7], we composed the new dataset using the stab, which is a single staccato note or chord that adds dramatic punctuation to a composition, used in modern electronic music. For clarity of the task, we have adjusted the length of each sample to 400ms. We configured a dataset with a sample rate of 44.1kHz and a bit depth of 16bit.

Since the lack of a large number of our sources, we performed three augmentations to obtain a sufficient amount of data. The augmentation performed are as follows: L-R channel change, time-stretching without pitch shift, and FIR filtering. We produced a total of 11,520 data that maintain the characteristics of the original data through these augmentations. Note that we did not add variations on the pitch because the data often have atonal properties.



(a) Side Distance (b) STSD

Fig. 2: Stereo audio representation. (a) shows the Lissajous figure and side distance, (b) shows the STSD.

4. EXPERIMENTS

In this section, we first propose our novel representation of stereo audio, *'Side distance' (D_{side}) and 'Short-time Side Distance (STSD)'*. Next, we introduce existing metrics for evaluating sample generation, then propose our new metrics for evaluating stereo image generation by combining existing metrics with side distance and STSD. We then compare the proposed method using M-S channels with the previous method(GANSynth [3]) which consumes L-R channels.

4.1. Representation for stereo audio

The goal of the generative model is to learn to produce samples that look similar to the ones on which it has been trained. Therefore, in order to evaluate the generative model, a proper distance metric between the generated samples and the samples from the dataset is essential. Until now, however, a metric for measuring the similarity between stereo images has never been suggested. In this section, we propose a novel representation of stereo audio that allows us to define the distance metric between them.

4.1.1. Side distance

When y is a stereo audio of length T, we define *'side distance'* D_{side} as following:

$$D_{side}(y) = \frac{\sqrt{2}}{2}\left(\max_{t\in[0,T)}[y_L(t)-y_R(t)] - \min_{t\in[0,T)}[y_L(t)-y_R(t)]\right) \tag{1}$$

where y_L and y_R are left and right channel of stereo audio y. For robustness, we use the 0.95, 0.05 quantiles of $y_L(t) - y_R(t)$ instead of maximum and minimum. The side distance can be viewed as an index indicating how far the given stereo audio is spread as visualized in Fig. 2a

4.1.2. Short-time Side Distance (STSD)

Although side distance is an indicator of how wide the given stereo audio is spread in the auditory space, it cannot express the change of the stereo image over time. Therefore, we introduce the concept of *'Short-time Side Distance (STSD)'* to capture the characteristics of the stereo image over time. The STSD is a sequence of side distance of a windowed signal. The procedure to obtain STSD is to divide a longer time signal into shorter segments of equal length and then compute side distance on each shorter segment. Formally,

$$STFT(y(t)) = D_{side}(y(\tau)w(t)) \tag{2}$$

where $w(t)$ is a window function which is nonzero for short period of time. As depicted in Fig.2b, STSD contains the change in the side distance of the windowed short fragment of signal over time.

4.2. Evaluation metrics

Following prior work, we use earth mover's distance (EMD), proposed by [13] to measure the similarity between two stereo audios' STSDs. Formally, EMD is defined as follows:

$$\text{EMD}(s_1, s_2) = \min_{\phi:s_1\to s_2} \sum_{x\in s_1} \|x - \phi(x)\|_2$$

where s_1 and s_2 are two distributions and ϕ is a bijection between them. Note that s_1 and s_2 can be any distribution. One can use STSD of stereo audio as s_1 and s_2.

Let S_g be the set of generated stereo audios and S_r be the set of reference audios with $|S_g| = |S_r|$. To evaluate generative models, we consider the three metrics, MMD, COV which are introduced by [14] and 1-NNA proposed by [15].

- **Coverage (COV)** measures the fraction of stereo audios in the reference set that are matched to at least one stereo audio in the generated set. For each stereo audio in the generated set, its nearest neighbor in the reference set is marked as a match:

  $$\text{COV}(S_g, S_r) = \frac{|\{\arg\min_{Y\in S_r} D(X,Y) | X \in S_g\}|}{|S_r|},$$

  where $D(X, Y)$ is a distance metric between two stereo audios. While coverage can detect mode collapse, it does not evaluate the quality of generated stereo audios.

- **Minimum matching distance (MMD)** is proposed to complement coverage as a metric that measures quality. For each stereo audio in the reference set, the distance to its nearest neighbor in the generated set is computed and averaged:

  $$\text{MMD}(S_g, S_r) = \frac{1}{|S_r|}\sum_{Y\in S_r} \min_{X\in S_g} D(X,Y),$$

  where $D(X, Y)$ is a distance metric between two stereo audios.

	COV (↑)	MMD (↓)	1-NNA (↓)
GANSynth [3]	34.55	1.42	76.22
Ours	**45.31**	**1.05**	**68.66**

Table 1: Quantitative comparison results of stereo audio generation on our dataset. The best results are marked in bold. Note that MMD is multiplied by 10^3.

- **1-nearest neighbor accuracy (1-NNA)** is proposed by Lopez-Paz and Oquab [15] for two-sample tests, assessing whether two distributions are identical. Let $S_{-X} = S_r \cup S_g - \{X\}$ and N_X be the nearest neighbor of X in S_{-X}. 1-NNA is the leave-one-out accuracy of the 1-NN classifier with given distance metric:

$$\text{1-NNA}(S_g, S_r) = \frac{\sum_{X \in S_g} \mathbb{1}[N_X \in S_g] + \sum_{Y \in S_r} \mathbb{1}[N_Y \in S_r]}{|S_g| + |S_r|},$$

  where $\mathbb{1}[\cdot]$ is the indicator function. For each sample, the 1-NN classifier classifies it as coming from S_r or S_g according to the label of its nearest sample. If S_g and S_r are sampled from the same distribution, the accuracy of such a classifier should converge to 50% given a sufficient number of samples. The closer the accuracy is to 50%, the more similar S_g and S_r are, and therefore the better the model is at learning the target distribution.

As a definition of COV, MMD, 1-NNA, and $D(X, Y)$ can be any distance metric between two audio samples. We use EMD(STSD(X), STSD(Y)) as a distance metric $D(X, Y)$.

4.3. Experimental results

4.3.1. Quantitative results

For a fair comparison, we trained both our model and GANSynth [3] with exactly same the hyperparameters (including network architectures, learning rate, epochs) in our training set. The only difference during the training scheme between GANSynth [3] and the proposed method is the channel encoding. Our proposed method consumes two-channel audio that is reparameterized from L-R representation to M-S representation while GANSynth [3] takes input as raw L-R represented stereo audios. The quantitative results are reported in Table 1. We observe that it is improved significantly by simply reparameterize the stereo audio in terms of every evaluation metrics (COV, MMD, 1-NNA).

4.3.2. Comparing audio quality to concurrent work

To better compare our method with concurrent work, we perform a subjective analysis over the stereo audios generated by both methods. In Fig. 3, we show the percentages of participants based on how they voted for the plausibility comparisons between ours and GANSynth [3]. The study employed 15 participants with different backgrounds - 5 audio researchers, 10 non-technical subjects. The participants are asked to choose the better audio between ours and other work. We can observe that our method received better results on both the audio researcher group and the non-technical subject group.

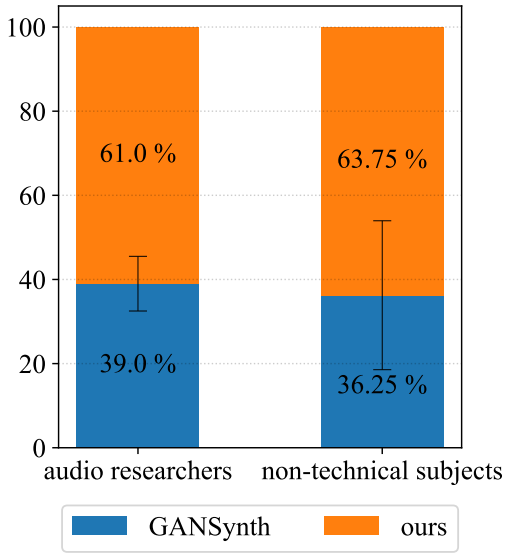


Fig. 3: Listening evaluation for audio quality to concurrent work. Our method received better results on both subject group, especially it is quite significant for audio researchers group ($p < 0.01$).

5. CONCLUSION

In this paper, we propose a novel, yet simple, training scheme for a stereo audio generative model. Also, we introduce a new dataset composed of two-channel stereo audios with rich stereo images. By reparameterizing the L-R channel into the M-S channel, the experiments on the proposed dataset demonstrate that our proposed training scheme gives promising results on every evaluation metrics (MMD, COV, 1-NNA). Furthermore, the study on human evaluations shows that our method is superior to the conventional method in terms of audio quality.

6. REFERENCES

[1] Toshiyuki Okano, Leo L Beranek, and Takayuki Hidaka, "Relations among interaural cross-correlation co-

efficient (iacc e), lateral fraction (lf e), and apparent source width (asw) in concert halls," *The Journal of the Acoustical Society of America*, vol. 104, no. 1, pp. 255–265, 1998.

[2] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," 2016.

[3] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts, "GANSynth: Adversarial neural audio synthesis," in *International Conference on Learning Representations*, 2019.

[4] Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5916–5920.

[5] Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial audio synthesis," in *International Conference on Learning Representations*, 2019.

[6] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3617–3621.

[7] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proceedings of the 34th International Conference on Machine Learning*, Doina Precup and Yee Whye Teh, Eds., International Convention Centre, Sydney, Australia, 06–11 Aug 2017, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1068–1077, PMLR.

[8] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[9] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 116–120.

[10] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5745–5749.

[11] Marco Fink, Sebastian Kraft, and Udo Zölzer, "Downmixcompatibe conversion from mono to stereo in time- and frequency-domain," in *in Proc. of the 18th Int. Conference on Digital Audio Effects*, 2015.

[12] J. D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1992, vol. 2, pp. 569–572 vol.2.

[13] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, 1998, pp. 59–66.

[14] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas J Guibas, "Learning representations and generative models for 3d point clouds," *arXiv preprint arXiv:1707.02392*, 2017.

[15] David Lopez-Paz and Maxime Oquab, "Revisiting Classifier Two-Sample Tests," *arXiv e-prints*, p. arXiv:1610.06545, Oct. 2016.

A. ADDITIONAL IMPLEMENTATION DETAILS

In this section we provide several additional details that are not provided in the main sections.

Generator	Output Size	k_{width}	k_{height}	$k_{filters}$	Nonlinearity
concat(Z, pitch)	(1, 1, 264)	-	-	-	-
conv2d	(2, 16, 256)	2	16	256	PN(LReLU)
conv2d	(2, 16, 256)	3	3	256	PN(LReLU)
upsample 2x2	(4, 32, 256)	-	-	-	-
conv2d	(4, 32, 256)	3	3	256	PN(LReLU)
conv2d	(4, 32, 256)	3	3	256	PN(LReLU)
upsample 2x2	(8, 64, 256)	-	-	-	-
conv2d	(8, 64, 256)	3	3	256	PN(LReLU)
conv2d	(8, 64, 256)	3	3	256	PN(LReLU)
upsample 2x2	(16, 128, 256)	-	-	-	-
conv2d	(16, 128, 256)	3	3	256	PN(LReLU)
conv2d	(16, 128, 256)	3	3	256	PN(LReLU)
upsample 2x2	(32, 256, 256)	-	-	-	-
conv2d	(32, 256, 128)	3	3	128	PN(LReLU)
conv2d	(32, 256, 128)	3	3	128	PN(LReLU)
upsample 2x2	(32, 512, 128)	-	-	-	-
conv2d	(32, 512, 64)	3	3	64	PN(LReLU)
conv2d	(32, 512, 64)	3	3	64	PN(LReLU)
upsample 2x2	(32, 1024, 64)	-	-	-	-
conv2d	(32, 1024, 32)	3	3	32	PN(LReLU)
conv2d	(32, 1024, 32)	3	3	32	PN(LReLU)
generator output	(32, 1024, 4)	1	1	4	Tanh
Discriminator	**Output Size**	k_{width}	k_{height}	$k_{filters}$	**Nonlinearity**
spectrogram image	(32, 1024, 4)	-	-	-	-
conv2d	(32, 1024, 32)	1	1	32	LReLU
conv2d	(32, 1024, 32)	3	3	32	LReLU
conv2d	(32, 1024, 32)	3	3	32	LReLU
downsample 1x2	(32, 512, 32)	-	-	-	-
conv2d	(32, 512, 64)	3	3	64	LReLU
conv2d	(32, 512, 64)	3	3	64	LReLU
downsample 1x2	(32, 256, 64)	-	-	-	-
conv2d	(32, 256, 128)	3	3	128	LReLU
conv2d	(32, 256, 128)	3	3	128	LReLU
downsample 2x2	(16, 128, 128)	-	-	-	-
conv2d	(16, 128, 256)	3	3	256	LReLU
conv2d	(16, 128, 256)	3	3	256	LReLU
downsample 2x2	(8, 64, 256)	-	-	-	-
conv2d	(8, 64, 256)	3	3	256	LReLU
conv2d	(8, 64, 256)	3	3	256	LReLU
downsample 2x2	(4, 32, 256)	-	-	-	-
conv2d	(4, 32, 256)	3	3	256	LReLU
conv2d	(4, 32, 256)	3	3	256	LReLU
downsample 2x2	(2, 16, 256)	-	-	-	-
concat(x, minibatch std.)	(2, 16, 257)	-	-	-	-
conv2d	(2, 16, 256)	3	3	256	LReLU
conv2d	(2, 16, 256)	3	3	256	LReLU
discriminator output	(1, 1, 1)	-	-	1	-

Table 2: Full description of our neural network architecture.

B. CONTRIBUTION OF WRITERS

Name	Task	Contribution
Seongrae Kim		-
Changwoon Choi		-
Sangwoo Han		-