

SOURCE MIXING AND SEPARATION ROBUST AUDIO STEGANOGRAPHY

Naoya Takahashi, Mayank Kumar Singh, Yuki Mitsufuji

Sony Group Corporation, Japan

ABSTRACT

Audio steganography aims at concealing secret information in carrier audio with imperceptible modification on the carrier. Although previous works addressed the robustness of concealed message recovery against distortions introduced during transmission, they do not address the robustness against aggressive editing such as mixing of other audio sources and source separation. In this work, we propose for the first time a steganography method that can embed information into individual sound sources in a mixture such as instrumental tracks in music. To this end, we propose a time-domain model and curriculum learning essential to learn to decode the concealed message from the separated sources. Experimental results show that the proposed method successfully conceals the information in an imperceptible perturbation and that the information can be correctly recovered even after mixing of other sources and separation by a source separation algorithm. Furthermore, we show that the proposed method can be applied to multiple sources simultaneously without interfering with the decoder for other sources even after the sources are mixed and separated.

Index Terms—Steganography, Watermarking, Source separation

1. INTRODUCTION

Audio steganography [1–3] is the science of concealing secret messages inside a host audio called a *carrier* in such a way that the concealment is unnoticeable to human ears. Recently, deep neural networks (DNNs) have been used as a steganographic function for hiding data inside images to achieve high *capacity* [4–7]. Kreuk et al. successfully adopted a DNN-based steganographic approach for audio by concealing a message in the short-time Fourier transform (STFT) domain while considering the distortion caused by the inverse STFT with a mismatched phase [2]. Although the method demonstrates robustness against some types of distortion that can be introduced during data transmission, the method does not assume that the carrier could be aggressively edited. Therefore, it is difficult to conceal messages in each source in a mixture, such as individual instrumental tracks in music, and recover the messages from the source-separated sounds of the mixture. In this work, we address this problem for the first time and propose a DNN-based steganographic method that works even after source mixing and separation.

Recently, as the source separation accuracy has been considerably improved owing to advances in DNN-based methods [8–15], separated sources are being widely used for many purposes such as karaoke, creating new contents using separated sources, and training models using a dataset that consists of separated sounds [16, 17]. Consequently, it is becoming more important to focus on the value and functionality of sound sources, which may be mixed with other sources and later separated by a source separation algorithm. The proposed method addresses this issue and can be used for various

applications. For instance, secret communication, the main focus of stenography, can be extended to source-wise communication where messages are concealed in sound sources and mixed with other unknown sources. Recipients who are aware of the presence of messages can decode them from individual separated sources. Sound source creators can conceal information such as musical notes of the source, captioning of the source, copyright information, or any unrelated information. In the case that copyright information is concealed for ownership protection, the method is also called watermarking. The protection of creators' rights against the abuse of separated sources is also becoming increasingly important and our proposed method addresses this problem.

In this work, we focus on musical sources. Our goal is to enable creators to conceal information in sound sources independently. Therefore, we mainly focus on the imperceptibility of the modification of the sources and the robustness against source mixing and separation. The contributions of this work are fivefold:

1. We propose concealing messages inside a source that will be mixed with other *unknown* sources and then separated by the source separation method. We call the proposed method source mixing and separation robust audio steganography (MSRAS).
2. We propose a DNN-based concealer to conceal messages in the time domain to avoid distortions caused by phase mismatch and enable simple end-to-end concealer and decoder optimization through the source separation model.
3. We further propose curriculum learning, which is shown to be essential to train the concealer and decoder through the source separation model.
4. We empirically show that MSRAS can recover a message from both unprocessed and separated sources with high accuracy while the modifications for the concealment are hardly detectable by human ears. We also show the robustness of the proposed method against other types of noise.
5. We further show that MSRAS can be applied to multiple sources in a mixture to conceal messages independently without interfering with the messages in other sources. This enables creators to hide messages in each source independently without knowing other sources that they will be mixed with.

2. RELATED WORKS

As mentioned above, steganography is closely related to watermarking. While the main goal of steganography is secret communication and it focuses on imperceptibility, watermarking focuses more on robustness and is typically used for ownership protection and verification. Various audio watermark approaches have been proposed, such as patchwork [18, 19], spread spectrum [20, 21], echo-hiding [22, 23], support vector regression [24, 25], and singular value decomposition [26]. Recently, deep-learning-based methods have been proposed for image watermarking and steganography [6].

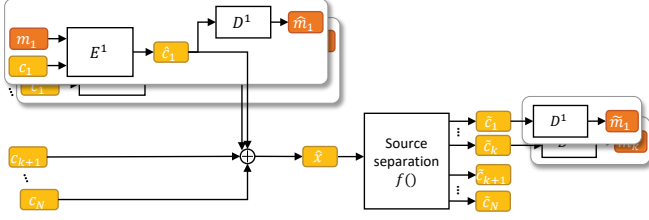


Fig. 1. Model overview.

Our work is also related to adversarial examples, where DNNs are shown to *detect* imperceptible features that can be designed to manipulate network prediction [27, 28]. Recently, adversarial examples have been explored in audio source separation [15].

3. PROPOSED METHOD

3.1. Audio steganography

The steganographic model consists of a concealer $E(c, m) = \hat{c}$, which conceals a message m inside a carrier audio signal c , and a decoder D which recovers the message from the embedded carrier as $D(\hat{c}) = \hat{m}$. The goal of the concealer and decoder is to minimize the message reconstruction error $d_m(m, \hat{m})$, while regularizing the modification of the carrier to be minimal. The model is trained by minimizing the loss function L defined as

$$L(c, m) = \lambda_c d_c(c, \hat{c}) + \lambda_m d_m(m, \hat{m}), \quad (1)$$

where λ_c and λ_m denote the weights, and $d_c(c, \hat{c})$ is the metric used to measure the similarity between the original and embedded carriers. In [2], the $l1$ norm is used for both d_c and d_m .

3.2. Incorporating source mixing and separation

We extend the method described in the previous section to enable message concealment in individual sources of a mixture so that the message can be recovered from the separated sources, as shown in Fig. 1. For simplicity, we assume that source separation is performed on the basis of the source type such as the instrument type in music. Given N sources c_1, \dots, c_N of the mixture $x = \sum_{n=1}^N c_n$, we conceal messages $m_{i \in \Gamma}$ into k of the N sources. Without loss of generality, we choose the first k sources as the ones containing messages ($\Gamma = [1, \dots, k]$). The concealer E^i takes the source c_i and message m_i as inputs and it outputs the embedded source $\hat{c}_i = E^i(c_i, m_i)$ independently from other sources. The embedded and non-embedded sources are mixed to form the mixture \hat{x} , which can be a final product to be distributed (such as music). Source separation $f(\hat{x})$ is then applied to obtain the separation \tilde{c}_i . Our goal is to recover the messages from both \hat{c}_i and \tilde{c}_i using the decoder $D^i()$. The loss function for source i becomes

$$L^{(i)} = \lambda_1^i d_c(c_i, \hat{c}_i) + \lambda_2^i d_m(m_i, \hat{m}_i) + \lambda_3^i d_m(m_i, \tilde{m}_i), \quad (2)$$

where $\tilde{m}_i = D^i(f(\hat{x})_i)$ denotes the message recovered from the i th separated source $f(\hat{x})_i$. Note that since \tilde{m} depends on all sources in the mixture \hat{x} , the loss $L^{(i)}$ is also related to other concealers $E^{j \neq i}$. One way to train all encoders and decoders is to use the summation of all losses $L^{all} = \sum_{i \in \Gamma} L^{(i)}$ universally. However, this strategy unnecessarily promotes the dependence on other unobserved sources for each concealer and failed to learn any steganographic function. Therefore, we instead use only $L^{(i)}$ to train the model for source i by freezing other concealers and alternately train the models for each

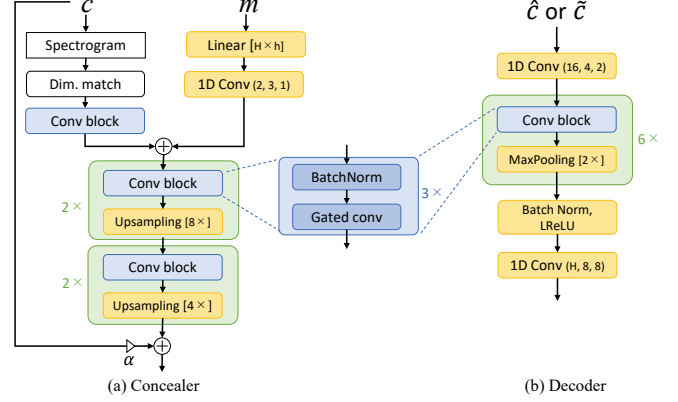


Fig. 2. TCD model architecture. Conv (o, κ, s) denotes the convolution with o filters, kernel size κ , and stride s .

source as shown in Algorithm 1. This training scheme promotes the independence of each model from other sources and avoids interference by knowing the concealment *strategies* of other concealers. This allows both the sender and the receiver of the message to work on the source without knowing the other sources.

Algorithm 1 Multi-source concealer decoder learning

```

1: for (# iteration) do
2:   sample  $c_1, \dots, c_N$  and  $m_1, \dots, m_k$ 
3:   for  $i = 1, \dots, k$  (source types) do
4:     // conceal messages and mix
5:      $\hat{x} = \sum_{j=1}^k E^j(c, m_j) + \sum_{j=k+1}^N c_j$ 
6:     // decode messages from  $i$ th embedded source and separation
7:      $\hat{m}_i = D^i(\hat{c}_i)$ ,  $\tilde{m}_i = D^i(f(\hat{x})_i)$ 
8:     freeze model parameters of  $E^{j \neq i}$ ,  $D^{j \neq i}$  and update the parameters of  $E^i$  and  $D^i$  using  $L^{(i)}$  in (2)
9:   end for
10: end for

```

3.3. Concealing message in time domain

Creuk et al. conceal and decode the message in the spectrogram domain [2], where the embedded spectrogram \hat{C} and the phase of the original carrier $\angle C$ are used to apply the inverse STFT to recover the time-domain signal. The mismatch of the magnitude and phase produces distortion in the embedded spectrogram \hat{C} . This problem is addressed by using the distorted spectrogram during the decoder training to model the distortion.

Recently, generative adversarial networks have been shown to generate a high-quality waveform from a mel-spectrogram [29]. Motivated by this work, we propose to directly conceal the message in the time domain. To this end, we propose a network architecture for a time domain concealer as shown in Fig. 2. The concealer projects the message to h -dimensional embeddings, and a 1D convolution is applied to it. The carrier (possibly multi-channel) signal is first converted to a spectrogram and the time and frequency dimensions are downsampled/interpolated to match the dimensions of the message embeddings. After a *conv block*, which comprises three stacks of Batch Norm and gated convolution [30], the output is merged with the message embeddings. Then, the conv block and upsampling layers are alternately applied four times to match the time dimension of the input waveform. The upsampling layers consist of the transposed convolution and we halve the number of channels at each layer. The

Table 1. Curriculum for model training.

Step	λ_1	λ_2	λ_3	α	d_c	# iteration
1	1	0	0	0	MDS	500
2	0.1	1	0	0	MDS	2500
3	0.1	1	1	0	MDS	5000
4	0.1	1	1	0→1	MDS	2000
5	0.0002	1	1	1	MDSR	5000

output of the final layer is then mixed with the input carrier weighted by the mixing ratio α . The decoder consists of a 1D convolution, followed by six stacks of conv blocks and max pooling layers, a Batch Norm, leaky ReLU non-linearity, and finally another 1D convolution. We refer to the proposed architecture as the time-domain concealer and decoder (TCD). In our experiment, we consider byte data or, more specifically, a sequence of H -dimensional one-hot-vectors, with a frame rate of T as the message. Therefore, the output of the decoder is fed to a softmax to obtain the posterior over the H classes. However, the method can be directly applied to other message data such as an H -band mel-spectrogram with a frame rate of T .

3.4. Curriculum learning

Since source separation is a highly nonlinear process due to it being an ill-posed problem, learning the TCD through the source separation model is extremely challenging. The models are unable to learn the steganographic function from scratch by directly using (2). To mitigate this difficulty, we propose curriculum learning as follows:

Step 1: We begin by training the concealer to generate audio that sounds similar to the carrier because the embedding signal must pass through the source separation model to enable message recovery from the separated source. To this end, we introduce an auxiliary loss function called a multi-resolution downsampled spectral (MDS) loss and use it as the metric d_c :

$$MDS(c, \hat{c}) = \sum_{m \in M} \|DS(S^m(c)) - DS(S^m(\hat{c}))\|^2, \quad (3)$$

where DS denotes the average pooling function with a kernel size of $(32, 1)$, S^m is a function used to compute a spectrogram with STFT parameter m , and $M = [1024, 2048, 4096, 8192]$ is the set of STFT window lengths. The idea of MDS is to promote a natural sound by employing a multi-resolution STFT loss [31] while allowing local spectral differences by downsampling the spectrogram. We set the mixing ratio α in the concealer to zero to avoid the trivial solution.

Step 2: Once the concealer starts to generate sounds similar to the source, we start to conceal the message. However, we first focus on the message recovery from \hat{c} only.

Step 3: After the decoder learns how to recover the message from the embedded source \hat{c} , we introduce source separation to the training criteria.

Step 4: We gradually increase α in Fig. 2 to one. This helps to minimize the perturbation from the original carrier.

Step 5: To further ensure the imperceptibility of the perturbation, we switch $d_c()$ from MDS to the multi-resolution downsampled spectrogram ratio (MDSR) defined as

$$MDSR(c, \hat{c}) = \sum_{m \in M} \|DS(S^m(c - \hat{c})) / DS(S^m(c))\|. \quad (4)$$

The MDSR promotes the ratio of the downsampled spectrogram of the carrier and the perturbation to be small. Thus, when the carrier contains high energy in a time-frequency band, the perturbation can

Table 2. SNR and accuracy of different methods.

Model	SNR [dB]		Accuracy [%]	
	\hat{c}	\tilde{c}	$D(\hat{c})$	$D(\tilde{c})$
Hide&Speak [2]	30.5	25.9	100.0	36.0
TCD trained with (1)	35.3	40.3	100.0	3.6
Hide&Speak+ $d_m(m_i, \tilde{m}_i)$	31.6	25.3	99.6	74.7
MSRAS ($\lambda_2 = 0$)	35.3	37.8	61.5	100.0
MSRAS (Proposed)	35.9	34.5	100.0	100.0

also have higher energy in that band, which approximately considers the masking effect. The curriculum is summarized in Table 1.

3.5. Increasing robustness

To improve the robustness against noise, we use three types of data augmentation during step 5 to improve the robustness of the model. (i) *Channel masking*: We randomly mask half the length of the carrier signal from one of the channels (we use stereo audio in our experiment). (ii) *Additive Noise*: We add Gaussian noise with $\sigma = 0.001$. (iii) *Random EQ*: A low-pass filter with a random cutoff frequency (15~20 kHz) is applied to half of the samples in the batch.

4. EXPERIMENTS

4.1. Setup

Experiments are conducted on the MUSDB18 dataset [32], which contains 100 and 50 songs for the *train* and *test* sets, respectively. For each song, four sources (*bass, drums, other, vocals*) and their mixture recorded in stereo format at 44.1 kHz are available. Unless otherwise noted, we use the *drum* track as the carrier to conceal the message. We consider text data as the message. We encode the 26 letters of the alphabets and one end-token to 27-dimensional one-hot-vectors every 23.2 ms; thus, the message capacity is about 43 characters per second. Randomly sampled character sequences are used for both training and testing. As the source separator, we use Demucs [13], which is an open-source music source separation library and performs separation in the time domain. We use the provided pre-trained weights without any modification. As the metric of message d_m , we use the cross-entropy loss. Models are trained on the *train* set using the Adam optimizer with a learning rate of 0.001 and a batch size of 12, and are evaluated on the *test* set.

4.2. Objective evaluation

We evaluate the proposed method by comparing with baselines. We adopt the state-of-the-art method in [2], which was originally proposed for hiding a spectrogram in another spectrogram, by adding a linear layer at the beginning of the concealer and the end of the decoder to match the dimension of the message to that of the spectrogram. We also consider three other baselines: (i) the method in [2] is extended to incorporate the loss on the message recovered from the separation expressed as (2), (ii) the proposed time-domain model, the TCD, is trained using the conventional loss (1), (iii) the proposed model is trained only on the separated source ($\lambda_2 = 0$). To evaluate the distortion of sources, we report the signal-to-noise ratio $SNR(signal, noise)$ on both the embedded source and separated source as $SNR(c, c - \hat{c})$ and $SNR(f(c), f(c) - \hat{c})$, respectively. Generally, SNR and accuracy are in the relationship of trade-off. We tune weight parameters λ such that SNR of the embedded source becomes higher than 30 dB as the perturbation becomes hardly audible around that level. The results are shown in Table 2. Both Hide&Speak [2] and the TCD achieve high accuracy with a high

Table 3. Subjective test on perceptibility of the perturbations.

Evaluated source	Accuracy [%]
Embedded source (\hat{c})	55.5
Separated source (\tilde{c})	53.9

Table 4. Comparison of different curricula.

Step	SNR [dB]		Accuracy [%]	
	\hat{c}	\tilde{c}	$D(\hat{c})$	$D(\tilde{c})$
2,3,4,5	54.1	52.1	6.3	6.3
1,3,4,5	74.3	70.9	6.3	6.3
1,2,4,5	27.2	26.3	100.0	100.0
1,2,3,5	27.4	28.2	100.0	100.0
1,2,3,4	18.3	20.9	100.0	100.0
All	35.9	34.5	100.0	100.0

SNR on the embedded source \hat{c} . However, they fail to decode the message after it is mixed with other sources and separated (\tilde{c}). By incorporating the loss on the recovered message from the separation ($d_m(m_i, \tilde{m}_i)$) into Hide&Speak [2], the decoding accuracy from the separated source is improved, but it remains significantly lower than the accuracy on $D(\hat{c})$. The proposed method achieves 100% accuracy on both the embedded and separated sources with high SNRs. This highlights the robustness of the TCD against source mixing and separation. Interestingly, when we train the TCD to recover the concealed message only from the separated source ($\lambda_2 = 0$), we obtain 61.5% accuracy on the embedded source even though 100% accuracy is obtained on the separated source. We also train Hide&Speak in the same setting of $\lambda_2 = 0$; however, it fails to learn any steganographic functions and the accuracy is the same as that of random guessing (1/27=3.7%). These results further validate the effectiveness of the proposed TCD.

Since both the proposed method and Hide&Speak baseline obtain 100% accuracy on \hat{c} , the difference is not visible. We further compare these models on a harder task by increasing the type of letters to 96. With the SNR on \hat{c} being around 36 dB, Hide&Speak and the proposed method obtain 99.3% and 98.8% accuracy on $D(\hat{c})$, and 8.1% and 97.7% on $D(\tilde{c})$, respectively. This results show that the proposed method performs competitively on the embedded sources, yet performs much more robustly on the separated sources.

4.3. Subjective detectability test

We also conduct a subjective test to evaluate the perceptibility of the perturbation. An ABX test is performed for both the embedded source \hat{c} and separated source \tilde{c} , where in case of \hat{c} , X is the original source c and either A or B is the same as X and the other is \tilde{c} , whereas in case of \tilde{c} , X is the separation of the mixture of the original sources. Subjects are asked to identify which one of A or B is the same as X, and allowed to listen to the samples many times. Forty-four audio engineers evaluate 10 samples of three second audio for each case, resulting in 440 evaluations. As shown in Table 3, the accuracy of correctly identifying the unmodified source is close to the chance rate (50%); thus, we conclude that the distortion caused by the message concealment is hardly perceptible for human ears.

4.4. Curriculum learning

Next, we show the effectiveness of the proposed curriculum learning. To this end, we omit one of the steps in the curriculum and compare the SNR and accuracy. When we omit the step, we extend

Table 5. Multiple models.

Instruments	SNR [dB]		Accuracy [%]	
	\hat{c}	\tilde{c}	$D(\hat{c})$	$D(\tilde{c})$
drums	27.9	30.8	100.0	100.0
vocals	32.3	29.9	86.8	85.6

Table 6. Robustness of models trained with and without data augmentation evaluated with different types of distortion.

Distortion	w/o D.A.	w/ D.A.
AGN ($\sigma = 0.001$)	99.9	97.2
Ch. drop	3.4	97.0
EQ	100.0	97.1
MP3 compression 320k	3.4	95.0
MP3 compression 128k	3.7	94.9

the number of iterations at the next step to match the total number of iterations for fair comparison. As shown in Table 4, steps 1 and 2 are essential for the TDC to learn the steganographic function. Omitting step 3, 4, or 5 does not deteriorate the accuracy, however, SNR decreases. Therefore, we conclude that steps 3 to 5 help to learn a highly imperceptible concealment function.

4.5. Interference from multiple models

We also test the case where multiple sources contain embedded messages. We train models for *drums* and *vocals* tracks as described in Sec. 3.2. The results are shown in Table 5. The accuracy of message recovery from *drums* separation remains high even if we include the concealer for *vocals*. Although the accuracy of *vocals* separation is relatively low compared with that of *drums*, the accuracy is similar to that on the embedded source ($D(\hat{c}) \approx D(\tilde{c})$). This shows that the messages can be concealed in multiple sources and recovered from each separated source without interfering with the other models.

4.6. Robustness against noise and edits

Finally, we investigate the robustness against different types of perturbation, namely, additive Gaussian noise (AGN), a channel drop that masks one of the channels, five-band-equalization with random gain from -3 to 3 dB (EQ), and MP3 compression. The noises are applied to the mixture \hat{x} , and the separation of the distorted mixture is tested. The accuracy of two models trained with and without the data augmentation described in Sec. 3.5 is shown in Table 6. The model trained without the data augmentation is robust against additive noise and equalization but vulnerable to data loss and signal compression. The data augmentation is shown to greatly improve the robustness against these types of noise and edits.

5. CONCLUSION

We propose audio steganography that is robust against source mixing and separation, where messages are concealed in some sources individually and the messages are recovered after mixing with other sources and separation of the mixture. To this end, we propose curriculum learning for training a time domain concealer and decoder. Experimental results confirm the effectiveness of the proposed method from different perspectives. Future works include generalizing the method to unseen separation models, testing on other domain signals such as a mixture of speeches and environmental sounds, and exploring the ability to evade steganalysis methods.

6. REFERENCES

- [1] D. C. Kar and C. J. Mulkey, "A multi-threshold based audio steganography scheme," *Journal of Information Security and Applications*, vol. 23, pp. 54–67, 2015.
- [2] F. Kreuk, Y. Adi, B. Raj, R. Singh, and J. Keshet, "Hide and speak: Towards deep neural networks for speech steganography," in *Proc. Interspeech*, 2020.
- [3] J. Wu, B. Chen, W. Luo, and Y. Fang, "Audio steganography based on iterative adversarial attacks against convolutional neural networks," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2282–2294, 2020.
- [4] S. Baluja, "Hiding images in plain sight: Deep steganography," in *Proc. NeurIPS*, 2017.
- [5] J. Hayes and G. Danezis, "Generating steganographic images via adversarial training," in *Proc. NeurIPS*, 2017.
- [6] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," in *Proc. ECCV*, 2018.
- [7] S.-P. Lu, R. Wang, T. Zhong, and P. L. Rosin, "Large-capacity image steganography based on invertible neural networks," in *Proc. CVPR*, 2021.
- [8] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. ISMIR*, 2017, pp. 745–751.
- [9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *Trans. Audio, Speech, and Language Processing*, 2019.
- [10] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. IWAENC*, 2018.
- [11] J. H. Lee, H.-S. Choi, and K. Lee, "Audio query-based music source separation," in *Proc. ISMIR*, 2019.
- [12] J.-Y. Liu and Y.-H. Yang, "Dilated convolution with dilated GRU for music source separation," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.
- [13] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.
- [14] N. Takahashi, P. Sudarsanam, N. Goswami, and Y. Mitsufuji, "Recursive speech separation for unknown number of speakers," in *Proc. Interspeech*, 2019.
- [15] N. Takahashi and Y. Mitsufuji, "Densely connected multidilated convolutional networks for dense prediction tasks," in *Proc. CVPR*, 2021.
- [16] J.-Y. Liu, Y.-H. Chen, Y.-C. Yeh, and Y.-H. Yang, "Score and lyrics-free singing voice generation," in *Proc. International Conference on Computational Creativity*, 2020.
- [17] S. Basak, S. Agarwal, S. Ganapathy, and N. Takahashi, "End-to-end lyrics recognition with voice to singing style transfer," in *Proc. ICASSP*, 2020.
- [18] N. K. Kalantari, M. A. Akhaee, S. M. Ahadi, and H. Amin-davar, "Robust multiplicative patchwork method for audio watermarking," *Trans. Audio, Speech, and Language Processing*, vol. 17, no. 6, 2009.
- [19] Y. Xiang, W. Zhou, and S. Nahavandi, "Patchwork-based audio watermarking method robust to desynchronization attacks," *Trans. Audio, Speech, and Language Processing*, vol. 22, no. 9, 2014.
- [20] A. Valizadeh and Z. J. Wang, "An improved multiplicative spread spectrum embedding scheme for data hiding," *Trans. Information Forensics Security*, vol. 7, no. 4, 2012.
- [21] Z. Liu and A. Inoue, "Audio watermarking techniques using sinusoidal pattern based on pseudorandom sequence," *Trans. Circuits System*, vol. 13, no. 8, 2003.
- [22] Y. Xiang, D. Peng, I. Natgunanathan, and W. Zhou, "Effective pseudonoise sequence and decoding function for imperceptibility and robustness enhancement in time-spread echo based audio watermarking," *Trans. Multimedia*, vol. 13, no. 1, 2011.
- [23] O. T.-C. Chen and W.-C. Wu, "Highly robust, secure, and perceptual-quality echo hiding scheme," *Trans. Audio, Speech and Language Processing*, vol. 16, no. 3, 2008.
- [24] X. Wang, W. Qi, and P. Niu, "A new adaptive digital audio watermarking based on support vector regression," *Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, 2007.
- [25] D. Lakshmi, R. Ganesh, S. R. Marni, R. Prakash, and P. Arul-mozhivarman, "SVM based effective watermarking scheme for embedding binary logo and audio signals in images," in *TENCON IEEE Region 10 Conference*, 2008.
- [26] P. K. Dhar and T. Shimamura, "Blind SVD-based audio watermarking using entropy and log-polar transformation," *Journal of Information Security and Applications*, 2015.
- [27] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. ICLR*, 2014.
- [28] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in *Proc. NeurIPS*, 2019.
- [29] K. Kumar, R. Kumar, T. de Boissiere, L. Gestein, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Proc. NeurIPS*, 2019.
- [30] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. ICML*, 2017.
- [31] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, 2020.
- [32] A. Liutkus, F.-R. Stöter, and N. Ito, "The 2018 signal separation evaluation campaign," in *Proc. LVA/ICA*, 2018.