Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Special Track on AI for Good

AudioQR: Deep Neural Audio Watermarks For QR Code

Xinghua Qu*, **Xiang Yin**, **Pengfei Wei**, **Lu Lu**, **Zejun Ma**
Speech and Audio Team, Bytedance AI Lab
{xinghua.qu, yinxiang.stephen, pengfei.wei, lulu.0314, mazejun}@bytedance.com

Abstract

Image-based quick response (QR) code is frequently used, but creates barriers for the visual impaired people. With the goal of "AI for good", this paper proposes the AudioQR, a barrier-free QR coding mechanism for the visually impaired population via deep neural audio watermarks. Previous audio watermarking approaches are mainly based on handcrafted pipelines, which is less secure and difficult to apply in large-scale scenarios. In contrast, AudioQR is the first comprehensive end-to-end pipeline that hides watermarks in audio imperceptibly and robustly. To achieve this, we jointly train an encoder and decoder, where the encoder is structured as a concatenation of transposed convolutions and multi-receptive field fusion modules. Moreover, we customize the decoder training with a stochastic data augmentation chain to make the watermarked audio robust towards different audio distortions, such as environment background, room impulse response when playing through the air, music surrounding, and Gaussian noise. Experiment results indicate that AudioQR can efficiently hide arbitrary information into audio without introducing significant perceptible difference. Our code is available at https://github.com/xinghua-qu/AudioQR.

1 Introduction

According to the report[1] of World Health Organization, there are 285 million people with visual impairment, and 39 million people being completely blind around the world. Due to the vision disability, this particular population faces many difficulties in their daily work and life. One of the most significant difficulties is scanning QR code images for the purpose of payment, identification, information retrieval, etc. This is mainly caused by the current vision based design, viz., the QR codes are displayed on images. Therefore, without external assistance, a visually impaired person may hardly realize there is QR code around, let alone accurately scan it.

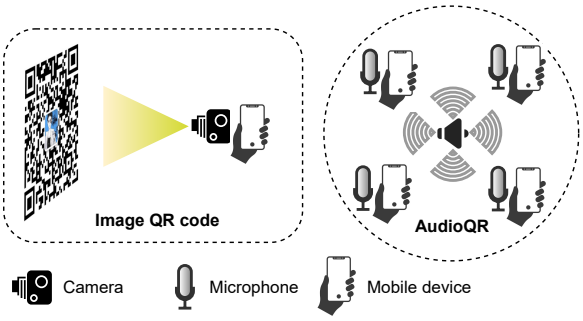


Figure 1: Comparison between image QR and AudioQR

Motivated by the principle of responsible AI, we propose the AudioQR, a new QR scanning mechanism using audios as carrier instead of images. AudioQR can embed the QR code information into an audio signal imperceptibly, and extract the QR code precisely and robustly. In this case, people (not limited to the visually impaired population) can use the microphone on the mobile to receive the audio played in the air for obtaining the embedded QR code. Compared with the current image based QR code, AudioQR has several advantages. 1) The image based QR code requires the direction alignment between camera and displayed QR code as shown in Figure 1. In contrast, AudioQR only needs the microphone to be inside the sphere neighborhood of the device that broadcasts the audio. 2) When the environment is in darkness or with inadequate lighting, the image based QR code would become hard to use, while the AudioQR will not be affected. Keeping these advantages in mind, we also notice that AudioQR is not perfect as discussed in Section 4. Therefore, our aim is not to replace/remove the current image based QR code system. Instead, we hope AudioQR together with the existing image based QR code could make the QR scanning more user-friendly and inclusive.

The key technique of AudioQR is audio watermarking that hides QR code into any audio through adding human imperceptible watermarks. Currently, the studies towards audio watermarking are mainly dominated by heuristic or handcrafted pipelines (i.e., traditional mathematical algorithms),

*Corresponding author xinghua.qu@bytedance.com

[1]https://www.emro.who.int/control-and-preventions-of-blindness-and-deafness/announcements/global-estimates-on-visual-impairment.html

which makes them less secure and not general enough, thus being hard to be deployed in large scale media platforms. Specifically, many previous approaches utilize a certain rule to hide watermark in audios. For instance, the least significant bit substitution (LSBS) [Hua *et al.*, 2016] replaces the last bits of the binary representation of waveform values. Some other approaches, such as the concatenation of discrete wavelet transform and singular value decomposition (SVD) [Al-Haj, 2014], directly operate in the frequency domain [Karajeh *et al.*, 2019]. However, for watermark extraction, these transform/decomposition based approaches usually need to save a specific key (e.g., UV matrices in SVD) for each watermarked audio, thus not being general enough. This accordingly makes them hard to be applied in large scale media platform (e.g., Tiktok) where millions of audios uploaded daily. **In contrast**, the end-to-end training pipeline for digital watermarking is more promising due to its "*one-model for all*" nature, which is well studied in computer vision tasks [Zhu *et al.*, 2018; Tancik *et al.*, 2020; Qu *et al.*, 2023] **yet not** in *speech and audio domain*. Based on a comprehensive survey [BYRNES *et al.*, 2021], there is no work exploring deep learning frameworks for end-to-end audio watermarking.

We build the AudioQR framework as an end-to-end encoder-decoder based pipeline as shown in Figure 2. Specifically, the encoder includes two modality specific encoders and one fusion encoder. To enforce the added watermark imperceptible for human being's auditory system, we train the AudioQR encoder with the losses from both time domain (i.e., l_1 norm of added watermark) and frequency domain (i.e., the l_1 distance between mel-spectrograms before and after adding the watermark). Moreover, we design a multi period QR decoder architecture with each period to handle a specific range of periodic signals from the watermarked audio. To boost the robustness of AudioQR towards real-world audio distortions (such as environment background, room impulse response, music surrounding, and Gaussian noise), we propose a stochastic distortion chain as data augmentation.

Our contributions can be summarized as bellow.

- Given the difficulties of current image based QR code system towards the population with vision impairment, we propose AudioQR, the first end-to-end audio watermarking pipeline that can imperceptibly hide QR code into arbitrary audios and recover the corresponding QR code precisely and robustly.
- We design the AudioQR encoder as two modality specific encoder and one fusion encoder, and the AudioQR decoder as a multi-period QR decoders. To achieve the robustness towards real-world audio distortions, we design a stochastic distortion chain as data augmentation.
- Experimental results based on LJSpeech, ESC-50, MusicNet453 and BUT Speech@FIT Reverb Database show that our AudioQR framework could accurately and robustly hide QR code in an imperceptible fashion.

2 Related Work

Audio watermarking has been a topic of study for more than twenty years, which predominantly employs traditional signal processing techniques [Hua *et al.*, 2016]. The first systematic study on traditional audio watermarking was conducted in the early 1990s [Cox *et al.*, 1997]. Since then, numerous audio watermarking techniques have been proposed, and they can be broadly categorized into two main groups: time-domain embedding and transform domain embedding [Hua *et al.*, 2016]. In time-domain based techniques, the watermark is directly added to the waveform using specific rules, such as least significant bit substitution (LSBS), which replaces the last binary bit of a waveform value [Chadha and Satam, 2013]. In echo hiding, the rule entails adding attenuated echoes to a carrier audio [Hua *et al.*, 2015]. While these rule-based methods are straightforward, they are not secure and ad hoc in nature. For example, if the watermark is added using LSBS, it can be easily removed by applying the same rule inversely. Similarly, in echo hiding, the echo kernels can be easily detected using cepstral analysis [Hua *et al.*, 2016], rendering the method insecure and hard to be deployed widely.

In contrast, transform domain-based audio watermarking methods consist of two steps: (1) transforming the waveform into another domain using algorithms like discrete Fourier transform (DFT) [Kang *et al.*, 2010], discrete wavelet transform (DWT) [Wang *et al.*, 2013], singular value decomposition (SVD) [Lei *et al.*, 2012], and their concatenations [Al-Haj, 2014]; and (2) embedding the watermark in the transformed domain and reconstructing the watermarked waveform using the inverse transform. While these transform domain-based methods offer improved imperceptibility and robustness compared to traditional time-domain approaches, they are still ad hoc and not generally extendable to large-scale media platforms. This is primarily due to the need to save data for inverse transforms. For example, in the SVD-based methods [Al-Haj, 2014], to enable watermark extraction, the U and V matrices must be saved for each audio slice, which can scale to billion level. Similarly, the methods based on DFT [Hua *et al.*, 2016] need to preserve the symmetric property of frequency domain samples within $[-\pi, \pi)$ to achieve reconstructed waveforms after inverse transform.

In summary, traditional watermarking algorithms still have limitations, such as difficulties in deployment on large-scale media platforms and requiring expert knowledge and experience in designing the handcrafted watermarking pipeline. In recent years, deep learning based watermark has been investigated, but they mainly focus on image-based tasks [Zhu *et al.*, 2018; Tancik *et al.*, 2020].

According to [Begum and Uddin, 2020], *there are no existing works exploring deep learning framework for end-to-end audio watermarking*. Given that, this paper is to fulfill this gap via providing *AudioQR*, *the first end-to-end pipeline for imperceptible and robust audio watermarking*. There are few reasons that make the end-to-end audio watermarking more challenging, viz., 1) the sequential data type of waveform is totally different from the images, which restricts directly applying existing machine learning based image watermarking approaches; 2) human auditory system is more complicated than visual system [Qin *et al.*, 2019]; and 3) hiding perturbation in image files are easier [Schönherr *et al.*, 2018], as images do not have temporal dependencies.

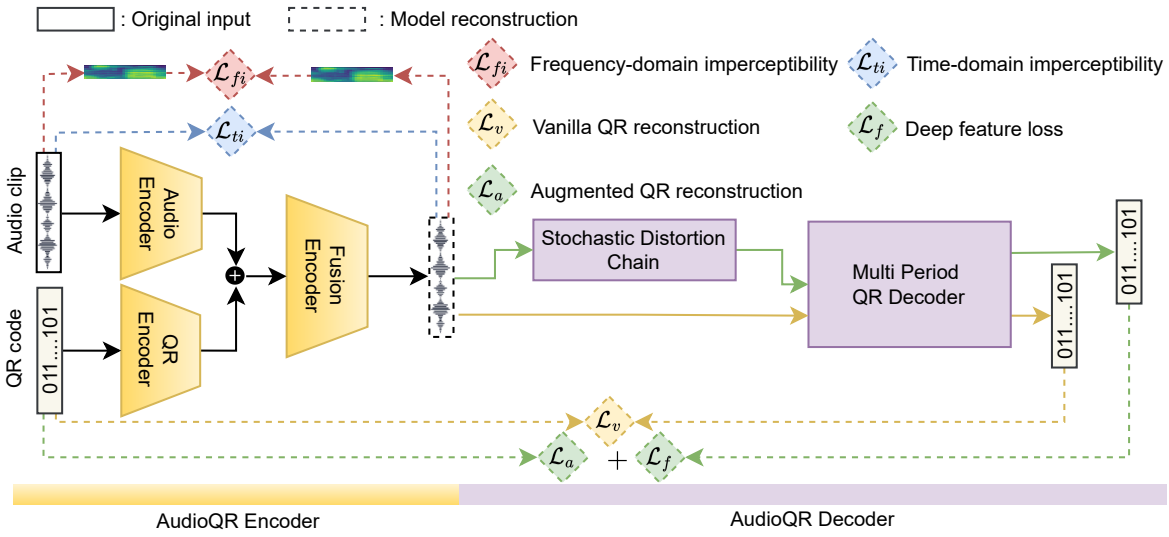


Figure 2: The training framework of AudioQR system.

3 Method

The AudioQR framework is composed of two fundamental parts, as illustrated in Figure 2: the AudioQR encoder and the AudioQR decoder. The AudioQR encoder includes three sub-components: an audio encoder, a QR encoder, and a fusion encoder. On the other hand, the decoder component consists of two distinct pathways: 1) vanilla QR decoding process through the multi-period decoder, 2) augmented decoding involving the stochastic distortion chain.

The AudioQR framework comprises several subcomponents that collectively enable the seamless integration of QR code information into audio signals. Specifically, the audio encoder and QR encoder generate encoded audio and QR code representations, respectively. The fusion encoder combines these representations into a unified feature space. The decoder component plays a critical role in retrieving the embedded QR code information, even when the audio signal is subject to various distortions. To facilitate the training of a robust decoder, the framework employs a stochastic distortion chain that simulates a range of audio perturbations. As a result, the AudioQR framework provides a comprehensive solution for embedding and retrieving QR codes in audio signals, even when confronted with diverse distortions.

3.1 AudioQR Encoder

The aim of the AudioQR encoder is to effectively encode z into a given audio signal x, where z indicates the QR message. The output of this process is the QR code embedded audio x'. The dimensions of x and z are $B \cdot L$ and $B \cdot D$, respectively, where L represents the length of the padded waveform, D represents the length of the embedded QR code, and B denotes the batch size. Both x and x' share the same dimension. The AudioQR encoder is comprised of three subcomponents: the audio encoder, the QR encoder, and the fusion encoder, which are visually represented in Figure 2 as $\mathbf{E}_A$, $\mathbf{E}_{QR}$, and $\mathbf{E}_F$, respectively. In general, the AudioQR encoder pipeline can be formulated as

$$x' = \mathbf{E}_F(\mathbf{E}_A(x) \oplus \mathbf{E}_{QR}(z)), \tag{1}$$

where $\oplus$ indicates the concatenation operation.

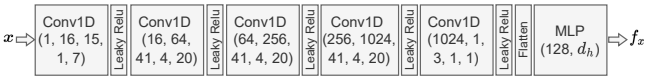


Figure 3: The architecture of the audio encoder. The Conv1D layer setting follows the format of (input channels, out channels, kernel size, stride, padding).

Audio Encoder

The audio encoder, depicted in Figure 3, is composed of five 1-dimensional convolutional layers, each employing a leaky Relu activation function, followed by a fully connected layer. The detail setting of each layer is given in Figure 3. To expedite the training process, we leverage weight normalization [Salimans and Kingma, 2016], as a reparameterization technique that disentangles the weight tensor's magnitude from its direction in each 1-dimensional convolutional layer. The input x is a sliced waveform with length 8192, which aligns with the training settings in speech synthesis, as detailed in [Kim *et al.*, 2021]. The output representation f_x generated by the audio encoder $\mathbf{E}_A$, is concatenated with the QR code representation f_z. This concatenation facilitates embedding the QR code into the audio signal.

QR Encoder

The QR encode contains two fully connected layers with Relu activations. The input of QR encoder is a multi-dimensional binary vector that carries the QR message z. The network's output is a latent representation that has the same dimension as the output produced by the audio encoder. The primary objective of the QR encode is to map the input QR message

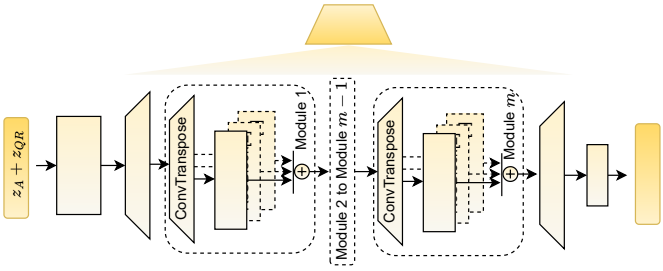


Figure 4: The neural network structure of fusion encoder $\mathbf{E}_F$

to a latent representation, allowing it to be combined with the output generated by the audio encoder. This concatenation enables the embedding of the QR message into audio signals. Overall, the QR encoder plays a crucial role in facilitating the secure transmission of information through audio signals.

Fusion Encoder

In general, the target of fusion encoder is to map the mixed representation from audio encoder and QR encoder to a waveform perturbation. The fusion encoder $\mathbf{E}_F$ is designed as a stack of convolutional layers and multi-receptive field fusion modules, as illustrated in Fig 4. This design is inspired by HiFi-GAN [Kong *et al.*, 2020], a generative adversarial network (GAN) based speech synthesis model using Mel-spectrogram. In specific, $\mathbf{E}_F$ utilizes the mixed latent representation $z_A + z_{QR}$ as input and up-samples it through transposed convolutions until the output waveform δx that shares the same dimension as its original waveform x.

In order to increase the versatility of the model to accommodate QR codes of varying dimensions, a linear projection layer is initially employed to map the QR code to a unified latent dimension. Thereby, $\mathbf{E}_F$ upsamples $z_A + z_{QR}$ for m times to match the temporal resolution of the input waveform x together with the pre- and post- convolutions. To restrict the perturbation of the output waveform to a reasonable magnitude, a weighted hyperbolic tangent activation function is applied. This is in line with the observation that the added audio watermark is usually much smaller than the original waveform. Namely, permitting the model to produce significant waveform perturbations would therefore impede training efficiency. In our setting, we set the weight as 0.1.

3.2 Multi-Period QR Decoder

The objective of the QR decoder $\mathbf{D}_{mp}$ is to effectively retrieve the QR information that has been embedded within the watermarked audio signal x'. The entire decoding process can be represented as

$$z' = \mathbb{D}_{mp}(x') \tag{2}$$

Since speech audio signals are comprised of a variety of sinusoidal signals, each with differing periods, it is necessary to identify the diverse periodic patterns that underlie the audio data. In light of this, multi-period QR decoders are proposed, in which each decoder is specialized to handle a specific range of periodic signals from the watermarked audio. By partitioning the input signal in this manner, it is possible to extract the hidden QR code more efficiently, as each decoder focuses on a distinct subset of the underlying periodic patterns presented in the watermarked audio. This results in an improved decoding performance, as the decoder architecture is specifically designed to account for the complex and varied periodicity exhibited in the watermarked audio.

More specifically, the multi-period QR decoder comprises a mixture of sub-decoders, each of which receives only equidistantly spaced samples of the watermarked audio, with the spacing controlled by the period parameter p. This design affords the ability to decode QR codes across various scales and structures. The philosophy behind here is also reflected by previous discriminator structure in GAN based TTS systems (e.g., HiFiGAN [Kong *et al.*, 2020]). More specifically, each sub-decoder is composed of a stack of strided and grouped convolutional layers with leaky ReLU activation. At the end of each sub-decoder, no activation function is applied. Subsequently, the output from multiple sub-decoders is stacked and subjected to mean pooling and a sigmoid function. Finally, the reconstructed QR code z' can be obtained.

3.3 Robustness Improvements via Stochastic Distortion Chain

The combination of the AudioQR encoders and the multi-period QR decoders forms the vanilla pipeline of QR code embedding and extraction. Nevertheless, this is not sufficient for real-world deployment since it fails to take care of the audio distortions that are typically introduced by external sources, which could have a detrimental effect on the utility of the AudioQR pipeline. In order to address this issue, we consider the threat model introduced below. The threat model entails identifying and assessing potential sources of audio distortion that may impact the effectiveness of the basic AudioQR pipeline. Thereby, we propose the stochastic distortion chain based data augmentation in order to improve its robustness towards common real-world audio distortions.

Threat Model

The threat model primarily concerns the potential distortions that could arise during the deployment of the AudioQR encoder and decoder. Essentially, the AudioQR encoder emits a unique audio signal containing a QR code, which is then extracted by the AudioQR decoder. However, during the transmission of the audio signal through the air, external factors may introduce distortions to the watermarked audio, causing variations in the extracted QR code. As a result, these distortions can have a detrimental impact on the overall accuracy and reliability of the AudioQR system. It is, therefore, important to carefully evaluate and mitigate potential distortions in order to enhance the robustness of the system.

Given this, we comprehensively consider four distinct types of audio distortions to ensure the effectiveness of our AudioQR system, including 1) environment background distortions (e.g., animal barking, urban traffic noise, machine enginer sound, external rain sound), 2) room impulse response distortions in different palces (e.g., shopping mall, living room, kitchen, coffee shop, office, hospital), 3) music surrounding distortions, and 4) Gaussian noise distortion. It is noteworthy that these four types of distortions are

closely relevant in the real-world scenarios where AudioQR systems are deployed. For instance, when a vision-impaired customer uses a mobile phone to receive the QR-embedded audio played by the bar counter in a coffee shop, the audio signal can be distorted by the surrounding noise, room acoustics, machine noise, and music played in the shop. Although there are other metrics for evaluating audio watermark robustness (e.g., mp3 compression, band pass filtering), they are not closely related to the AudioQR scenario we consider. Hence, we focus on the real-world-driven threats that are specific to the AudioQR scenario we described above. The robustness in such setting is less investigated in previous studies. Given that, we propose a stochastic distortion chain for data augmentation during training in order to improve the robustness of our AudioQR system.

Stochastic Distortion Chain

To enhance the robustness of AudioQR system for practical real-world deployment, we propose to use a stochastic distortion chain as data augmentations. Data augmentation has been widely applied in model training for different purpose, such as adversarial training [Rebuffi *et al.*, 2021] and augmented representation learning [Chen *et al.*, 2020]. In digital watermarking, there are also studies using data augmentation for enhancing the robustness toward input variations. For instance, [Qin *et al.*, 2019] utilized the data augmentation of room impulse response to simulate the audio played through the air, but their work focused on audio adversarial attacks, instead of our audio watermarking task. [Qu *et al.*, 2023] utilized the data augmentation with randomized connections of different image corruptions during the anti-forwarding watermark training. However, they are for computer vision area, thus being different from our setting in audio.

In general, the four types of audio distortions for data augmentation training can be represented as δ_e for environment background distortion, δ_r for room impulse response distortion, δ_m for music surrounding distortion, and δ_g for Gaussian noise distortion. Under the stochastic distortion chain, the probabilities of each distortion are p_e, p_r, p_m, p_g, respectively. Therefore, the decoding process of the augmented QR decoding can be represented as

$$\begin{aligned} z'_{aug} &= \mathbf{D}_{mp}(x' + \delta_{aug}) \\ \delta_{aug} &= p_e \circ \delta_e + p_r \circ \delta_r + p_m \circ \delta_m + p_g \circ \delta_g, \end{aligned} \tag{3}$$

where $p_e \circ \delta_e$ means the distortion δ_e is sampled with probability p_e. Similar meaning applies to $p_r \circ \delta_r$, $p_m \circ \delta_m$, and $p_g \circ \delta_g$. Such stochastic concatenation of different distortions is motivated by two factors: 1) the real-world audio distortions exist in the stochastic concatenation manner; 2) the stochatic combination of different distortions can increase the generalization ability of robustness.

3.4 Loss Functions

There are totally two types of loss functions for training AudioQR encoder and decoder jointly, including 1) the audio watermark imperceptibility loss functions, and 2) the QR code reconstruction loss functions.

The Audio Watermark Imperceptibility Loss

The audio watermark imperceptibility loss mainly aims to minimize the similarity between the original audio and the watermarked audio, as shown in Figure 2. Such imperceptibility is achieved via two loss functions, viz., $\mathcal{L}_{ti}$ for time-domain imperceptibility and $\mathcal{L}_{fi}$ for frequency-domain imperceptibility.

More specifically, $\mathcal{L}_{ti}$ minimizes the magnitude of the add-on watermark in time domain, viz.,

$$\mathcal{L}_{ti} = \mathbb{E}_x ||x' - x||_1. \tag{4}$$

Based on our experiments, the $L1$ distance minimization between x' and x directly on waveform is usually not enough for ensuring the imperceptibility. Given that, we involve another imperceptibility loss from frequency domain as

$$\mathcal{L}_{fi} = \mathbb{E}_x ||\phi(x') - \phi(x)||_1, \tag{5}$$

where $\phi(\cdot)$ is the function that transforms the waveform to spectrogram with Mel scale in frequency domain. Such a transformed distance between audios has also been proved to enable more efficient and stable training in previous vocoder study [Kong *et al.*, 2020] for translating audio spectrogram to waveform. In general, the spectrogram contains less noisy information than original waveform given the filtering provided by the Fourier transform. Moreover, Mel scale is also involved to better reflect the human perception range in frequency. Therefore, the distance calculation based on mel speatrogram and original waveform could provide a comprehensive measure about the imperceptibility quantification.

QR Code Reconstruction Loss

The QR reconstruction losses aim to decode the QR code from the watermarked audio x'. Given two QR decoding pipelines as shown by Figure 2, there are two different types of QR reconstruction losses, namely, $\mathcal{L}_v$ for vanilla QR reconstruction indicated by the yellow pipeline, and $\mathcal{L}_a$ for augmented reconstruction indicated by the green pipeline. In particular, $\mathcal{L}_v$ and $\mathcal{L}_a$ can be represented as

$$\begin{aligned} \mathcal{L}_v &= \mathbb{E}_{x'} ||\mathbf{D}_{mp}(x') - z||_1, \\ \mathcal{L}_a &= \mathbb{E}_{x'} ||\mathbf{D}_{mp}(x' + \delta_{aug}) - z||_1, \end{aligned} \tag{6}$$

where z is the ground truth QR message that serves as the reconstruction target. δ_{aug} is given in Eq. 3.

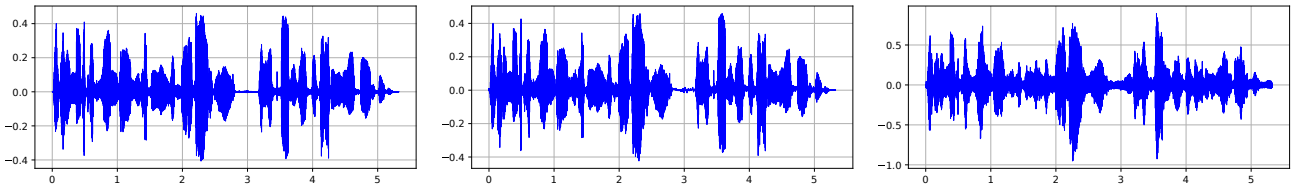
Moreover, in computer vision, to better enable imperceptibility of pixel watermark, the perceptual similarity loss based on the average learned perceptual image patch similarity (LPIPS) distance [Zhang *et al.*, 2018] is usually utilized. Motivated by so, we propose the audio deep feature loss $\mathcal{L}_f$ that is calculated in the multi-period encoder. Let $f_{x'}$ indicates the deep features from $\mathbf{D}_{mp}$ with input x'. Similarly, $f_{x'+\delta_{aug}}$ means the corresponding deep features from input $x' + \delta_{aug}$. Therefore, $\mathcal{L}_f$ can be represented as

$$\mathcal{L}_f = \mathbb{E}_{x'} ||f_{x'} - f_{x'+\delta_{aug}}||_1. \tag{7}$$

In summary, the training loss $\mathcal{L}_{total}$ for our end-to-end AudioQR system is

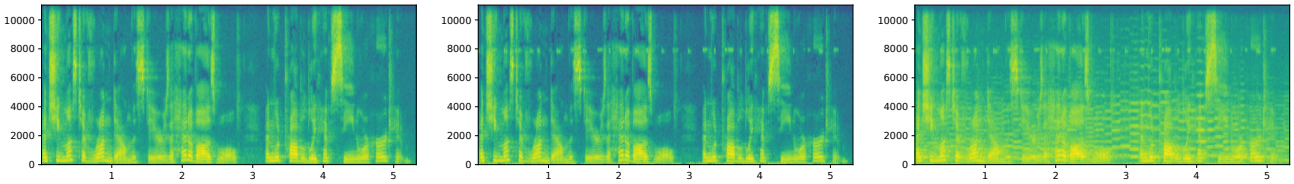
$$\mathcal{L}_{total} = \lambda_{imp}(\mathcal{L}_{ti} + \mathcal{L}_{fi}) + \lambda_{mr}(\mathcal{L}_v + \mathcal{L}_a + \mathcal{L}_f), \tag{8}$$

where λ_{imp} and λ_{mr} are weights for the watermark impermeability losses and QR reconstruction losses, respectively.



(a) The non-watermarked waveform (b) The watermarked waveform (c) The watermarked & distorted waveform

Figure 5: The comparison of waveform from the non-watermarked audio, watermarked audio, and watermarked & distorted audio.



(a) The non-watermarked spectrogram (b) The watermarked spectrogram (c) The watermarked & distorted spectrogram

Figure 6: The comparison of spectrogram from the non-watermarked audio, watermarked audio, and watermarked & distorted audio.

Algorithms	**BER**	**ACC**
DWT [Xiang, 2011]	0.00%	100%
[Wang *et al.*, 2016]	0.00%	100%
SS-SNR-HS [Su *et al.*, 2018]	0.00%	100%
[Li *et al.*, 2018]	0.12%	99.88%
[Hsu *et al.*, 2020]	1.23%	98.78%
GA-DT [Wu *et al.*, 2021]	0.10%	99.90%
AudioQR w/o Aug	0.01%	99.99%
AudioQR Aug	0.02%	99.98%

Table 1: Accuracy of audio watermark decoding.

4 Experiments

4.1 Dataset

We evaluate the effectiveness and robustness of our audio QR system on four datasets, i.e., LJSpeech, ESC-50, MusicNet and BUT Speech@FIT Reverb Database. LJSpeech dataset contains 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books. The dataset separation for training and test follows the setting in [Kim *et al.*, 2021]. In calculating the mel spectrogram, we use Short-time Fourier transform (STFT) to convert waveform to linear spectrogram first. In such STFT process, the FFT size, window size and hop size of the transform are set to 1024, 1024 and 256, respectively. Thereafter, we get 80 bands mel-scale spectrograms by introducing a mel-filterbank to linear spectrograms. The sampling rate used on LJSpeech is 22050.

In the stochastic distortion chain module, we use ESC-50 dataset [Piczak, 2015] for environmental background distortion, and MusicNet dataset for music surrounding distortion. In particular, ESC-50 dataset contains 2000 5-second environmental audio recordings for environmental sound classification. MusicNet contains 330 freely-licensed classical music recordings. Moreover, we involve BUT Speech@FIT Reverb Database to simulate different reverberations in different places, such as living room, kitchen, coffee shop, shopping mall, office, hospital.

4.2 Evaluation Metrics

We use bit error rate (BER) to represent the error of decoding QR code from watermarked audios. The corresponding accuracy (ACC) can be calculated as $100 - BER$, which indicates the percentage ratio of correctly reconstructing the pre-embedded QR code. Moreover, to evaluate the imperceptibility of the QR code embedded audios, we involve the signal to noise (SNR) ratio that is computed by

$$SNR = 10 \cdot log \frac{\sum x_i^2}{\sum (x' - x)^2}. \tag{9}$$

4.3 Implementation Settings

In our experiments, we set the input dimension of audio x to 8192. We utilize the Adam optimizer with a learning rate of $2e^{-4}$, and the exponential decay rates of moment estimates are set at 0.8 and 0.99. The loss weights λ_{imp} and λ_{mr} in Eq. 8 are set to 1. During the training process, the audio QR code capacity can be chosen flexibly based on the requirements. We opt for 50 bits for each 8192 audio clip, resulting in a capacity of around 135 per second, based on the sampling rate of 22050. However, we also find that using 100 bits per clip (resulting in the capacity of 270) would not significantly impact the performance. The probabilities p_e, p_r, p_m, p_g of the four distince audio distortions are set to be 0.75.

4.4 Accuracy Evaluation

As shown in Table 1, our well trained model could achieve 99.99% bit recovery accuracy on LJSpeech test dataset, which means that the embedded QR code can be precisely recovered. Such an ability of achieving near/equal 100% BER is also observed in many other audio watermarking algorithms, such as DWT based algorithms [Xiang, 2011; Li *et al.*, 2018], derivative free optimization [Su *et al.*, 2018;

	No Distortion	RIR	Environment Background	Music	Gaussian Noise	All concatenated
AudioQR w/o Aug	99.99%	48.26%	49.62%	49.80%	50.88%	50.31%
AudioQR with Aug	99.98%	97.84%	96.33%	98.84%	98.64%	93.50%

Table 2: The robustness evaluation based on RIR, environment background distortion, music background distortion, Gaussian noise distortion , and the concatenation of all distortions.

Algorithms	**SNR**
DWT [Xiang, 2011]	23.98 db
SS-SNR-HS [Su *et al.*, 2018]	24.94 db
[Wang *et al.*, 2016]	22.44 db
[Li *et al.*, 2018]	24.34 db
[Hsu *et al.*, 2020]	27.7 db
GA-DT [Wu *et al.*, 2021]	25.04 db
AudioQR w/o Aug	**49.50** db
AudioQR Aug	**31.84** db

Table 3: Imperceptibility (SNR) evaluation and comparison. The larger the better of SNR value.

Wu *et al.*, 2021]. However, our method is more computational efficient. In comparison, the SS-SNR-HS method [Su *et al.*, 2018] takes an average computation time of around 574 seconds per audio watermark, while our method only requires inference through the decoder network, which takes less than 1 second. Moreover, our algorithm could achieve better imperceptibility as shown below.

4.5 Imperceptibility Evaluation

We conducted an evaluation of the imperceptibility of our added QR code watermark in audios based on SNR, and compared our results with those from previous baselines, as presented in Table 3. Our methods yielded significantly better SNRs than previous baselines in general. Specifically, our AudioQR model, trained without using data augmentation, achieved an impressive 49.50 SNR. Even with data augmentation, AudioQR still achieved a respectable 31.84 SNR. In contrast, most previous baselines were only able to achieve SNRs between 20 and 30. The highest SNR value achieved by the baselines we compared against was 27.7. The human perceptible threshold [Al-Haj, 2014] is typically set at 20 dB. Given this threshold, it becomes clear that our method is significantly more imperceptible than previous baselines. When AudioQR is trained without data augmentation, the SNR is greatly larger than this threshold. Even with data augmentation, our model still achieves a relatively high SNR (i.e., 31.84 db), indicating a minimal perceptual impact on the audio quality. These results indicate the effectiveness of AudioQR in achieving high imperceptibility.

4.6 Robustness Evaluation

We evaluate the robustness of AudioQR based on the four types of audio distortions illustrated in Section 3.3. In specific, during evaluation, we apply each audio distortion and their concatenation to the watermarked audio x' with randomly selects strength from $[10db, 20db]$. The comparison between AudioQR with and without data augmentation is shown in Table 2. In general, we can easily find that with introducing the stochastic distortion chain based data augmentation, the robustness of the AudioQR model has been significantly improved. In particular, without using data augmentation, the QR code recovery accuracy decreases from 99.99% to around 50% on all kinds of distortions. In contrast, after introducing data augmentation in AudioQR training, the accuracy under distortions exceeds 93.5%. Moreover, we also note that the concatenation of all distortions impacts the accuracy of AudioQR with data augmentation most. This is reasonable because the concatenation of distortions usually brings stronger impact compared to the independent case.

4.7 Audio Watermark Analysis

To further analyze the pattern of the embedded QR watermark, we plot the waveform (Figure 5) and spectrogram (Figure 6) from three different cases, i.e., a) non-watermark, b) watermarked, and c) watermarked & distorted. Comparing cases a) and b) in both Figure 5 and Figure 6, we can easily find that the added watermark is quite imperceptible from both time domain and frequency domain. Conversely, comparing cases b) and c) in Figure 5&6, we can observe significant differenence of both waveform and spectrogram. This reflects again that AudioQR can precisely extract QR code from imperceptible watermark. Moreover, the strong distortions cannot destroy our model's such ability for extracting the imperceptible QR code embedding.

4.8 Limitations and Discussions

Although our AudioQR works efficiently with a significant better imperceptibility, there are still challenges for real-world deployment. One of the challenges is how to handle the situation when different audio QR codes are played simultaneously, since the decoder would be confused on which one should output. Such an issue also happens in image QR code scanning. An intuitive solution is to choose decoding results based on audio strength. Another challenge is about the security of deployed QR systems. Namely, a malicious party may steal the model based on input-output queries. We believe that merging the audio encoding with the audio synthesis training could significantly avoid this issue, where the data pair is not available.

5 Conclusion

This paper propose AudioQR as a new QR coding mechanism considering that the current image based QR code scanning is naturally unfriendly towards the population with vision impairment. not only precisely embed and recover any random QR code, but also behave robustly when different real-world distortions exist. Therefore, we present an end-to-end encoder-decoder based framework for audio QR scanning. We carefully design the model structure of each modules. The results showcase that AudioQR can precisely and robustly recovers the hidden QR code.

References

[Al-Haj, 2014] Ali Al-Haj. An imperceptible and robust audio watermarking algorithm. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):1–12, 2014.

[Begum and Uddin, 2020] Mahbuba Begum and Mohammad Shorif Uddin. Digital image watermarking techniques: a review. *Information*, 11(2):110, 2020.

[BYRNES *et al.*, 2021] OLIVIA BYRNES, HU WANG, CONGBO MA, MINHUI XUE, and QI WU. Data hiding with deep learning: A survey unifying digital watermarking and steganography. *J. ACM*, 1(1), 2021.

[Chadha and Satam, 2013] Ankit Chadha and Neha Satam. An efficient method for image and audio steganography using least significant bit (lsb) substitution. *International Journal of Computer Applications*, 2013.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020.

[Cox *et al.*, 1997] Ingemar J Cox, Joe Kilian, F Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, 1997.

[Hsu *et al.*, 2020] Chih-Yu Hsu, Shu-Yi Tu, Chao-Tung Yang, Ching-Lung Chang, and Shuo-Tsung Chen. Digital audio signal watermarking using minimum-energy scaling optimisation in the wavelet domain. *IET Signal Processing*, 14(10):791–802, 2020.

[Hua *et al.*, 2015] Guang Hua, Jonathan Goh, and Vrizlynn LL Thing. Time-spread echo-based audio watermarking with optimized imperceptibility and robustness. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2):227–239, 2015.

[Hua *et al.*, 2016] Guang Hua, Jiwu Huang, Yun Q Shi, Jonathan Goh, and Vrizlynn LL Thing. Twenty years of digital audio watermarking—a comprehensive review. *Signal Processing*, 128:222–242, 2016.

[Kang *et al.*, 2010] Xiangui Kang, Rui Yang, and Jiwu Huang. Geometric invariant audio watermarking based on an lcm feature. *IEEE Transactions on Multimedia*, 13(2):181–190, 2010.

[Karajeh *et al.*, 2019] Huda Karajeh, Tahani Khatib, Lama Rajab, and Mahmoud Maqableh. A robust digital audio watermarking scheme based on dwt and schur decomposition. *Multimedia Tools and Applications*, 78(13):18395–18418, 2019.

[Kim *et al.*, 2021] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning (ICML)*, pages 5530–5540. PMLR, 2021.

[Kong *et al.*, 2020] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:17022–17033, 2020.

[Lei *et al.*, 2012] Baiying Lei, Yann Soon, Feng Zhou, Zhen Li, and Haijun Lei. A robust audio watermarking scheme based on lifting wavelet transform and singular value decomposition. *Signal Processing*, 92(9):1985–2001, 2012.

[Li *et al.*, 2018] Jin-feng Li, Hong-Xia Wang, Tao Wu, Xing-ming Sun, and Qing Qian. Norm ratio-based audio watermarking scheme in dwt domain. *Multimedia Tools and Applications*, 77:14481–14497, 2018.

[Piczak, 2015] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia (MM)*, pages 1015–1018, 2015.

[Qin *et al.*, 2019] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International Conference on Machine Learning (ICML)*, pages 5231–5240. PMLR, 2019.

[Qu *et al.*, 2023] Xinghua Qu, Alvin Chan, Yew-Soon Ong, Pengfei Wei, Xiang Yin, Caishun Chen, Zhu Sun, and MA Zejun. Only for you: Deep neural anti-forwarding watermark preserves image privacy. *openreview*, 2023.

[Rebuffi *et al.*, 2021] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:29935–29948, 2021.

[Salimans and Kingma, 2016] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016.

[Schönherr *et al.*, 2018] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *arXiv preprint arXiv:1808.05665*, 2018.

[Su *et al.*, 2018] Zhaopin Su, Guofu Zhang, Feng Yue, Lejie Chang, Jianguo Jiang, and Xin Yao. Snr-constrained heuristics for optimizing the scaling parameter of robust audio watermarking. *IEEE Transactions on Multimedia*, 20(10):2631–2644, 2018.

[Tancik *et al.*, 2020] Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2126, 2020.

[Wang *et al.*, 2013] Xinkai Wang, Pengjun Wang, Peng Zhang, Shuzheng Xu, and Huazhong Yang. A norm-space, adaptive, and blind audio watermarking algorithm by discrete wavelet transform. *Signal Processing*, 93(4):913–922, 2013.

[Wang *et al.*, 2016] XY Wang, QL Shi, SM Wang, and HY Yang. A blind robust digital watermarking using invariant exponent moments. *AEU-International Journal of Electronics and Communications*, 70(4):416–426, 2016.

[Wu *et al.*, 2021] Qiuling Wu, Aiyan Qu, Dandan Huang, and Lejun Ma. Robust and blind audio watermarking scheme based on genetic algorithm in dual transform domain. *Mathematical Problems in Engineering*, 2021:1–14, 2021.

[Xiang, 2011] Shijun Xiang. Audio watermarking robust against d/a and a/d conversions. *EURASIP Journal on Advances in Signal Processing*, 2011:1–14, 2011.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.

[Zhu *et al.*, 2018] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 657–672, 2018.