Abstract—Existing reversible audio watermarking (RAW) techniques are often vulnerable to intentional or even unintentional attacks on the cover object. This article proposes a robust RAW scheme based on lattices, which is referred to as Meet-in-the-Middle Embedding (MME). In MME, the lattice quantization errors are properly scaled and added back to the quantized host signals such that the receiver can estimate the cover. Scaling factor serves as a key factor to the reversibility of MME, whose feasible range is rigorously justified. Both theoretically and experimentally, we demonstrate the superiority of MME to improved quantization index modulation (IQIM) in terms of signal-to-watermark ratio (SWR) and generalized signal-to-noise ratio (GSNR). Moreover, simulations show that MME also outperforms other state-of-the-arts in SWR, objective difference grade (ODG), and bit error rate (BER).

Index Terms—Reversible audio watermarking (RAW), lattices, signal-to-watermark ratio (SWR), bit error rate (BER).

I. INTRODUCTION

ReVERSIBLE data hiding (RDH) is a technique where a payload is embedded into host data such that the consistency of the host is perfectly preserved and the host data are restored during the extraction of the payload [1]. While data hiding is a broader concept that encompasses various techniques of concealing data within other data [2], [3], RDH specifically refers to techniques that allow the embedded data to be extracted without any loss of the original host data [4]. Nowadays, audio is widely spread on the Internet and has been one of the main digital multimedia carriers. Its associated Reversible Audio Watermarking (RAW) techniques are beneficial for archiving, transmitting, and authenticating high-quality audio data that contains metadata, secret data, and so forth [5].

Similar to other types of multimedia, a critical step of RAW is to design a proper reversible function, in which the common practice is to embed some clues of the cover object into the watermarked object, such that the cover object can be restored. Conventional RDH methods are mostly designed from the perspective of “statistics” and the reversible functions are based on expansion-shifting operations [6], such as difference expansion (DE) [7], prediction-error expansion (PEE) [8], [9], histogram shifting (HS) [10]. If the neighboring elements of the cover are not correlated, or the cover has a uniform distribution, these statistics-based approaches seem to be invalid. Moreover, the robustness and imperceptibility of the statistics-based approaches are not quite competitive.

In recent years, RDH has been investigated from the perspective of “geometry” [11], [12], [13], [14], where the reversible functions are based on quantization index modulation (QIM) [15]. These approaches are more capable of handling the floating-point numbers, as the rationale of QIM works with real-valued numbers. Ko et al. [12] developed a reversible watermarking algorithm based on nested QIM via the iterative process that reduces the normalized error. Peng et al. [13] made an improvement in RDH for one dimensional scalar-QIM by maintaining the relative distance, and extended it into $N$-dimensional region nesting model [14]. These methods are collectively referred to as improved-QIM (IQIM), as is named in [13]. It is noteworthy that applying IQIM to digital audio is straightforward, although they generally fail to meet the demand of high robustness and imperceptibility. The reason of no robustness is that the scaling factor of IQIM is fixed as $1 - \frac{1}{2^n}$. The reason of high distortion is more involved: i) The constructed difference vector is rather sub-optimal. ii) The embedding method is based on the integer lattice $\mathbb{Z}^n$, whose distortion performance is worse than other optimal lattices.

The robustness of the embedded message is also important as the cover object may go through some attacks such as additive noises. In this regard, robust RAW has been investigated in [16], [17], [18]. To be concise, Nishimura et al. [16] proposed a robust RAW scheme by combining quantization index modulation (QIM) and amplitude expansion, and also a scheme based on error-expansion of linear prediction [17] for segmental audio and G.711 speech. Liang et al. [18] proposed a robust RAW scheme based on high-order difference statistics. Nevertheless, these works overemphasize the robustness of embedded messages, but...
Finally, we make a comprehensive comparison between MME and other state-of-the-art RAW methods, and the signal-to-watermark ratio (SWR) and generalized signal-to-noise ratio (GSNR) between MME and IQIM are compared. The objective difference grade (ODG), and bit error rate (BER) are calculated. For the reversibility of MME hinges on the properly choosing the scaling factor, we derive the feasible range of the scaling factor.

Second, we analyze the performance of MME in terms of distortion and robustness. More specifically, we compare the signal-to-watermark ratio (SWR) and generalized signal-to-noise ratio (GSNR) between MME and IQIM [13].

Finally, we make a comprehensive comparison between MME and other state-of-the-art RAW methods, and the numerical simulations show that MME excels in the robustness and imperceptibility features such as SWR, GSNR, objective difference grade (ODG), and bit error rate (BER). The simulated results ideally align with our theoretical analyses.

Notation: Matrices and column vectors are denoted by uppercase and lowercase boldface letters. We write $|\cdot|$ for the floor function, and $|\cdot|$ for the absolute value.

II. PRELIMINARIES

A. QIM

The one-bit scalar-QIM is perhaps the most popular version of QIM in the data hiding community [22]. Its rationale can be explained by a one-bit embedding example in Fig. 1. Denote circle and cross positions in Fig. 1 as two sets $\Lambda_0$ and $\Lambda_1$, respectively. Given a host/cover sample $s \in \mathbb{R}$ and a one-bit message $m \in \{0, 1\}$, the watermarked value is simply moving $s$ to the nearest point in $\Lambda_0$ when $m = 0$, and to the nearest point in $\Lambda_1$ when $m = 1$. Define $Q(s) = \Delta|s/\Delta|$ with $\Delta$ being a step-size parameter. Then the embedding process can be described as

$$s_{\text{QIM}} \triangleq Q_m(s) = Q(s - d_m) + d_m, \ m \in \{0, 1\},$$

where $d_0 = -(\Delta/4)$, $d_1 = \Delta/4$, $\Lambda_0 = d_0 + \Delta \mathbb{Z}$ and $\Lambda_1 = d_1 + \Delta \mathbb{Z}$.

Assume that the transmitted $s_{\text{QIM}}$ has undergone the contamination of an additive noise term $n$, then at the receiver’s side we have: $y = s_{\text{QIM}} + n$. A minimum distance decoder is therefore given by

$$\hat{m} = \arg\min_{m \in \{0, 1\}} \left[ \min_{s \in \Lambda_m} |y - s| \right].$$

II. IQIM

Inspired by conventional QIM, a reversible version of QIM called improved-QIM (IQIM) has been proposed in [13]. Its embedding and extraction processes can be described as follows.

1) Embedding: Step i. Calculate the number of quantization interval $\gamma$ and relative distance of $s$ in the $\gamma^{\text{th}}$ quantization interval $r$ by

$$\gamma = \left[ \frac{s}{\Delta} \right], \ r = s - 2^b \gamma \Delta,$$

where $b$ is the bit length of a watermark $m$, and $\Delta$ is again the step size of quantization.

Step ii. To embed the watermark $m$, the final embedded point $s_{\text{IQIM}}$ is set as

$$s_{\text{IQIM}} = 2^b \gamma \Delta + m \Delta + \frac{r}{2^b}.$$
\[ \gamma_w \text{ and the relative distance } r_w \text{ by} \]
\[ \begin{cases} \gamma_w = \left\lfloor \frac{y\Delta}{2^b} \right\rfloor \\ r_w = y - \gamma_w\Delta \end{cases} \tag{5} \]

**Step ii.** Estimate the watermark \( \hat{m} \) by
\[ \hat{m} = \gamma_w - 2^b \left\lfloor \frac{\gamma_w}{2^b} \right\rfloor. \tag{6} \]

**Step iii.** Estimate the cover \( \hat{s} \) by
\[ \hat{s} = 2^b \left\lfloor \frac{\gamma_w}{2^b} \right\rfloor \Delta + 2^b r_w. \tag{7} \]

If \( y \) is noiseless, one may verify that \( \hat{m} = m \) and \( \hat{s} = s \). For noisy \( y \), although \( \hat{s} \) becomes inaccurate, \( \hat{m} \) may still be reliable if the noise pollution is small enough. In addition, IQIM can also be described by an \( N \)-dimensional integer set \( \mathbb{Z}^N \), which is no more than independently performing embedding and extraction in each dimension. For instance, the embedding function w.r.t. the watermark \( m \) is given by
\[
\begin{align*}
\text{IQIM} = & \beta \begin{bmatrix}
\gamma_1 \\
\vdots \\
\gamma_N
\end{bmatrix} \top + \frac{\Delta m}{\beta} \\
& + (1 - \beta) \hat{s},
\end{align*}
\tag{8}
\]

where \( \beta = 1 - \frac{1}{2^b} \) denotes the scaling factor in IQIM.

### C. Lattices

This article will use some concepts from lattices to arrive at simpler description of the algorithm and more elegant analysis. Hereby we review some basic concepts of lattices and nested lattice codes.

An \( N \)-dimensional lattice in \( \mathbb{R}^N \) is defined as \( \Lambda = \{ Gz \mid z \in \mathbb{Z}^N \} \), where the full-ranked matrix \( G \in \mathbb{R}^{N \times N} \) represents the generator matrix of \( \Lambda \). For any \( x \in \mathbb{R}^N \), finding its closest vector in a lattice \( \Lambda \) is referred to as the closest vector problem (CVP). Based on efficient algorithms to solve CVP \cite{[19]}, a nearest neighbor quantizer \( Q_\Lambda(\cdot) \) is defined as
\[ Q_\Lambda(x) = \arg\min_{\lambda \in \Lambda} \| x - \lambda \|. \tag{9} \]

Details of some optimal low-dimensional lattices for quantization are listed in Table I. Based on the CVP quantizer we also define \( \text{mod}(x, \Lambda) = x - Q_\Lambda(x) \).

The set of vectors in \( \mathbb{R}^N \) that are closer to \( \lambda \in \Lambda \) than other vectors in \( \Lambda \) is called the Voronoi region \( \mathcal{V}_\lambda \), i.e.,
\[ \mathcal{V}_\lambda = \{ x : Q_\Lambda(x) = \lambda \}. \tag{10} \]

The fundamental Voronoi region is represented as
\[ \mathcal{V}_\Lambda = \{ x : Q_\Lambda(x) = 0 \}. \tag{11} \]
and the volume of the fundamental region is
\[ \text{Vol}(\mathcal{V}_\Lambda) = \int_{\mathcal{V}_\Lambda} \text{d}x = |\text{det} G|. \]  

The packing radius \( r_{\text{pack}}(\Lambda) \) represents the radius of the largest sphere contained in \( \mathcal{V}_\Lambda \)
\[ r_{\text{pack}}(\Lambda) = \frac{1}{2} \min_{\lambda \in \Lambda \setminus \{0\}} \| \lambda \|. \]  

The covering radius \( r_{\text{cov}}(\Lambda) \) refers to the radius of the smallest sphere containing \( \mathcal{V}_\Lambda \)
\[ r_{\text{cov}}(\Lambda) = \max_{\mathbf{x} \in \mathbb{R}^N} \min_{\lambda \in \Lambda} \| \mathbf{x} - \lambda \|. \]  

In the lattice quantization, a typical and efficient method is using the nested lattices based shaping, called coset coding [19, 22]. Two lattices \( \Lambda_f \) and \( \Lambda_c \) are regarded as nested when these two lattices have a inclusion relation, i.e., \( \Lambda_c \subset \Lambda_f \). The lattice \( \Lambda_f \) is called the fine/coding lattice, and its subset \( \Lambda_c \) is called the coarse/shaping lattice. The generator matrices of \( \Lambda_c \) and \( \Lambda_f \) which are expressed by \( G_c \) and \( G_f \), respectively, have a relation
\[ G_c = G_f J, \]  

where the sub-sampling matrix \( J \) represents nesting matrix. When \( J \) is an identity matrix, it denotes a self-similar shaping [23] operation. When \( J \) is a diagonal matrix with different elements, it denotes a rectangular shaping [24] operation.

III. THE PROPOSED METHOD

A. Lattice-QIM Reformulation

Following [22, 25, 26], we reformulate QIM for arbitrary lattices and payloads from the perspective of lattice quantization and coset coding, based on which the description of the proposed RAW scheme can also be simpler.

1) Embedding: Step i. Choose a pair of nested lattices \( \Lambda_f \) and \( \Lambda_c \subset \Lambda_f \) in \( \mathbb{R}^N \). The fine lattice \( \Lambda_f \) can be decomposed as the union of \( |\text{det} J| \) cosets of the coarse lattice \( \Lambda_c \)
\[ \Lambda_f = \bigcup_{i=0}^{|\text{det} J| - 1} \Lambda_i = \bigcup_{\mathbf{d}_i \in \Lambda_f \setminus \Lambda_c} (\mathbf{d}_i + \Lambda_c), \]  

where each coset \( \Lambda_i = \Lambda_c + \mathbf{d}_i \) is a translated coarse lattice and \( \mathbf{d}_i \) is called the coset representative of \( \Lambda_i \).

Step ii. Perform labeling to associate the \( i \in \{0, 1, \ldots, |\text{det} J| - 1\} \) to the \( i \)th message vector \( \mathbf{m}_i \in \mathcal{M} \).

Step iii. Consider a cover/host vector denoted as \( s \). To embed a message vector \( \mathbf{m}_i \) over \( s \), the watermarked cover vector is set as
\[ s_{\text{QIM}} = W(s) = \Lambda_c(s) + (1 - \alpha)(s - \Lambda_c(s)). \]  

2) Decoding: Let the (possibly) noisy observation of \( s_{\text{QIM}} \) be \( y \).

Step i. Estimate the coset representative \( \hat{\mathbf{d}}_i \) of \( \Lambda_i \) by
\[ \hat{\mathbf{d}}_i = \text{mod}(Q_{\Lambda_f}(y), \Lambda_c). \]  

Step ii. Perform inverse labeling to turn \( \hat{\mathbf{d}}_i \) into the estimated message vector \( \hat{\mathbf{m}}_i \).

The extracted message vector is correct as long as \( Q_{\Lambda_f}(y) = s_{\text{MME}} \). Moreover, the information transmission rate per dimension of Lattice-QIM is simply
\[ R = \frac{1}{N} \log |\text{det} J|. \]  

B. Meet-in-the-Middle Embedding

Based on the above, we observe that there exists a difference vector \( e \) between the cover vector \( s \) and its quantized watermarked vector \( s_{\text{QIM}} \), i.e.,
\[ e = s - Q_{\Lambda_c}(s). \]  

Obviously the information about \( e \) is lost if we only use \( Q_{\Lambda_c}(s) \) as the watermarked vector.

Notice that QIM has certain error tolerance capability. If we treat \( e \) as the “beneficial noise” and add it back to \( Q_{\Lambda_c}(s) \), then the information about the cover \( s \) can be maintained, and the scheme becomes reversible. The tricky part about adding “beneficial noise” is that, they should be properly scaled to meet several demands. First, the scaled \( e \) should be small enough such that it does not go beyond the Voronoi region of the fine lattice. Second, the scaled \( e \) should not be too small to avoid exceeding the used representation accuracy of numbers. The flowchart of the algorithm is shown in Fig. 3 and explained as follows.

Definition 1 (MME Embedding). The watermarked cover vector is set as
\[ s_{\text{MME}} = W(s) = \Lambda_c(s) + (1 - \alpha)(s - \Lambda_c(s)) = \alpha Q_{\Lambda_c}(s) + (1 - \alpha)s, \]  

in which \( \alpha \) is a chosen scaling factor such that \( (1 - \alpha)(s - Q_{\Lambda_c}(s)) \in \mathcal{V}_{\Lambda_f} \).

Definition 2 (Noiseless MME Extraction). If the receiver’s side has \( s_{\text{MME}} \), the estimated cover vector can be accurately restored by
\[ \hat{s} = W^{-1}(s_{\text{MME}}) = \frac{s_{\text{MME}} - Q_{\Lambda_c}(s_{\text{MME}})}{1 - \alpha} + Q_{\Lambda_f}(s_{\text{MME}}) = \frac{1}{1 - \alpha} s_{\text{MME}} - \frac{\alpha}{1 - \alpha} Q_{\Lambda_f}(s_{\text{MME}}). \]  

In this case, the message vector can also be accurately decoded by using the decoder of QIM.

In the noisy setting, consider an additive white Gaussian noise (AWGN) channel in the form of \( y = s_{\text{MME}} + \mathbf{n} \), where the entries of \( \mathbf{n} \) admit a standard Gaussian distribution \( N(0, \sigma_n^2) \). In this case, accurately restoring the cover vector is impossible, but the message vector may be accurately decoded by using the decoder of QIM. Compared with IQIM, MME is advantageous in the following aspects.

- As shown in Figs. 4 and 5, the difference vector of MME can be confined to a small Voronoi cell (hence the name “meet-in-the-middle”) while IQIM cannot. In the distortion analysis we will shown that this brings certain performance gains.
- MME can be applied to general lattices while IQIM is only feasible for the standard integer lattice.
Fig. 3. Details of the proposed Meet-in-the-Middle Embedding method.

C. Correctness Analysis

1) Reversibility: For the embedding function (21), its definition field \( \mathcal{A} \) is noted as
\[
\mathcal{A} = \{ s | s \in \mathbb{R}^N \}. \tag{23}
\]
And the value field \( \mathcal{B} \) is defined as
\[
\mathcal{B} = \{ \lambda_{\text{MME}} \lambda_{\text{MME}} = \mathcal{W}(s), s \in \mathcal{A} \}. \tag{24}
\]

Fig. 4. Comparison of embedding process between IQIM and MME in one-dimension with different host vector \( s \). (a) \( s = 0.5 \), (b) \( s = 1.5 \).

Fig. 5. Comparison of embedding process between IQIM and MME in two-dimension with \( s = (1.61, 0.73) \).

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Hereby we introduce a theorem about the one-to-one correspondence between the cover and the watermarked cover.

**Theorem 1 (Noiseless Recovery):** The function \( \mathcal{W} \) in MME is bijective.

*Proof:* It suffices to prove that \( \mathcal{W} \) is both injective and surjective. “Injective” means no two elements in the domain of the function gets mapped to the same image, i.e., for any cover vectors \( s_a, s_b \in A \),

\[
s_a \neq s_b \Rightarrow \mathcal{W}(s_a) \neq \mathcal{W}(s_b).
\]

With \( s_a \neq s_b \), we analyze the relationship between their output \( \mathcal{W}(s_a) \) and \( \mathcal{W}(s_b) \) in three cases:

1. \( s_a \) and \( s_b \) belong to two different Voronoi regions \( V_{\Lambda_a} \) and \( V_{\Lambda_b} \) of the coarse lattice \( \Lambda_c \). It follows from \( V_{\Lambda_a} \cap V_{\Lambda_b} = \emptyset \) that \( \mathcal{W}(s_a) \neq \mathcal{W}(s_b) \).
2. \( s_a \) and \( s_b \) belong to the same Voronoi region in \( \Lambda_c \) but different watermarks \( (\Lambda_a, (s_a)) \neq (\Lambda_b, (s_b)) \). Since the “self-noise” does not go beyond the Voronoi cell of a fine lattice, i.e., \( (1 - \alpha)(s - Q_{\Lambda_a}(s)) \in V_{\Lambda_f} \), then obviously \( \mathcal{W}(s_a) \neq \mathcal{W}(s_b) \).
3. \( s_a \) and \( s_b \) belong to the same Voronoi region in \( \Lambda_c \) and watermark \( (\Lambda_a, (s_a)) = (\Lambda_b, (s_b)) \). Since \( s_a \neq s_b \), one has \( (1 - \alpha)s_a \neq (1 - \alpha)s_b \), and they are still unequal after both adding \( Q_{\Lambda_a}(s) \).

Combining the above three cases, the injection is proved.

“Surjective” means that any element in the range of the function is hit by the function. We prove this by using contradiction. If \( \mathcal{W} \) is not surjective, i.e., \( \exists s_{\text{MME}} \in B \), \( \forall s \in A \), \( s.t. \mathcal{W}(s) \neq s_{\text{MME}} \), then we obtain \( s_{\text{MME}} \notin B \) which contradicts (24).

The above theorem says that \( \mathcal{W} \) is a reversible function, where each element of one set is paired with exactly one element of the other set. Thus the correctness of the estimated covers is guaranteed.

In the noisy scenario of \( n \neq 0 \) in \( y = s_{\text{MME}} + n \), we argue that the cover vector is also approximately reversible if \( Q_{\Lambda_f}(\hat{y}) = 0 \), where

\[
\hat{y} = (1 - \alpha)e + n
\]

is referred to as the composite noise vector at the receiver’s side.

To be concise, by using the estimation function

\[
\hat{s}_{\text{noisy}} = \mathcal{W}^{-1}(y) = \frac{y - Q_{\Lambda_f}(y)}{1 - \alpha} + Q_{\Lambda_f}(y),
\]

we have

\[
\hat{s}_{\text{noisy}} - s = \frac{1}{1 - \alpha}n.
\]

It says that, if the composite noise \( \hat{y} \) is small enough to ensure the correctness of the message, then the distance from the noisy estimated cover and the clean cover is only \( \|1/(1 - \alpha)n\| \).

2) **Correctness of the Estimated Messages:** We have the following theorem about the robustness of the watermarking scheme. Since the “self-noise” is within the boundary of the Voronoi region, the estimated messages in the noiseless setting is correct for sure. More generally in the noisy setting, if the composite noise is small, the correctness of the estimated messages is also guaranteed.

**Theorem 2:** If \( \hat{y} \) satisfies

\[
Q_{\Lambda_f}(\hat{y}) = 0,
\]

then the estimated message by using MME is correct.

*Proof:* At the receiver’s side, the noisy observation can be written as \( y = Q_{\Lambda_f}(s) + \hat{y} \). Since \( Q_{\Lambda_f}(s) \in \Lambda_f = \cup_{d_i \in \Lambda_f} (d_i + \Lambda_c) \), \( Q_{\Lambda_f}(y) \) can be written as

\[
Q_{\Lambda_f}(y) = Q_{\Lambda_f}(Q_{\Lambda_f}(s)) + Q_{\Lambda_f}(\hat{y}).
\]

Then we have

\[
\begin{align*}
\text{mod}(Q_{\Lambda_f}(y), \Lambda_c) &= \text{mod}(\text{mod}(Q_{\Lambda_f}(Q_{\Lambda_f}(s)), \Lambda_c) + \text{mod}(Q_{\Lambda_f}(\hat{y}), \Lambda_c), \Lambda_c) \\
&= \text{mod}(d_i + Q_{\Lambda_f}(\hat{y}), \Lambda_c).
\end{align*}
\]

Thus \( \text{mod}(Q_{\Lambda_f}(y), \Lambda_c) = d_i \) if \( Q_{\Lambda_f}(\hat{y}) = 0 \). Since each coset representative \( d_i \) corresponds a unique message \( m_i \), the lemma is proved.

**D. Setting the Scaling Factor**

In this subsection, we will discuss the feasible range of \( \alpha \).

1) **Lower Bound:** The lower bound of \( \alpha \) is determined by ensuring the correctness of estimated messages. As mentioned in Definition 1, one should have \( (1 - \alpha)(s - Q_{\Lambda_a}(s)) \in V_{\Lambda_f} \). This implies that

\[
(1 - \alpha)V_{\Lambda_a} \subseteq V_{\Lambda_f},
\]

It suffices to have

\[
(1 - \alpha)r_{\text{cov}}(\Lambda_c) \leq r_{\text{pack}}(\Lambda_f),
\]

which means

\[
\alpha \geq 1 - \frac{r_{\text{pack}}(\Lambda_f)}{r_{\text{cov}}(\Lambda_c)}.
\]

For the special case of self similar shaping, a tighter bound can be derived. To be precise, let \( \Lambda_c = \Gamma \Lambda_f \). It follows from \( (1 - \alpha)\Gamma V_{\Lambda_f} \subseteq V_{\Lambda_f} \) that

\[
\alpha \geq 1 - 1/\Gamma.
\]

Fig. 6 depicts the distribution of the watermarked cover vector \( s_{\text{MME}} \) based on the general bound in (34) and the special-case bound in (35).

2) **Upper Bound:** The upper bound of \( \alpha \) is related to the correct extraction of the covers. Recall that the audio data is stored as integer for conventional Pulse-code modulation (PCM) and single precision floating-point numbers for the latest 32-bit floating format audio [27], and it cannot represent arbitrary real-valued numbers without loss of accuracy. Overflow occurs when the magnitude of a number exceeds the range allowed by the size of the bit field. To avoid the effect of overflow, we need to have

\[
\alpha \leq 1 - 2^{-(L+1)},
\]
where \( L \) denotes the bit length. For example, we have \( \alpha \leq 1 - 2^{-24} \) when using 32-bit floating format audio, and \( \alpha \leq 1 - 2^{-17} \) when using PCM.

IV. PERFORMANCE METRICS

A. Embedding Distortion

To evaluate the embedding distortion, we adopt the signal to watermark ratio (SWR) metric defined as

\[
\text{SWR} (\text{dB}) = 10 \times \log \left( \frac{\sigma_w^2}{\sigma_w^2} \right),
\]

(37)

where \( \sigma_h^2, \sigma_w^2 \) represent the power of the host and the additive watermark, respectively. For MME, we have

\[
\sigma_w^2 = E(\|w\|^2) = \alpha^2 E(\|e\|^2) = \alpha^2 N G(\Lambda_c) \text{Vol}(V_{\Lambda_c})^{2/N},
\]

(38)

in which \( G(\Lambda_c) \) represents the normalized second moment of the chosen coarse lattice \( \Lambda_c \), and (38) is derived from the widely adopted flat-host assumption [28]. For the simplest setting of \( \Lambda_f = \Delta Z^N \) and \( \Lambda_c = \Delta^2 2^b Z^N \), since \( G(\Delta^2 2^b Z^N) = 1/12 \), we have

\[
\sigma_w^2 = \frac{N \sigma_e^2 2^{2b} \Delta^2}{12}.
\]

(39)

**Proposition 1:** By setting \( \Lambda_f = \Delta Z^N \), \( \Lambda_c = \Delta^2 2^b Z^N \), and \( \alpha = \beta \), MME has a larger SWR than IQIM.

**Proof:** Based on the inequality that

\[
x - 1 < |x| \leq x,
\]

(40)

the embedding distortion of IQIM can be bounded as

\[
\sigma_{w,\text{IQIM}}^2 = E \left( \beta^2 \left\| s - \left[ 2^b \Delta [\gamma_1 \cdots \gamma_N] + \frac{\Delta m}{\beta} \right] \right\|^2 \right)
\]

\[
= \beta^2 \frac{N}{2^b} \sum_{q=1}^{2^b} \sum_{m_i=q=0}^{2^b-1} \int_0^{2^b \Delta} \left[ 2^b \Delta q + \frac{m_i q \Delta}{\beta} - q \right]^2 f(s_q) ds_q
\]

\[
> \frac{\beta^2}{2^b} \sum_{q=1}^{N} \sum_{m_i=q=0}^{2^b-1} \int_0^{2^b \Delta} \left[ m_i q \Delta - 2^b \Delta \right]^2 f(s_q) ds_q
\]

\[
= \frac{(2^b)^2 (2 \cdot 2^b - 1)}{6 (2^b - 1)} N \beta^2 \Delta^2.
\]

(41)

Letting \( \alpha = \beta \) in (39), we have

\[
\sigma_{w,\text{IQIM}}^2 - \sigma_w^2 > \frac{\beta^2}{2^b} \frac{N \beta^2 \Delta^2}{12 (2^b - 1)} > 0.
\]

(42)

As IQIM has a larger denominator in SWR, the proposition is proved.

**Proposition 1** justifies the advantage of MME over IQIM: by using a better meet-in-the-middle approach to construct the “beneficial noise”, it enjoys a larger SWR.

B. Robustness of Messages

In addition to the distortion/imperceptibility metric, it is also necessary to evaluate the robustness of MME against additive noises. Following [22], we define the generalized signal-to-noise ratio (GSNR) as

\[
\text{GSNR} = \frac{4r^2 \text{pack}(\Lambda_f)}{E(\|\hat{w}\|^2)},
\]

(43)

where \( \hat{w} \) is the composite noise vector defined in (26).

Due to the independence of \( e \) and \( n \), we have

\[
E(\|\hat{w}\|^2) = E(\|\delta - \hat{d}\|^2)
\]

\[
= (1 - \alpha)^2 E(\|e\|^2) + E(\|n\|^2)
\]

\[
= (1 - \alpha)^2 N G(\Lambda_c) \text{Vol}(V_{\Lambda_c})^{2/N} + N \sigma_n^2.
\]

(44)

By substituting (44) into (43), the GSNR of MME can be expressed as

\[
\text{GSNR} = \frac{4r^2 \text{pack}(\Lambda_f)}{(1 - \alpha)^2 N G(\Lambda_c) \text{Vol}(V_{\Lambda_c})^{2/N} + N \sigma_n^2}.
\]

(45)

Since the energy of the self-noise \((1 - \alpha)e\) is smaller than that of IQIM, its composite-noise also features a smaller energy. Therefore, MME has a larger GSNR than IQIM.

V. SIMULATIONS

We carry out some simulations in this section to verify the effectiveness of MME. The simulation setups are summarized as follows.

**Datasets:** In the experiments, two online datasets [30] and [31] are adopted, which contain ringtone, natural sound, absolute music, songs, and speeches. They consist of 5 and 70 bipolar 16-bits Waveform Audio File Format (WAV) files, respectively, which are labeled as No. 1001-1005 and 2001-2070. The cover vectors are generated from these datasets, and the messages are generated from uniform random distributions.

**Distortion Metrics:** In addition to the SWR metric, the objective difference grade (ODG) [29] is also introduced in the...
TABLE II
THE LEVEL OF OBJECTIVE DIFFERENCE GRADE SCORES [29]

| ODG | Quality   | Impairment            |
|-----|------------|-----------------------|
| 0   | Excellent  | Imperceptible         |
| -1  | Good       | Perceptible but not annoying |
| -2  | Fair       | Slightly annoying     |
| -3  | Poor       | Annoying              |
| -4  | Bad        | Very annoying         |

experiments, which is a popular metric to assess accurately the objective auditory perception and is produced by the Perceptual Evaluation of Audio Quality (PEAQ) measure standardized in the ITU-R BS. 1387 [32]. The ODG scores are computed by the software downloaded from [33], which is the Matlab version of Evaluation of Audio Quality (EAQUAL) [34]. ODG gives a score in the range of [−4.5, 0.5]. As shown in Table II, the audio quality is said to be excellent, good, fair, poor, and bad based on the respective ODG scores.

Robustness Metrics: In addition to the GSNR metric, we also introduce the bit error rate (BER) metric to exactly show the decoding error rates. BER is defined as

\[ BER = \frac{\sum_{k=1}^{M} \sum_{q=1}^{N} (x_{k,q} \oplus x'_{k,q})}{MN}, \]  

where \( x \) and \( x' \) separately represent the original and estimated messages, while \( q \) and \( k \) are dimensional and sample identifiers respectively.

Benchmarks and Settings: The state-of-the-art methods in [13], [18] and [16] are adopted as the benchmarks, which are respectively named as “IQIM”, “Liang” and “Nishimura”. For IQIM, Nishimura and the proposed MME, unless stated otherwise, we set \( \alpha = 0.6569, \Delta = 2000, \Lambda_f = \Delta Z^2 \), and \( \Lambda_c = \Delta 2^R Z^2 \). In addition, “Liang” is configured with parameters \( S = 300, T = 7500, G = 20000 \). The above parameters are chosen to achieve the best distortion-robustness tradeoff. The effects of different parameters would be demonstrated in Section V-B.

A. Imperceptibility Performance

Figs. 7(a) and (b) have respectively depicted the average value of SWR and ODG for the whole dataset. The blue and black bars represent the confidence interval (CI) and the range of standard deviation (SD). Fig. 7(a) shows that the proposed method attains the highest SWR of 32.9891 dB, which is better than the 31.6769 dB of Liang, and the 28.3758 dB of IQIM. Moreover, the Nishimura method fails to work in this setting. Regarding ODG, the proposed method, IQIM and Nishimura work well, and they all belong to the class of QIM-based methods. More in-depth comparisons are made with the most related IQIM method hereby. Fig. 8 shows the performance of SWR and ODG with different code rate \( R \). We have the following observations: i) For the SWR performance, MME is strictly better than IQIM regardless of the code rate \( R \). ii) In terms of ODG, when the performance is above the bar of “fair”, MME generally outperforms IQIM, but falls short when both methods become unreliable.

The advantages of using more general lattices for MME are further examined in Table III. In addition to the simple integer lattice \( Z \), the optimal lattices in 2, 4, and 8 dimensions are respectively \( A_2 \), \( D_4 \) and \( E_8 \). The Conway-Sloane type labeling technique [25] can be used for these lattices. Specifically, we have the following observations from the table. i)MME outperforms IQIM in all the metrics (SWR and ODG), regardless of the chosen type of lattices. For instance, the SWR of proposed lower bound in \( A_2 (R = 1) \) is 34.3771 dB, which is greater than the
28.3758 SWR in IQIM. ii) Choosing larger-dimensional optimal lattices is more desirable. Steady improvements can be observed by increasing the dimensions of the lattices from $\mathbb{Z}$ to $A_2$, $D_4$ and $E_8$. iii) Since the fine lattices are fixed to feature unit volumes while the coarse lattices grow as the code rate $R$ increases, the SWR/ODG decreases with $R$.

B. Robustness Performance

With the same setting as in Fig. 7(a), we further investigate the robustness performance of the watermarking algorithms against AWGN attacks. Fig. 9 has depicted the BER comparison for MME, Liang and IQIM. The Nishimura method has been excluded as its robustness performance is weak. The reason for MME’s better BER performance is that, by adjusting the scaling factor $\alpha$, its watermarked vector can be tuned further away from the border of the decoding regions.

Fig. 10(a) and (c) have drawn the GSNR for MME and IQIM for various SNR and $\alpha$. The GSNR of MME can go well beyond $10^{12}$, while IQIM flattens at about $3 \times 10^{11}$. Fig. 10(b) and (d) are also attached to show the better GSNR performance.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
of MME is not generated by scarifying the SWR performance. The simulation results above are remarkable: compared with other QIM-variants, MME can simultaneously achieve smaller distortion and better robustness.

Since both the BER and SWR metrics are related to $\Delta$ in MME, we plot the trade-off relations of BER versus SWR in Fig. 11 by changing $\Delta$ and letting SNR = 25 dB. We have the following observations: i) The trade-off capability becomes better when $\Delta$ increases, where the curve rises smoothly when $\Delta = 500$ and become sharper when $\Delta$ increases. ii) The ratio of SWR/BER decreases when $\Delta$ increases, which means larger distortion and higher robustness is achieved with a larger $\Delta$.

C. Reversibility

Due to the sensibility of HAS, the human awareness of sounds from different sources is also diverse, which may cause a distinct feeling for heterogeneous sounds. To vividly demonstrate the actual effect of embedding, we draw the original, watermarked and restored signal of four typical audio host signals in Fig. 12. They are absolute music, songs, speeches and white noises, respectively. These host signals are labeled as No. 2002, 2003, 2040 and 2050. Judging by the naked eye, the waveforms of the original and the restored signals are basically the same. In terms of watermarked signals, MME achieves the SWRs of 38.0775 dB, 33.2487 dB, 34.0998 dB and 35.1670 dB for these host signals. Lastly, Fig. 12(d) shows that are basically no difference between the original cover signals and the restored cover signals.

For noisy recovering, the audio files of setting datasets are labeled as No. 1-75. We respectively calculate the theoretical and experimental average distances for each file under AWGN attack of SNR = 40, 60 and 80. Results are plotted in Fig. 13, in which all subfigures justify the correctness of (28).

D. Discussion

It is noteworthy that the AWGN model produces simple and tractable mathematical formulations which are useful for gaining insight into the underlying behavior of a system before these other phenomena are considered. For other types of attacks, such as lossy compression, the advantage of MME in the AWGN channel can be easily replicated by using ad-hoc techniques.

For demonstration purposes, we further implement MME in two transformed domains to illustrate its robustness to MP3 compression. The first example, referred to as DCT-MME, applies MME to the discrete cosine transform (DCT) coefficient of each block divided from audio signal. In the second example,
Fig. 13. Theoretical and experimental average distances between restored and original signal.

Fig. 14. Comparisons between MME, DCT-MME, LWT-MME and benchmarks.

we apply MME to the approximation coefficients vector generated by lifting wavelet transform (LWT), which is named as LWT-MME. The simulation results are depicted in Fig. 14. We observe from Fig. 14(c) that DCT-MME and LWT-MME enjoy a sharp BER descending rate, and they both outperform Liang’s approach when the compression rate becomes larger than around 192kbps. Fig. 14(a) and (b) depict the imperceptibility and AWGN robustness of the transformed domain MME, whose performance is similar to that of time-domain MME.

VI. CONCLUSION

In this article, a novel reversible audio watermarking method referred to as meet-in-the-middle embedding has been proposed. We have justified its correctness, and make comparisons with existing methods, which show that MME exhibits less distortion and better robustness. Theoretical analyses on the SWR and GSNR have been presented. Simulations show that the proposed method features a smaller amount of distortion and probability of decoding errors compared with IQIM and other representative methods in RAW.

ACKNOWLEDGMENTS

The authors are grateful to the reviewers for their constructive suggestions that improved the presentation and quality of this article.

REFERENCES

[1] Y. Shi, X. Li, X. Zhang, H. Wu, and B. Ma, “Reversible data hiding: Advances in the past two decades,” IEEE Access, vol. 4, pp. 3210–3237, 2016.
[2] M. Li, C. Shan, Z. Tian, X. Du, and M. Guizani, “Adaptive information hiding method based on feature extraction for visible light communication,” IEEE Commun. Mag., vol. 61, no. 4, pp. 102–106, Apr. 2023.
[3] Z. Wang, G. Feng, H. Wu, and X. Zhang, “Data hiding during image processing using capsule networks,” Neurocomputing, vol. 537, pp. 49–60, 2023.
[4] Y. Tang, S. Wang, C. Wang, S. Xiang, and Y. Cheung, “A highly robust reversible watermarking scheme using embedding optimization and rounded error compensation,” IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 4, pp. 1593–1609, Apr. 2023.
[5] G. Hua, J. Huang, Y. Q. Shi, J. Goh, and V. L. L. Thing, “Twenty years of digital audio watermarking - A comprehensive review,” Signal Process., vol. 128, pp. 222–242, 2016.
[6] X. Li and Z. Guo, “General expansion-shifting model for reversible data hiding,” in Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Ann. Summiit Conf., 2016, pp. 1–4.
[7] J. Tian, “Reversible data embedding using a difference expansion,” IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 8, pp. 890–896, Aug. 2003.
[8] D. M. Thodi and J. J. Rodríguez, “Prediction-error based reversible watermarking,” in Proc. IEEE Int. Conf. Image Process., vol. 16, no. 3, pp. 354–362, Mar. 2006.
[9] F. Peng, Z. Lin, X. Zhang, and M. Long, “Reversible data hiding in encrypted 2D vector graphics based on reversible mapping model for real numbers,” IEEE Trans. Inf. Forensics Secur., vol. 14, no. 9, pp. 2400–2411, Sep. 2019.
[10] L. Ko, J. Chen, Y. Shieh, H. Hsin, and T. Sung, “Nested quantization index modulation for reversible watermarking and its application to healthcare information management systems,” Comput. Math. Methods Med., vol. 2012, pp. 839161:1–839161:8, 2012.
received the BS, the MS, and the PhD degrees in electrical and electronic engineering from the South China University of Technology, Guangzhou, China, in 2011 and 2014, respectively, and the PhD degree from the Electrical and Electronic Engineering Department, Imperial College London, U.K., in 2018. He is currently an associate professor with the College of Cyber Security, Jinan University, Guangzhou, China. He is the recipient of the 2021 CIE Information Theory Society Yong-sung Award, and the 2020 super-star super-star supervisor award of the National Crypto-Math Challenge of China. His research interests include lattice codes, wireless communications, and information security.

Shanshang Lyu received the BS and MS degrees in electronic and information engineering from the South China University of Technology, Guangzhou, China, in 2011 and 2014, respectively, and the PhD degree from the Electrical and Electronic Engineering Department, Imperial College London, U.K., in 2018. He is currently an associate professor with the College of Cyber Security, Jinan University, Guangzhou, China. He is the recipient of the 2021 CIE Information Theory Society Yong-sung Award, and the 2020 super-star super-star supervisor award of the National Crypto-Math Challenge of China. His research interests include lattice codes, wireless communications, and information security.

Junren Qin received the BS and MS degrees from the University of International Relations, in 2017, and Jinan University, in 2023, China, respectively. He is currently working toward the PhD degree with Jinan University, Guangzhou, China. His research interests include lattice coding, wireless communications, and algebraic geometry. He has published a series of papers in Crypto, Eurocrypt, IEEE Transactions on Information Theory, etc. He was the recipient of the NSFC outstanding young scientist grant, in 2002.

Xingyuan Liang received the BS and MS degrees from Nanchang Hangkong University, in 2017, and Jinan University, in 2020, of China, respectively. He is currently working toward the PhD degree with Jinan University, Guangzhou, China. His current research interest is reversible data hiding.

Hao Chen received the PhD degree in mathematics from the Institute of Mathematics, Fudan University, in 1991. He is currently a professor with the College of Information Science and Technology/Cyber Security, Jinan University. His research interests include coding and cryptography, quantum information and computation, lattices, and algebraic geometry. He has published a series of papers in Crypto, Eurocrypt, IEEE Transactions on Information Theory, etc. He was the recipient of the NSFC outstanding young scientist grant, in 2002.