

Efficient Speech Watermarking for Speech Synthesis via Progressive Knowledge Distillation

Yang Cui Peter Pan Lei He Sheng Zhao

Microsoft, Beijing, China

{yancu, peterpan, helei, szhao}@microsoft.com

Abstract—With the rapid advancement of speech generative models, unauthorized voice cloning poses significant privacy and security risks. Speech watermarking offers a viable solution for tracing sources and preventing misuse. Current watermarking technologies fall mainly into two categories: DSP-based methods and deep learning-based methods. DSP-based methods are efficient but vulnerable to attacks, whereas deep learning-based methods offer robust protection at the expense of significantly higher computational cost. To improve the computational efficiency and enhance the robustness, we propose PKDMark, a lightweight deep learning-based speech watermarking method that leverages progressive knowledge distillation (PKD). Our approach proceeds in two stages: (1) training a high-performance teacher model using an invertible neural network-based architecture, and (2) transferring the teacher’s capabilities to a compact student model through progressive knowledge distillation. This process reduces computational costs by 93.6% while maintaining high level of robust performance and imperceptibility. Experimental results demonstrate that our distilled model achieves an average detection F1 score of 99.6% with a PESQ of 4.30 in advanced distortions, enabling efficient speech watermarking for real-time speech synthesis applications.

Index Terms—speech watermarking, text-to-speech synthesis, progressive knowledge distillation, speech privacy.

I. INTRODUCTION

With the rapid advancement of deep learning-based generative models, text-to-speech (TTS) and voice cloning technologies have reached unprecedented levels of realism. While these innovations enable high-fidelity speech synthesis, they also introduce significant challenges in copyright protection, content authentication, and privacy security. The increasing difficulty in distinguishing synthetic speech from natural speech raises concerns about misinformation, deepfake attacks, and unauthorized content reproduction.

Audio watermarking has emerged as a proactive and effective solution by embedding traceable, imperceptible signals within speech, enabling robust identification and authentication [1]. Traditional audio watermarking methods relied on digital signal processing (DSP) to embed watermarks in time domain or transformation domains, demonstrating resilience against common distortions such as noise addition or lossy compression. However, their reliance on handcrafted features makes them susceptible to advanced manipulations, including reverberation and desynchronization attacks [2].

To overcome these limitations, researchers have explored deep learning-based methods that enhance the imperceptibility, robustness, and capacity of audio watermarks. These methods

generally fall into two categories: generative watermarking and post-hoc watermarking.

Generative watermarking techniques directly embed watermarks into the speech synthesis process, ensuring that all generated speech inherently carries a watermark. While this approach eliminates the need for additional processing, the embedded watermark is model-specific and lacks adaptability across different synthesis models. For example, Traceable-Speech [3] embeds watermarking into a TTS model, ensuring high imperceptibility while remaining resilient to audio clipping attacks. Similarly, [4] introduces a method for watermarking audio generative models at the latent representation level, ensuring robust detection. To enhance flexibility, HiFi-GANw [5] fine-tunes a HiFi-GAN vocoder [6] by jointly optimizing speech quality and watermark extraction loss, enabling robust watermark retrieval across various synthesis models. Building on the collaborative watermarking framework, [7] enhances the robustness to both traditional and neural audio codecs by employing a HiFi-GAN vocoder and an anti-spoofing detector. In another generative framework, GROOT [8] embeds watermarks during audio synthesis using diffusion models, achieving high extraction accuracy even under compound post-processing attacks.

On the other hand, post-hoc watermarking applies watermarking techniques after speech synthesis, offering flexible integration but requiring additional processing. As a representative example, WavMark [9] introduces an invertible neural network (INN)-based approach, allowing flexible watermark embedding and detection with high robustness. Extending this approach, DRAW [10] enhances resilience against desynchronization and replay attacks in a dual-decoder framework. While early approaches mainly ensured basic robustness, recent work increasingly targets advanced manipulations. DeAR [11] and DeepAWR [12] employ deep learning to improve the robustness against re-recording distortions. To further mitigate manipulation attacks, a method proposed in [13] adopts an encoder-decoder framework in the discrete wavelet transform (DWT) domain. Additionally, a multi-scale transformer model [14] extracts acoustic features to strengthen resistance against desynchronization. MaskMark [15] embeds a secret key vector through a multiplicative spectrogram mask to improve robustness against neural network-based transformations. Along with efforts on robustness, efficiency has also been explored. AudioSeal [16] employs an EnCodec [17]-based generator-detector architecture for fast, localized watermark de-

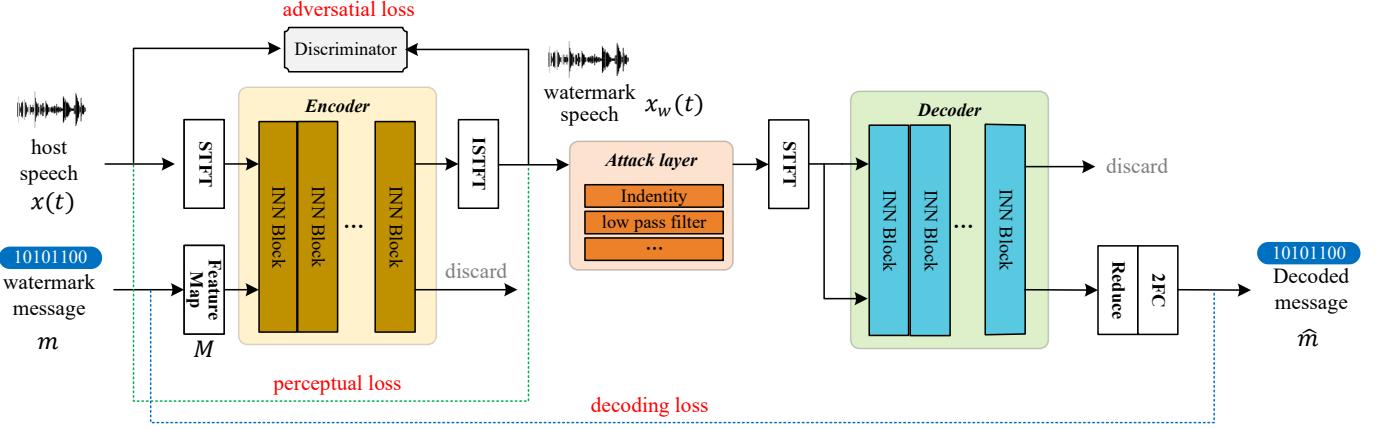


Fig. 1. Overview of the Proposed Watermark Framework.

tection at the sample level. By integrating psychoacoustic masking with a compact model design, SilentCipher [18] enhances message capacity and robustness while maintaining imperceptibility. Operating on discrete latent representations, DiscreteWM [19] employs vector-quantized autoencoders to embed watermarks, enhancing robustness while offering flexible capacity and efficient detection.

Deep learning-based speech watermarking offers strong robustness but often comes with high computational costs, limiting its practicality in real-time applications. In contrast, traditional DSP-based watermarking is computationally efficient but lacks resilience against sophisticated attacks. To enhance both efficiency and robustness, we propose PKDMark, a Progressive Knowledge Distillation (PKD) framework that combines the robustness of deep models with the efficiency of lightweight architectures, enabling practical deployment in real-world speech synthesis and transmission systems. The key contributions of the paper are summarized as follows:

(1) The robustness of an INN-based speech watermarking system against advanced signal distortions is enhanced by introducing a complex domain message feature map and a multi-objective optimization strategy that jointly improves perceptual quality and decoding accuracy.

(2) A PKD-based method is introduced to transfer the watermarking capability from a teacher model with strong robustness and high perceptual quality. Through progressive distillation, the lightweight student model incrementally learns these capabilities, enabling efficient real-time watermark embedding.

(3) Furthermore, the proposed framework exhibits strong scalability and generalizability, offering a viable solution for a broad class of deep learning-based post hoc watermarking systems beyond speech synthesis.

II. RELATED WORK

A. WavMark

WavMark [9] is a deep learning-based audio watermarking framework built on an invertible neural network (INN), en-

abling shared parameters for both watermark embedding and decoding. It encodes up to 32 bits per second while maintaining high imperceptibility and a low average bit error rate (BER) of 0.48% across ten types of attacks. WavMark further supports automatic localization, allowing precise watermark detection without external reference. Its robustness and domain versatility make it well-suited for voice cloning detection, copyright protection, and broadcast authentication.

B. Knowledge Distillation

Knowledge distillation (KD) [20] is a well-known neural model compression method where a smaller student model is generally supervised by a larger teacher model to obtain a competitive performance. Progressive knowledge distillation (PKD) [21] is an extension of traditional knowledge distillation, designed to enhance the knowledge transfer process from teacher to student. Instead of transferring all knowledge at once, PKD gradually refines the student model over intermediate learning stages. This approach has been demonstrated to be effective in improving the performance of large language model [22], image classification and generation [23]–[25] and information retrieval [26]. Different from prior PKD studies, the proposed PKD approach explores knowledge distillation in the context of speech watermarking for the first time. By jointly optimizing both the teacher and student models, it enables more effective knowledge transfer and improves the student model's ability to embed robust watermarks.

III. METHOD

A. Overview of framework

The proposed method adopts an *encoder–attack–decoder* pipeline to enable robust speech watermarking, where an intermediate attack layer simulates real-world distortions to enforce resilience against manipulations. An overview of the proposed method framework is shown in Figure 1. The process begins by converting the input host speech signal $x(t)$ into a complex-valued spectrogram $X \in \mathbb{C}^{T \times F}$ via Short-Time Fourier Transform (STFT), where T denotes the number of

time frames and F the number of frequency bins. A binary message $m \in \{0, 1\}^K$ of length K is mapped and expanded to a complex feature map $M \in \mathbb{C}^{T \times F}$. The encoder network fuses X and M to generate a watermarked spectrogram X_w , which is subsequently transformed back to the time domain using inverse STFT, yielding the watermarked speech signal $x_w(t) = \text{ISTFT}(X_w)$. To simulate real-world distortions, an attack layer applies various of distortions before message decoding, ensuring watermark resilience against different types of manipulations. The distorted speech $x_a(t)$ is then transformed into a complex STFT spectrogram \hat{X}_a and passed through the decoder to recover a decoded feature map \hat{M} . Finally, the decoded message \hat{m} is obtained by reducing \hat{M} along the temporal dimension and feeding the reduced vector through two feedforward layers.

Building on this foundation, our main contribution is a two-stage training strategy that enhances robustness, imperceptibility, and efficiency. In the first stage, a teacher model is trained for end-to-end audio watermarking using an invertible neural network (INN), guided by a multi-objective optimization scheme. In the second stage, a lightweight student model is trained via progressive knowledge distillation (PKD), where supervision gradually transitions from the teacher's guidance to the student's own predictions. This design enables efficient watermark embedding while preserving the strong robustness and high perceptual quality of the teacher watermarking system. For watermark detection, the teacher decoder is retained to ensure accurate and reliable message recovery.

B. Teacher Model

The teacher model is trained by jointly optimizing the encoder and decoder to balance imperceptibility and robustness, with an adversarial discriminator further enhancing speech fidelity. Compared to WavMark [9], our approach introduces a novel message feature map in the complex STFT domain, along with a multi-objective joint optimization strategy. By exploiting both amplitude and phase information, it alleviates distortions caused by phase-amplitude mismatches during ISTFT reconstruction. The message feature map embeds information in both the real and imaginary components, ensuring alignment with the time and frequency structures in the complex STFT domain of the original speech. Additionally, the multi-objective optimization ensures an optimal balance between imperceptibility and robustness, resulting in superior performance of the teacher model compared to WavMark.

1) *Model Architecture:* The teacher model utilizes an invertible neural network (INN) [9] in which the encoder and decoder share parameters, ensuring strict reversibility in both forward and inverse processes. Training procedure follows an encoder-attack-decoder framework, incorporating a signal processor, message feature mapping, encoder, decoder, attack layer, and discriminator. The detailed modules and training flow are as follows.

i. **Signal processor:** Input speech is converted to the STFT domain using a sliding window and FFT analysis, producing a

complex representation. The watermarked STFT spectrogram is transformed back to the time domain using ISTFT.

ii. **Message Feature Map:** An embedding layer maps each bit to a corresponding message embedding vector. A two-layer feedforward neural network transforms the embedding vectors to align with the frequency dimension F of complex STFT spectrogram X of the host speech. The hidden vectors are then summed and averaged, forming a compact representation \mathbf{h} . Finally, the transformed representation is repeated along the time and complex axes, ensuring it matches the time-frequency dimension of X , resulting in a complex message feature map.

$$\mathbf{M} = \text{Repeat}_T(\mathbf{h} + j \cdot \mathbf{h}) \quad (1)$$

where $\mathbf{M} \in \mathbb{C}^{T \times F}$ is the complex-valued message feature map aligned with the shape of the host speech STFT. The vector $\mathbf{h} \in \mathbb{R}^F$ represents the average output of a two-layer feedforward neural network applied to the message embeddings. The operator $\text{Repeat}_T(\cdot)$ denotes repeating and concatenating the vector \mathbf{h} along the time axis T times. Here, $j = \sqrt{-1}$ is the imaginary unit, and the same vector \mathbf{h} is used as both the real and imaginary part. This design enables the message embedding to jointly leverage both the real and imaginary parts of the complex-valued representation.

iii. **Encoder:** The encoder receives the complex STFT spectrogram of the original audio and the message feature map, transforming them through a few INN blocks, resulting in the complex watermarked STFT spectrogram.

iv. **Decoder:** The decoder performs the inverse process of the encoder. For simplicity, it takes the watermarked STFT as dual input to extract the watermark feature map. A reduction operation averages the hidden vectors across the time and complex axes. Finally, a two-layer feedforward neural network reconstructs the extracted watermark bits from the decoded message feature map.

v. **Attack Layer:** Fourteen types of distortions (detailed in Section IV.C) are introduced to interfere with the watermarked signal, comprising nine basic distortions and five advanced attacks. These attacks simulate practical conditions, including degradation, cropping, and noise, ensuring the watermark remains robust against basic and advanced challenges.

vi. **Discriminator:** The discriminator is to distinguish the watermarked speech and the original speech, providing adversarial feedback to the entire network by computing a cross-entropy loss.

2) *Multi-Objective Joint Optimization:* During the training of the teacher model, multiple loss terms are introduced and jointly optimized to balance imperceptibility, robustness, and decoding accuracy:

- **Perceptual Loss:** A multi-scale STFT loss [16] is adopted to measure the difference in the time-frequency domain between the watermarked speech and the original speech. Specifically, magnitude spectral differences are computed at various window sizes and frequency resolutions and then combined via weighted summation. The perceptual loss \mathcal{L}_{per} can be

formulated as:

$$\mathcal{L}_{\text{per}} = \sum_{i \in e} \alpha_i \|STFT_i(x(t)) - STFT_i(x_w(t))\|_2 \quad (2)$$

where $x(t)$ represents the original speech, $x_w(t)$ represents watermarked speech. $STFT_i$ is a normalized STFT with a window size of 2^i and a hop length of $2^{i/4}$. $e = 5, \dots, 11$ is the set of scales and α represents the set of scalar coefficients that balance the terms.

- **Decoding Loss:** The MSE loss is used to measure the difference between the decoded message and the original message. Additionally, another MSE loss is computed between the original and decoded message feature maps to ensure fine-grained alignment in the time-frequency domain, thereby ensuring accurate recovery of the embedded information. The decoding loss \mathcal{L}_{dec} is defined as:

$$\mathcal{L}_{\text{dec}} = \|m - \hat{m}\|_2 + \alpha \|\mathbf{M} - \hat{\mathbf{M}}\|_2 \quad (3)$$

where m denotes the embedded message and \hat{m} is the decoded message. \mathbf{M} represents the message feature map in the encoder, while $\hat{\mathbf{M}}$ denotes the corresponding feature map predicted in the decoder. The scalar α is a balancing factor that controls the contribution of the second item.

- **Adversarial Loss:** The adversarial loss is computed via the discriminator, which attempts to differentiate between watermarked audio $x_w(t)$ and the original audio x . The resulting cross-entropy loss drives the adversarial training process, thereby improving the fidelity of the watermarked audio.

$$\mathcal{L}_{\text{dis}} = \log(1 - d(x(t))) + \log(d(x_w(t))) \quad (4)$$

$$\mathcal{L}_{\text{adv}} = \log(1 - d(x_w(t))) \quad (5)$$

where \mathcal{L}_{dis} and \mathcal{L}_{adv} denote the loss functions of the discriminator and the encoder, respectively. $d(\cdot)$ denotes the output of the discriminator, representing the probability that the input is identified as watermarked audio.

The total loss $\mathcal{L}_{\text{total}}$ is computed as a weighed sum of the above loss terms:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{per}} \mathcal{L}_{\text{per}} + \lambda_{\text{dec}} \mathcal{L}_{\text{dec}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}, \quad (6)$$

where λ_{per} , λ_{dec} , and λ_{adv} are the corresponding weights of the perceptual loss \mathcal{L}_{per} , the decoding loss \mathcal{L}_{dec} and the adversarial loss \mathcal{L}_{adv} .

C. Student Model

To reduce model size and improve inference efficiency, direct knowledge distillation (DKD) is initially explored by jointly training the student encoder with the teacher decoder. However, experiments (detailed in Section IV.E) show that this approach falls short of the teacher model's performance, as it primarily transfers knowledge at the output layer while overlooking feature representation learning. To address this, we propose a progressive knowledge distillation (PKD) method in training, which gradually transfers watermark knowledge from the teacher to the student model.

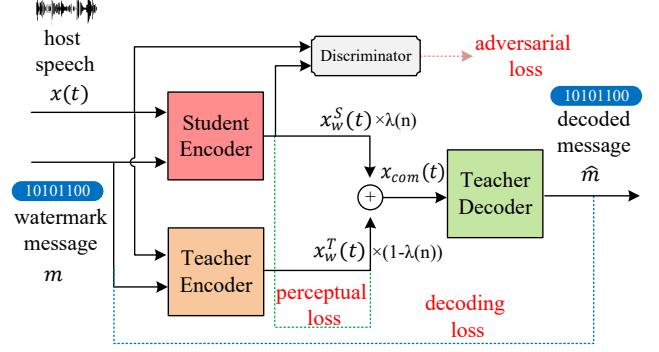


Fig. 2. Overview of the PKD Framework.

1) **Progressive Knowledge Distillation:** The overall PKD framework is illustrated in Figure 2. Specifically, the student encoder output $x_w^S(t)$ and the teacher encoder output $x_w^T(t)$ are linearly combined as follows:

$$x_{\text{com}}(t) = \lambda(n) \cdot x_w^S(t) + (1 - \lambda(n)) \cdot x_w^T(t), \quad (7)$$

where $\lambda(n)$ denotes the mixing factor of student model, with n representing the training step. The combined output X_{com} is decoded by the teacher decoder. During training, $\lambda(n)$ gradually increases from 0 to 1. Initially, When $\lambda(n) = 0$, the training framework is equivalent to the teacher model training framework. Conversely, when $\lambda(n) = 1$, it corresponds to the direct knowledge distillation framework.

Adopting a progressive strategy to increase $\lambda(n)$ enables the student model to strike a better balance between perceptual quality and decoding robustness. In the early training stage, a smaller $\lambda(n)$ ensures that gradient updates are primarily guided by the teacher network, providing a stable supervisory signal. This guidance helps the student network build a strong initial representation and prevents it from getting trapped in local optima. As training progresses and the student network improves, gradually increasing $\lambda(n)$ shifts the gradient updates to rely more on the student's own outputs. This transition encourages the student model to improve its decoding performance by optimizing its own representations while still benefiting from subtle guidance from the teacher network. In the later stage, when $\lambda(n) = 1$, the student model has effectively learned the teacher's representations and can be jointly optimized with the teacher decoder to further enhance performance.

2) **Joint optimization strategy:** The student model adopts the same attack layer, discriminator, message feature mapping, and adversarial loss as the teacher model. The main points of the joint optimization strategy are as follows:

- **Perceptual Loss:** Different from the first stage, the student perceptual loss is computed as the MSE between the watermark speech of student model $x_w^S(t)$ and the watermark speech of teacher model $x_w^T(t)$. This alignment ensures that the student model's watermark embedding closely follows the

TABLE I
PERFORMANCE COMPARISON WITH TEST SPEECH WATERMARKING METHODS. BEST PERFORMANCE IN EACH COLUMN IS HIGHLIGHTED IN BOLD.

Models	BPS(\uparrow)	PESQ(\uparrow)	SNR(dB)(\uparrow)	BER(%)(\downarrow)															
				MEAN	ND	GN	MF	MP3	LP	DS	QT	AS	EA	SS	FS	RA	PU	PD	
WavMark	32	4.31	39.99	8.58	0.19	1.17	3.24	0.69	1.07	1.92	1.96	0.32	0.21	2.70	3.69	5.56	49.46	47.94	
AudioSeal	16	4.28	27.53	18.66	0.00	0.56	52.12	59.50	47.09	50.69	51.25								
WavMark-16bps	16	4.29	41.56	2.43	0.54	1.14	2.52	0.84	0.78	1.26	1.32	0.54	0.60	3.25	3.06	9.56	5.23	6.49	
Teacher model	16	4.33	45.75	0.56	0.00	0.38	0.46	0.13	0.07	0.85	0.13	0.00	1.11	1.20	1.45	1.04	0.98		
Teacher-woMSSL	16	4.29	40.34	1.23	0.00	0.76	2.16	0.18	0.16	1.24	0.24	0.00	0.12	2.52	2.04	6.79	2.52	2.76	
Teacher-woCMFM	16	4.30	42.25	1.77	0.60	0.84	1.68	0.84	0.78	0.96	0.90	0.54	1.02	1.86	1.86	7.03	3.43	4.45	
DKD Student	16	4.08	44.29	1.94	0.00	2.41	1.50	0.91	2.93	6.38	0.39	0.07	0.20	3.13	3.19	2.93	4.10	2.54	
PKDMark	16	4.30	45.54	0.51	0.00	0.39	0.58	0.19	0.32	1.23	0.00	0.00	0.52	0.85	1.58	0.65	0.84		

teacher model's representation in the time-frequency domain, enhancing perceptual similarity and embedding consistency.

$$\mathcal{L}_{\text{per}}^S = \sum_{i \in e} \alpha_i \|STFT_i(x_w^S(t)) - STFT_i(x_w^T(t))\|_2 \quad (8)$$

- **Decoding Loss:** During joint optimization, the teacher decoder is utilized to decode the messages from the combined output $x_{\text{com}}(t)$ and compute the decoding loss. The MSE loss of encoder and decoder message feature map in first stage is not included during the distillation process.

$$\mathcal{L}_{\text{dec}}^S = \|m^S - \hat{m}^S\|_2 \quad (9)$$

- **Adversarial Loss:** The discriminator of the first stage is utilized in the second stage to provide the adversarial loss of student encoder $\mathcal{L}_{\text{adv}}^S$.

$$\mathcal{L}_{\text{adv}}^S = \log(1 - d(x_w^S(t))) \quad (10)$$

- **Optimization Strategy:** As in the first stage, the total loss $\mathcal{L}_{\text{total}}^S$ is computed as a weighted sum of the above three terms.

$$\mathcal{L}_{\text{total}}^S = \lambda_{\text{per}} \mathcal{L}_{\text{per}}^S + \lambda_{\text{dec}} \mathcal{L}_{\text{dec}}^S + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}^S, \quad (11)$$

To maintain training stability, the teacher encoder and decoder's parameters are initially kept fixed, with updates applied only to the student encoder. Once the student model gradually aligns with the teacher's performance as $\lambda(n)$ reaches 1, a final global fine-tuning step is performed. This approach effectively leverages knowledge from both the teacher encoder and decoder, allowing the student model to converge efficiently while preserving speech quality.

IV. EXPERIMENTS

A. Experimental setup

- **Datasets.** The experiments are conducted using a composite speech dataset totaling 1,000 hours. It comprises 400 hours from CommonVoice [27], 300 hours from LibriTTS [28], 200 hours from ASVspoof 2021 DF [29], and 100 hours of high-quality synthetic speech generated via Microsoft Azure Text-to-Speech. This combination provides diverse, natural, and multilingual speech, while ensuring applicability to both real-world speech and DeepFake detection scenarios. The dataset is split into 900/50/50 hours for training/validation/testing. All speech samples are processed at a 24 kHz sampling rate. The STFT and ISTFT operations are performed with an FFT size of 1,024 samples and a hop length of 600 samples.

- **Implementation.** The PKD framework begins with training a teacher model, followed by iterative PKD training to obtain a compact student model. The embedding capacity is set to 16 BPS, which is sufficient for speech synthesis applications. The teacher model comprises 8 INN blocks, each containing a 5-layer ResNet module with 32 channels per layer, while the student model is significantly smaller, with 2 INN blocks and 16 channels per layer. During PKD training, the mixing factor $\lambda(n)$ is linearly increased from 0.1 to 1 over 40k steps.

- **Evaluation Metrics.** Imperceptibility and robustness are assessed using both objective and subjective metrics. For imperceptibility, signal-to-noise ratio (SNR) quantifies watermarking distortion, while perceptual evaluation of speech quality (PESQ) [30] measures perceptual quality. Additionally, the subjective comparison mean opinion score (CMOS) captures human perception of quality differences. Robustness is evaluated using bit error rate (BER) to assess the reliability of extracted information and detection F1 score to measure accuracy in distinguishing between original and watermarked speech.

B. Imperceptibility

To evaluate imperceptibility, synthesized speech samples from both the teacher and student models are compared against the original recordings. PESQ scores reported in Table I indicate minimal quality differences between the teacher (4.33) and student (4.30) models. Furthermore, a CMOS test yields a score of -0.04 when comparing original recordings with student model samples, suggesting that PKDMark is nearly indistinguishable from the original speech. These results confirm that the distillation process effectively preserves high imperceptibility.

C. Robustness

The robustness is evaluated under various distortions, as summarized in Table I. The tested distortions include: 1) no distortion (ND); 2) Gaussian noise (GN); 3) median filtering (MF); 4) MP3 compression at 64kbps (MP3); 5) low pass filtering with 4kHz stop band (LP); 6) down sampling to 8kHz (DS); 7) quantization to 9bit (QT); 8) amplitude scaling (AS); 9) echo addition with 100ms delay (EA); 10) slow speed to 0.9x (SS); 11) fast speed to 1.1x (FS); 12) reverberation attack with RT60 = 200ms (RA); 13) pitch up shifting by +10% (PU); 14) pitch down shifting by -10% (PD). Distortions 1) to 9) are basic distortions, while 10) to 14) are advanced attacks, which

pose challenges even for some deep learning-based watermark methods [31] [32]. Two open-source models are selected for comparison: WavMark [9] and AudioSeal [16]. To make a fair comparison, WavMark-16bps is trained at a 16bps bit rate, with the remaining settings kept the same as the original WavMark.

The results presented in Table I demonstrate that PKDMark achieves an average BER of 0.51% across all various distortions, which is comparable to the teacher model's average BER of 0.56%. Furthermore, PKDMark exhibits superior robustness compared to WavMark and WavMark-16bps across all test distortions, especially under advanced attacks. In particular, PKDMark consistently exceeds WavMark (8. 58%) and WavMark-16bps (2. 43%) in all types of distortions.

Compared to AudioSeal, PKDMark achieves significantly better robustness under advanced distortions. While AudioSeal obtains perfect BERs (0.00%) in several basic distortions, its performance degrades drastically under more complex attacks such as time stretching and pitch shifting, resulting in a high average BER of 18.66%. In contrast, PKDMark maintains consistently low BERs in most distortion scenarios, highlighting its strong generalization capability under challenging conditions.

To further evaluate watermark reliability, watermark speech detection is performed on a test dataset of 1000 TTS samples, each lasting between 1 and 2 seconds and embedded with an 8-bit synchronization code from the 16-bit capacity. The same distortions in Table I are applied. A simple detection strategy is used: if at least 7 out of 8 bits match, the sample is classified as watermarked speech; otherwise, it is classified as original speech. The average detection F1 score reaches 99.6% across test distortions. These results confirm the robustness of our approach in preserving critical watermark message and ensuring reliable detection even under advanced manipulations.

D. Efficiency

Table II provides a summary of the efficiency gains achieved by PKDMark in terms of computational cost and real-time performance. The computational cost is measured in floating point operations per second (FLOPS). The teacher model requires 36.0 GFLOPS, whereas PKDMark reduces this to 2.3 GFLOPS, resulting in a 93.6% reduction. Real-time performance is evaluated across both GPU and CPU configurations. All evaluations are conducted using sequential single-threaded processing, without leveraging parallel or multi-threaded inference. On an NVIDIA T4 GPU, PKDMark achieves a real-time factor (RTF) of 0.006, representing an 84.7% improvement over the teacher model's RTF of 0.039. On an Intel i9-13900 CPU, PKDMark achieves an RTF of 0.020, marking an 87.1% improvement compared to the teacher model's 0.155. These results highlight the efficiency and practical deployment potential of PKDMark for real-time speech synthesis across diverse hardware environments.

TABLE II
COMPUTATIONAL COST AND REAL-TIME PERFORMANCE OF PKDMARK.

Model	GFLOPS	RTF (↓)	
		T4 (GPU)	i9-13900 (CPU)
Teacher Model	36.0	0.039	0.155
PKDMark	2.3	0.006	0.020
Reduction (%)	93.6%	84.7%	87.1%

E. Ablation study

The effectiveness of the proposed components is assessed through a series of ablation experiments, with results summarized in Table 1. For the teacher model, Teacher-woCMFM removes the proposed complex message feature map from the teacher model, resulting in an increased BER of 1.77% and reduced SNR of 42.25 dB. Teacher-woMSSL, which removes the multi-scale STFT loss, yields an increased BER of 1.23% and reduced SNR of 40.34 dB. These findings demonstrate that both the proposed complex feature map and multi-objective joint optimization are essential to improve the robustness and quality of the watermark.

For the distillation strategy, PKD Student (PKDMark) and DKD Student are trained under identical configurations, differing only in the distillation method. As shown in Table 1, PKD Student achieves superior performance in PESQ (4.30 vs. 4.08) and BER (0.51% vs. 1.94%). These results suggest that the PKD strategy enables more effective knowledge transfer, leading to a better balance between perceptual quality and watermark robustness.

V. CONCLUSIONS

In this paper, we propose PKDMark, an efficient speech watermarking method for real-time speech synthesis via progressive knowledge distillation (PKD). Our approach enhances an invertible neural network (INN)-based teacher model to embed watermarks that preserve high fidelity and demonstrate robustness against 14 types of distortions and attacks. By transferring watermarking knowledge from the teacher model to a compact student model, we achieve a significant reduction of 93.6% in computational cost. Ablation studies further confirm that the proposed enhancements contribute to superior performance relative to the baseline model, with PKD outperforming the DKD approach in both perceptual quality and robustness. These results demonstrate the effectiveness of the proposed PKDMark method, showing that it achieves superior robustness and imperceptibility particularly under advanced attacks compared to the two speech watermarking methods evaluated in this study. Future work will focus on expanding watermark capacity and strengthening robustness to more advanced adversarial conditions, such as neural codec compression and neural denoising attacks.

REFERENCES

- [1] G. Hua, J. Huang, Y. Q. Shi, J. Goh, and V. L. Thing, “Twenty years of digital audio watermarking—a comprehensive review,” *Signal processing*, vol. 128, pp. 222–242, 2016.

- [2] Z. Liu, Y. Huang, and J. Huang, "Patchwork-based audio watermarking robust against de-synchronization and recapturing attacks," *IEEE transactions on information forensics and security*, vol. 14, no. 5, pp. 1171–1180, 2018.
- [3] J. Zhou, J. Yi, T. Wang, J. Tao, Y. Bai, C. Y. Zhang, Y. Ren, and Z. Wen, "Traceablespeech: Towards proactively traceable text-to-speech with watermarking," in *Interspeech 2024 - Annual Conference of the International Speech Communication Association*. ISCA, 2024, pp. 2250–2254.
- [4] R. San Roman, P. Fernandez, A. Deleforge, Y. Adi, and R. Serizel, "Latent watermarking of audio generative models," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [5] X. Cheng, Y. Wang, C. Liu, D. Hu, and Z. Su, "Hifi-ganw: Watermarked speech synthesis via fine-tuning of hifi-gan," *IEEE Signal Processing Letters*, 2024.
- [6] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020, pp. 17022–17033.
- [7] L. Juvela and X. Wang, "Audio codec augmentation for robust collaborative watermarking of speech synthesis," in *ICASSP 2025 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [8] W. Liu, Y. Li, D. Lin, H. Tian, and H. Li, "Groot: Generating robust watermark for diffusion-model-based audio synthesis," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 3294–3302.
- [9] G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei, "Wavmark: Watermarking for audio generation," *arXiv preprint arXiv:2308.12770*, 2023.
- [10] B. Li, J. Chen, Y. Xu, W. Li, and Z. Liu, "Draw: Dual-decoder-based robust audio watermarking against desynchronization and replay attacks," *IEEE Transactions on Information Forensics and Security*, 2024.
- [11] C. Liu, J. Zhang, H. Fang, Z. Ma, W. Zhang, and N. Yu, "Dear: A deep-learning-based audio re-recording resilient watermarking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13201–13209.
- [12] G. Lin, W. Luo, P. Zheng, and J. Huang, "An audio watermarking method against re-recording distortions," *Pattern Recognition*, p. 111366, 2025.
- [13] S. Wen, Q. Zhang, T. Hu, and J. Li, "Robust audio watermarking against manipulation attacks based on deep learning," *IEEE Signal Processing Letters*, 2024.
- [14] C. Tong, I. Natgunanathan, Y. Xiang, J. Li, T. Zong, X. Zheng, and L. Gao, "Enhancing robustness of speech watermarking using a transformer-based framework exploiting acoustic features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [15] P. O'Reilly, Z. Jin, J. Su, and B. Pardo, "Maskmark: Robust neuralwatermarking for real and synthetic speech," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4650–4654.
- [16] R. San Roman, P. Fernandez, H. Elsahar, A. D'efossez, T. Furion, and T. Tran, "Proactive detection of voice cloning with localized watermarking," *ICML*, 2024.
- [17] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, vol. 2023, 2023, iCLR 2024 Journal Track.
- [18] M. K. Singh, N. Takahashi, W. Liao, and Y. Mitsufuji, "Silentcipher: Deep audio watermarking," in *Interspeech 2024 - Annual Conference of the International Speech Communication Association*, 2024, pp. 2235–2239.
- [19] S. Ji, Z. Jiang, J. Zuo, M. Fang, Y. Chen, T. Jin, and Z. Zhao, "Speech watermarking with discrete intermediate representations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.
- [20] G. Hinton, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [21] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International journal of computer vision*, vol. 129, no. 6, pp. 1789–1819, 2021.
- [22] X. Xu, M. Li, C. Tao, T. Shen, R. Cheng, J. Li, C. Xu, D. Tao, and T. Zhou, "A survey on knowledge distillation of large language models," *arXiv preprint arXiv:2402.13116*, 2024.
- [23] W. Shi, Y. Song, H. Zhou, B. Li, and L. Li, "Follow your path: a progressive method for knowledge distillation," in *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2021, pp. 596–611.
- [24] M. Rezagholizadeh, A. Jafari, P. Salad, P. Sharma, A. S. Pasand, and A. Ghodsi, "Pro-kd: Progressive distillation by following the footsteps of the teacher," in *COLING 2022 - International Conference on Computational Linguistics*. ICCL, 2022, pp. 4714–4727.
- [25] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *ICLR 2022 - International Conference on Learning Representations*. ICLR, 2022.
- [26] Z. Lin, Y. Gong, X. Liu, H. Zhang, C. Lin, A. Dong, J. Jiao, J. Lu, D. Jiang, R. Majumder *et al.*, "Prod: Progressive distillation for dense retrieval," in *Proceedings of the ACM Web Conference*, 2023, pp. 3299–3308.
- [27] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, R. Henretty, M. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.
- [28] H. Zen, V. Dang, and R. Clark, "Libritts: A corpus derived from librispeech for text-to-speech," in *Interspeech 2019 - Annual Conference of the International Speech Communication Association*. ISCA, 2019, pp. 1526–1530.
- [29] J. Yamagishi, M. Todisco, X. Wang, A. Nautsch, H. Delgado, N. Evans, T. Kinnunen, M. Sahidullah, V. Vestman *et al.*, "Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," in *Interspeech 2021 - Annual Conference of the International Speech Communication Association*. ISCA, 2021.
- [30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq): A new method for speech quality assessment of telephone networks and codecs," in *ICASSP 2001 - IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 2001, pp. 749–752.
- [31] C.-H. Wu, W. Ge, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "A comparative study on proactive and passive detection of deepfake speech," in *Interspeech 2025 - Annual Conference of the International Speech Communication Association*. ISCA, 2025.
- [32] Y. Özer, W. Choi, J. Serrà, M. K. Singh, W.-H. Liao, and Y. Mitsufuji, "A comprehensive real-world assessment of audio watermarking algorithms: Will they survive neural codecs?" in *Interspeech 2025 - Annual Conference of the International Speech Communication Association*. ISCA, 2025.